Received: 15 March 2023 | Revised: 11 May 2023 | Accepted: 6 June 2023

DOI: 10.1049/qtc2.12063

IET Quantum Communication

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Deep reinforcement learning-based routing and resource assignment in quantum key distribution-secured optical networks

Purva Sharma[1] | Shubham Gupta[1] | Vimal Bhatia[1,2] | Shashi Prakash[3]

[1]Department of Electrical Engineering, Indian Institute of Technology (IIT) Indore, Indore, India

[2]Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic

[3]Department of Electronics and Instrumentation Engineering, Institute of Engineering and Technology, Devi Ahilya University, Indore, India

Correspondence

Vimal Bhatia.
Email: vbhatia@iiti.ac.in

Funding information

Indo-US Science and Technology Forum, Grant/Award Number: IUSSTF/JC-089/2019; Ministry of Education, India

Abstract

In quantum key distribution-secured optical networks (QKD-ONs), constrained network resources limit the success probability of QKD lightpath requests (QLRs). Thus, the selection of an appropriate route and the efficient utilisation of network resources for establishment of QLRs are the essential and challenging problems. This work addresses the routing and resource assignment (RRA) problem in the quantum signal channel of QKD-ONs. The RRA problem of QKD-ONs is a complex decision making problem, where appropriate solutions depend on understanding the networking environment. Motivated by the recent advances in deep reinforcement learning (DRL) for complex problems and also because of its capability to learn directly from experiences, DRL is exploited to solve the RRA problem and a DRL-based RRA scheme is proposed. The proposed scheme learns the optimal policy to select an appropriate route and assigns suitable network resources for establishment of QLRs by using deep neural networks. The performance of the proposed scheme is compared with the deep-Q network (DQN) method and two baseline schemes, namely, first-fit (FF) and random-fit (RF) for two different networks, namely The National Science Foundation Network (NSFNET) and UBN24. Simulation results indicate that the proposed scheme reduces blocking by 7.19%, 10.11%, and 33.50% for NSFNET and 2.47%, 3.20%, and 19.60% for UBN24 and improves resource utilisation up to 3.40%, 4.33%, and 7.18% for NSFNET and 1.34%, 1.96%, and 6.44% for UBN24 as compared with DQN, FF, and RF, respectively.

KEYWORDS

deep reinforcement learning, optical network, quantum key distribution, routing and resource assignment

1 | INTRODUCTION

With the rise of various high security-hungry applications, such as finance, cloud-based and several other government services, the importance of optical network security is growing rapidly. This decade will be expected to witness a surge in quantum computers' availability and capability. This evolution of quantum computers is expected to easily break security of the existing and the future optical networks as their security is built on the conventional cryptographic algorithms [1–3]. Thus, to secure the data on optical networks, quantum key distribution (QKD) is proposed as a solution. Quantum keys enhance the security of optical networks, generated by using the QKD technique [4–9] as QKD is based on the fundamental principles of quantum mechanics, namely, the Heisenberg's uncertainty principle [10] and the quantum no-cloning theorem [11], instead of the computational complexity of algorithms [4, 5, 12]. These fundamental principles ensure that a third party trying to eavesdrop on a secret key is easily detected. QKD generates and distributes secret keys over an insecure communication channel using QKD protocols, such as BB84 [13, 14] and others [7, 15–17]. The generated secret keys are then used to encrypt/decrypt the information. The generated quantum keys are impossible to copy because of the fundamental principles of quantum mechanics [11]. This is an important advantage of QKD systems over current and any future public-key cryptography/classical systems/methods. Hence, QKD, when combined with optical networks, provides long-term security against security breaches compared to the conventional cryptographic systems/methods [4, 5].

A QKD-secured optical network (QKD-ON) involves the realisation of quantum signal channel for transmission of quantum bits, public interaction channel for verification of the exchanged key information (these two channels form a QKD system), as well as the traditional data channel for encrypted data transmission between the sender and the receiver [18–21]. A cost-efficient solution for deployment of QKD-ONs is to integrate QKD (quantum signal channel/public interaction channel) into existing optical networks (traditional data channel) using wavelength division multiplexing as shown in Figure 1 [18, 19]. In [22], for better transmission performance, it was proposed that the quantum signal channel can be placed at the highest frequency in the C-band with a large guard band of 200 GHz between the quantum signal channel and the two classical channels (traditional data channel/public interaction channel) [23]. However, the co-existence of quantum signal channel and the two classical channels introduce various networking challenges, such as routing, wavelength and time-slot allocation [18, 19], trusted repeater node placement [24], survivability [25], quantum key recycling [26], and QKD for multicast service [27] in QKD-ONs [5, 18].

The routing and wavelength assignment is one of the most important networking challenges of QKD-ONs since few network resources are reserved for quantum signal channel due to limited network resources in a single optical fiber. Therefore, in order to utilise the network resources more efficiently, optical time division multiplexing has been used to construct the quantum signal channel [18]. Hence, the modified routing and wavelength assignment problem of QKD-ONs is termed as routing, wavelength and time-slot assignment [19]. In addition, to prevent the encrypted data from the eavesdropper, quantum keys of QKD lightpath requests (QLRs) must be updated often. This unique updation/modification feature of keys enhances the security level of the QLRs, where resources in the quantum signal channel should be reassigned periodically to update the key of QLRs [19]. These diverse assignment and reassignment (during QLR modification because of unique key updation/modification feature) of network resources for establishment of QLR make routing, wavelength and time-slot assignment or routing and resource assignment (RRA) problem of QKD-ONs complex, challenging, and different from the conventional optical networks [18, 19]. Furthermore, RRA is one of the most important networking challenges of QKD-ONs [18], especially in dynamic traffic scenario, where lightpath requests are not known in advance.

In the recent years, comprehensive research has been conducted on different networking challenges of QKD-ONs [5, 18]. To solve the RRA problem of QKD-ONs, various schemes have been proposed in the literature [18, 19]. The RRA problem was first investigated in [19], and an integer linear programming model along with the heuristic algorithms was proposed to solve this problem in a static traffic scenario, where the set of QLRs is known a priori. In the context of dynamic traffic scenario, where the arrival and the departure of QLRs are not known, the RRA algorithm was proposed in [18] to address the resource assignment problem. Moreover, to prevent the encrypted QLR from eavesdroppers, periodic updation/modification of the secret key was introduced in [19]. In the literature, different security-level provisioning solutions were proposed in [18, 19] to solve the aforementioned problem. A new key on-demand approach with the quantum key pool construction technique over a software-defined optical network was proposed to maintain a balance between the security level and the resource utilisation efficiency [28]. In [29], a time-scheduled scheme with the quantum key pool technique by considering three sub-problems was presented for the purpose of providing sufficient secret keys across QKD-ONs. Furthermore, two new heuristic routing, wavelength and time-slot assignment algorithms were proposed based on the new node structure and auxiliary graph, where some trusted

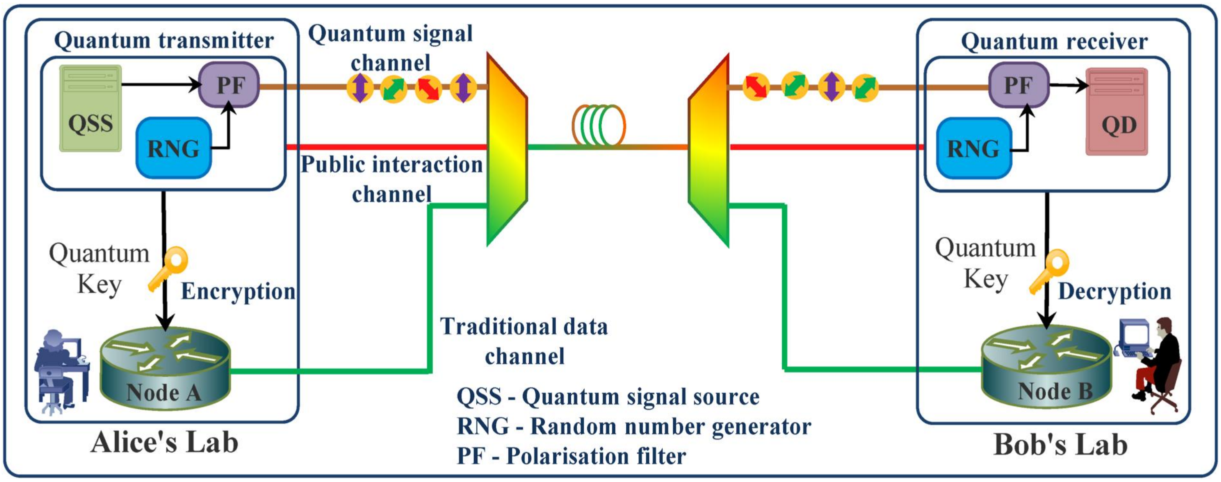


FIGURE 1 An illustration of a basic quantum key distribution-secured optical networks with three channels [5, 18, 19].

repeater nodes can be bypassed to reduce the wastage of quantum key resources [24]. In [30], for effective distribution of secret keys for multicast services (where data is transmitted from single source node to multiple destination nodes), a new distributed subkey-relay tree-based secure multicast-routing and key assignment technique was developed. However, as the dynamicity increases, the performance of the existing strategies becomes inefficient as these rely on fixed policies that focus on the immediate optimisation goals for the current network state and are unable to achieve the optimal solution to solve the dynamic RRA problem of QKD-ONs. Inspired by the recent advances in artificial intelligence/machine learning, which allow systems/machines to learn from historical data and make predictions to solve decision-making problems, we explored the capability of different options available in artificial intelligence/machine learning. Typically, machine learning requires data labelling to identify the raw data and add meaningful and informative labels that help the machine learning model to learn. In machine learning, the aim of training or validating the model with a labelled dataset is sometimes referred to as 'ground truth.' However, in this work, the problem of routing and resource assignment is a complex decision-making problem (explained in Section 1) and also the mapping of input and output variables or labelling of data is not feasible as it is based on the observation of QKD-ON's environment conditions.

Recently, reinforcement learning, one of the most important subfields of machine learning, has received extensive research attention as it is a feedback-based machine learning method. In reinforcement learning, an agent continuously interacts with the environment to make decisions, observe the results of decisions, and then automatically adjust its strategy based on the feedback of the previous decision to achieve the optimal/best policy. However, reinforcement learning, a learning process, is inapplicable and unsuitable for large and complex networks because it takes a lot of time to reach the optimal/best policy as it has to explore and learn about the whole system. As a result, reinforcement learning's applications are quite limited in practice. Deep learning has recently been introduced as a new ground-breaking method and has potential to overcome the limitations of reinforcement learning. Deep learning helps to design complex environments and extract important features, thereby reducing computation complexity. Deep learning is implemented using Neural Networks, thus opening a new era for improving the reinforcement learning algorithms' learning process. This combined form of reinforcement learning and deep neural networks (DNNs) is known as Deep Reinforcement Learning (DRL). DRL has ability to approximate optimal policy by employing DNNs for complex decision-making problems and improves the learning speed and performance of the reinforcement learning algorithms. As a result, the application of DRL [31, 32] has received intensive research interest in communication and networking to solve the complex decision-making problems [33, 34] and has become one of the most active areas of research in machine learning. However, in QKD-ONs, limited works have been reported using the application of DRL to address the RRA problem [35, 36].

The main contributions of this work are as follows:

(i) The DRL method is exploited to address the RRA problem in the quantum signal channel of QKD-ONs.
(ii) A RRA scheme based on DRL to select an optimal route and allocate suitable network resources during assignment and reassignment is proposed in this work.
(iii) The performance of the proposed DRL-based RRA scheme is compared with the deep-Q network (DQN) method and the two baseline schemes, namely, First Fit (FF) and Random Fit (RF).
(iv) Simulation results demonstrated that the proposed DRL-based RRA scheme outperforms the DQN and the two baseline schemes in terms of blocking probability (BP) and resource utilisation (RU) for both the considered networks.

The paper is structured as follows. Section 2 describes the system model along with the notation used. The concept of a proposed DRL-based RRA scheme, working principle of DRL framework for RRA in QKD-ONs, and modelling and training are presented in Section 3. The simulation results are discussed in Section 4. Section 5 concludes the paper.

2 | SYSTEM MODEL

The network topology of QKD-ON is modelled as $G(V_Q, E_Q, W_T, W_Q, K_T)$, where V_Q and E_Q are the sets of QKD-ON nodes and links, respectively. W_T and W_Q denote the total attainable wavelengths on each link and the number of wavelengths reserved for the quantum signal channel in the QKD-ONs, respectively. The total number of attainable time-slots on each quantum link is denoted by K_T. Moreover, the number of attainable time-slots, that is, K_T, is the same on each quantum link in the QKD-ONs for this work. A QLR is modelled as Q_t $(o_t, d_t, t_{arr}, t_{upd}, t_{dep}, T, Z^c_{t,k}, Z^m_{t,k})$, $Q_t \in Q$. o_t and d_t denote the source and destination node of a QLR, respectively. The arrival, update, and departure time of a QLR are represented by t_{arr}, t_{upd}, and t_{dep}, respectively. T is the secret key update period of a QLR. A set of total incoming QLRs over the QKD networks is represented by Q. Then, the number of specific QLRs (Q_s) can be determined by the expression as follows:

$$\left|Q_s\right| = \frac{\left|V_Q\right| * \left(\left|V_Q\right| - 1\right)}{2} \tag{1}$$

where $\left|V_Q\right|$ represents the total number of QKD-ON nodes in the network.

Let $Z^c_{t,k}$ and $Z^m_{t,k}$ denote the required number of secret key time slots for a QLR creation and modification, respectively. To establish Q_t, it is required to compute and select an end-to-end routing path $P_{ot,\ dt}$ from source-destination QKD-ON nodes and allocate network resources on each quantum link along the selected path $P_{ot,\ dt}$ using the proposed DRL-based RRA scheme. QLR is served if network resources are available on one of the pre-calculated paths ($K_{ot,\ dt}$) during assignment and reassignment, else QLR is blocked.

3 | PROPOSED DRL-BASED ROUTING AND RESOURCE ASSIGNMENT IN QKD-ON

3.1 | Proposed DRL-based RRA scheme

This subsection discusses the concept of a proposed DRL-based RRA scheme, where the objective is to maximise the number of QLRs, while reducing blocking and efficiently utilising network resources.

The proposed DRL-based RRA scheme jointly addresses the routing and resource assignment problem of the quantum signal channel. In this scheme, for routing, the DRL agent selects an optimal path based on the hop counts $H_t \in H$, where QLR will utilise less number of links, thereby resulting in more accommodation of QLRs in QKD-ONs. The resources on the selected path $P_{ot,\ dt}$ will be assigned using I candidate of the DRL-based RRA scheme. In this work, I candidate describes the process of assigning resources to QLR on the selected path $P_{ot,\ dt}$ during assignment and reassignment. The I candidate of DRL-based RRA scheme selects the closest available time slots to $Z^{c}_{t,k}$ during assignment (for QLR creation) and $Z^{m}_{t,k}$ reassignment (for QLR modification). The reason is that the selection of the closest available time slots reduces the wastage of network resources and enhances the possibilities of the available resources for the upcoming QLRs in the QKD-ONs, hence resulting in lower blocking of QLRs.

For ease of understanding, Figure 2 depicts the steps of I candidate of the proposed DRL-based RRA scheme during assignment. Consider a scenario in which a QLR AC arrives in the QKD-ON from the source node A to the destination node C, requires two time slots for assignment, and selects the best path A–B–C out of the pre-calculated K paths (Assume $K = 3$, then $K_{ot,\ dt}$ for a QLR AC ($K_{A,\ C}$) are A–B–C, A–E–D–C, A–E–B–C), as shown in Figure 2. The I candidate of the proposed DRL-based RRA scheme first calculates the sizes and initial indices of all the available I time-slot blocks on the corresponding quantum links represented by $Z^{a}_{t,k,i}$ and $Z^{b}_{t,k,i}$, where i represents the number of available I time-slot blocks of the selected path A–B–C, respectively. As shown in Figure 2, the calculated sizes ($Z^{a}_{t,k,i}$ [represented with the green dashed box in Figure 2]) and initial indices ($Z^{b}_{t,k,i}$ [represented with the red box in Figure 2]) of all the available I time slot blocks are $Z^{a}_{t,k,1} = \{3\}$, $Z^{a}_{t,k,2} = \{2\}$, $Z^{a}_{t,k,3} = \{4\}$, and $Z^{b}_{t,k,1} = (0)$, $Z^{b}_{t,k,2} = (4)$, $Z^{b}_{t,k,3} = (10)$, respectively. Based on the above mentioned criterion of the resource assignment and reassignment, the I candidate finds one of the closest I time-slot blocks represented by $Z_{closest}$, and assigns it to the QLR AC on the selected path A–B–C. In the above example, the $Z_{closest}$ on the basis of size is $Z^{a}_{t,k,2} = \{2\}$, and its initial index is $Z^{b}_{t,k,2} = (4)$ for the assignment of QLR AC (time slots filled with yellow colour) as shown in Figure 2. During reassignment, similar steps of I candidate of the proposed DRL-based RRA have been followed to improve the security level of QLR in QKD-ONs.

3.2 | Working principle of DRL framework for RRA in QKD-ON

This subsection describes the working principle of DRL framework to address the RRA problem of QKD-ONs.

Figure 3 illustrates the working principle of DRL-based RRA scheme for QKD-ONs. When the software-defined network (SDN) controller receives a QLR Q_t, it fetches the state representation, which includes the in-service QLRs, network topology, and resource utilisation information. The fetched information is fed into the DRL through a feature engineering module (*represented with a dashed red line in*

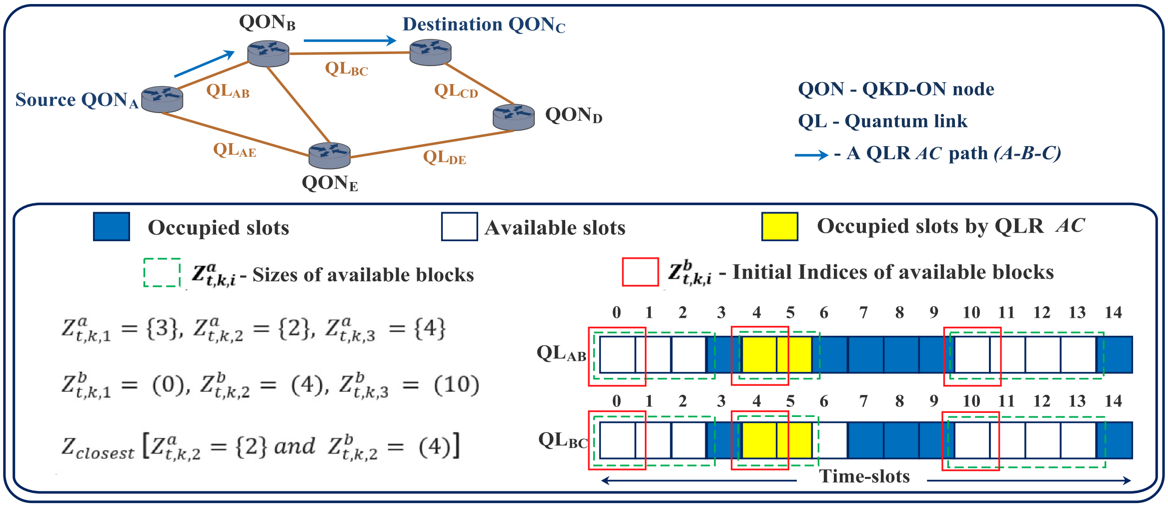

QON_B
Destination QON_C
Source QON_A
QL_AB
QL_BC
QL_CD
QON_D
QL_AE
QL_DE
QON_E
QON - QKD-ON node
QL - Quantum link
- A QLR AC path (A-B-C)
Occupied slots
Available slots
Occupied slots by QLR AC
$Z^{a}_{t,k,i}$ - Sizes of available blocks
$Z^{b}_{t,k,i}$ - Initial Indices of available blocks
$Z^{a}_{t,k,1} = \{3\}$, $Z^{a}_{t,k,2} = \{2\}$, $Z^{a}_{t,k,3} = \{4\}$
$Z^{b}_{t,k,1} = (0)$, $Z^{b}_{t,k,2} = (4)$, $Z^{b}_{t,k,3} = (10)$
$Z_{closest}\left[Z^{a}_{t,k,2} = \{2\}\ and\ Z^{b}_{t,k,2} = (4)\right]$
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
QL_AB
QL_BC
Time-slots

FIGURE 2 An illustration of I candidate of the proposed DRL-based RRA. DRL, deep reinforcement learning; RRA, routing and resource assignment.

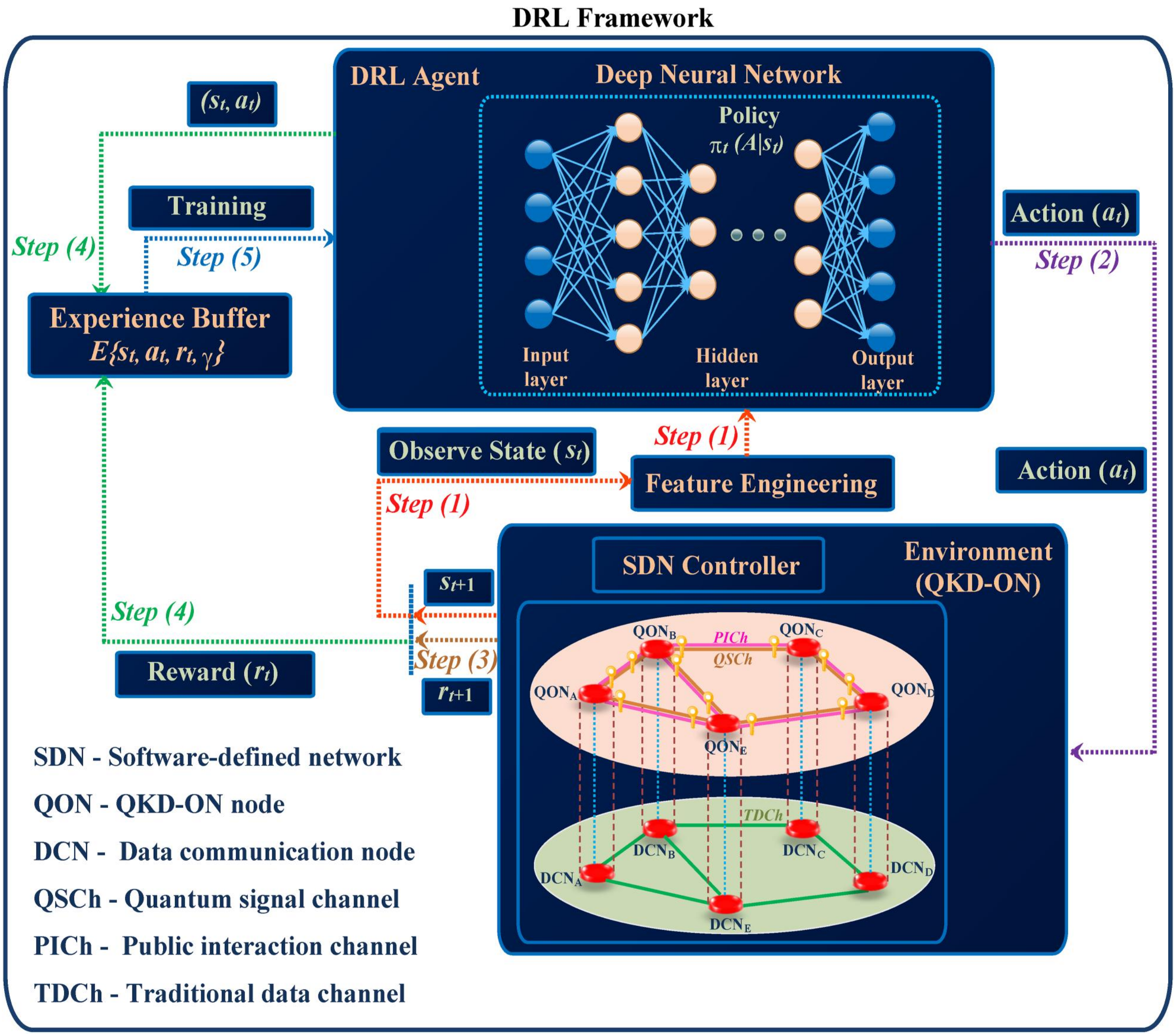


FIGURE 3 An illustration of the proposed DRL framework for RRA in QKD-ONs. DRL, deep reinforcement learning; QKD-ONs, quantum key distribution-secured optical networks; RRA, routing and resource assignment.

Figure 3), which generates the customised state s_t, where s_t represents the environment's state (*Step 1*). The DNN of DLR-based RRA reads the generated customised state data and takes action $a_t \in A$, where A is a definite action space. The DRL agent takes action according to the RRA policy π_t $(A\,|\,s_t, \theta)$, where A in this case defines a set of routing and resource assignment schemes for Q_t (where resources are assigned using the steps of I candidate of DRL-based RRA scheme (discussed in Section 3 [3.1]) and a set of DNNs' parameters is represented by θ. A policy π_t of RRA scheme generates a probability distribution over A. The SDN controller selects an action $a_t \in A$ based on the probability distribution and sets up the corresponding Q_t (*represented with a dashed purple line in* Figure 3) (*Step 2*). A reward system produces an immediate reward r_t for DRL-based RRA by receiving feedback from the previous RRA operation (*represented with a dashed brown line in* Figure 3) (*Step 3*). An experience E and a tuple $\{s_t, a_t, r_t, \gamma\}$ are stored in an experience buffer (*a dashed green line represents this action in* Figure 3) (*Step 4*), which will be used for training the DNN (*represented with a dashed blue line in* Figure 3) (*Step 5*) in order to achieve the optimal policy, where $\gamma \in [0, 1]$ is a discount factor. The objective of the DRL framework for RRA is to maximise the discounted cumulative reward G_t, which is defined as follows:

$$G_t = \sum_{j=0}^{\infty} \gamma^j \cdot \pi\left(a|s_{t+j}\right) \quad R\left(s_{t+j}, a\right) \tag{2}$$

3.3 | Modelling and training

This section first introduces DRL-based RRA modelling, which includes definitions of state representation, action, and reward, and then explains the training mechanism.

3.3.1 | Modelling

State

The state s_t is defined as a vector, expressed in Equation (3). It contains information of Q_t and the current network resource utilisation state, as well as the key feature of the I candidate of DRL-based RRA to provision Q_t during assignment and reassignment. When the number of possible pre-calculated paths ($K_{ot,\ dt}$) between o_t and d_t is smaller than K, we assign $\{\{Z^a_{t,k,i}, Z^b_{t,k,i}\}\,|_{\,i\,\in\,[1,\ I]}, Z^c_{t,k}, Z^m_{t,k}, Z^{Total}_{t,k}\,|_{\,k\,\in\,[1,\ K]}\}$ as an array of -1 ($\forall k > K_{ot,\ dt}$) in order to maintain the state's s_t format consistent.

$$s_t\{o_t, d_t, \left\{\left\{Z^a_{t,k,i}, Z^b_{t,k,i}\right\}|_{i\in[1,I]}, Z^c_{t,k}, Z^m_{t,k}, Z^{Total}_{t,k}|_{k\in[1,K]}\right\} \quad (3)$$

Action

For each Q_t to be served, the DRL agent selects a routing path from the pre-calculated candidate paths (K) and performs resource assignment and reassignment according to the I candidate of DRL-based RRA on the selected path. Therefore, the action space A includes $K.I$ actions.

Reward

The designed reward function $R(s_t, a_t)$ depends on two factors, that is, successful provision of Q_t and hop counts of selected path $P_{ot,\ dt}$. DRL-based RRA receives an immediate positive reward $r_t = +X$ on successful provisioning of Q_t, otherwise $r_t = -X$. Additionally, more positive rewards will be received if the agent selects a path having less hop counts.

3.3.2 | Training

This work uses the DRL algorithm, namely, proximal policy optimisation (PPO) [37], for training RRA. A DQN method [38] is also utilised to compare the proposed DRL-based RRA scheme based on PPO. DQN algorithms employ Q-learning to determine the best action to be taken in a given state and a deep neural network to estimate the Q-value function. PPO is a first-order policy gradient optimisation algorithm, which ensures that the policy update is not too large. A large step in a policy update results in learning a bad policy and may lead to instability during training.

In general, DRL environments are modelled as Markov Decision Process attributed to their finite state transitions [31]. However, DRL-based RRA comprehends infinite possible state transitions based on incoming Q. Hence, it is difficult to model it as the Markov Decision Process. Therefore, in this work, an experience buffer has been created of length N, where the experience $E\{s_t, a_t, r_t, \gamma\}$ generated after provisioning of each Q_t will be stored. The proposed DRL-based RRA has the following steps: (1) During initialisation, the experience buffer is cleared, (2) the QKD-ON state updates by releasing the network resources of the expired Q_t and the previous Q_t for reassignment, (3) the state model of Q_t modelled as s_t defined in Section 3 (3.3.1), (4) when the experience buffer is filled with N samples, DRL-RRA invokes training based on the working principle of the DRL framework for RRA as discussed in Section 3 (3.2). During the training process, a ϵ-greedy strategy has been employed for exploration and exploitation, and (5) at the end, discounted cumulative reward G_t has been calculated using Equation (2), and the buffer will be emptied.

4 | RESULTS AND DISCUSSION

4.1 | Simulation setup

In this work, two popularly used different sizes of network, namely 14-node The National Science Foundation Network (NSFNET) and 24-node UBN24 [39], are used to evaluate the performance of DRL-based RRA in comparison with the DQN method and the baseline schemes, namely FF and RF. A short-distance QKD network, where the maximum distance between the two end nodes is less than the distance that can accomplish the point-to-point QKD mechanism is assumed. In this work, 80 wavelengths with 50 GHz channel spacing for three types of channels, namely, traditional data channel, quantum signal channel, and public interaction channel in QKD-ONs, are considered. Same number of wavelengths are reserved for both the quantum signal channel and the public interaction channel [19]. Therefore, this work analyses the performance of DRL-based RRA, DQN, and the baseline schemes only for the quantum signal channel. In this work, a dynamic RRA problem is considered in which QLRs are randomly generated according to Poisson distribution between source and destination nodes following uniform traffic distribution.

During the training, the hyper-parameters γ (determines how much DRL agents care about rewards, $\gamma \in [0, 1]$) and the learning rate (controls how quickly the model is adapted to the RRA problem) are set to 0.95 and 10^{-3}, respectively. DNNs used ReLU as an activation function in the hidden layers because it allows models to learn faster and perform better. Adam algorithm [40] is used as an optimiser because of its fast computation time and simple parameter tuning and is considered as a default optimiser for most applications. The simulations of the proposed DRL-based RRA, DQN, and baseline schemes are performed with a customised Python-based simulator. This simulator uses NetworkX to design the graph representation of the network model and Pytorch-based preferred RL library for the DRL algorithms. The implementation of QKD-ON system model, QLR model, and simulations is performed using Python.

4.2 | Evaluation results

4.2.1 | Training

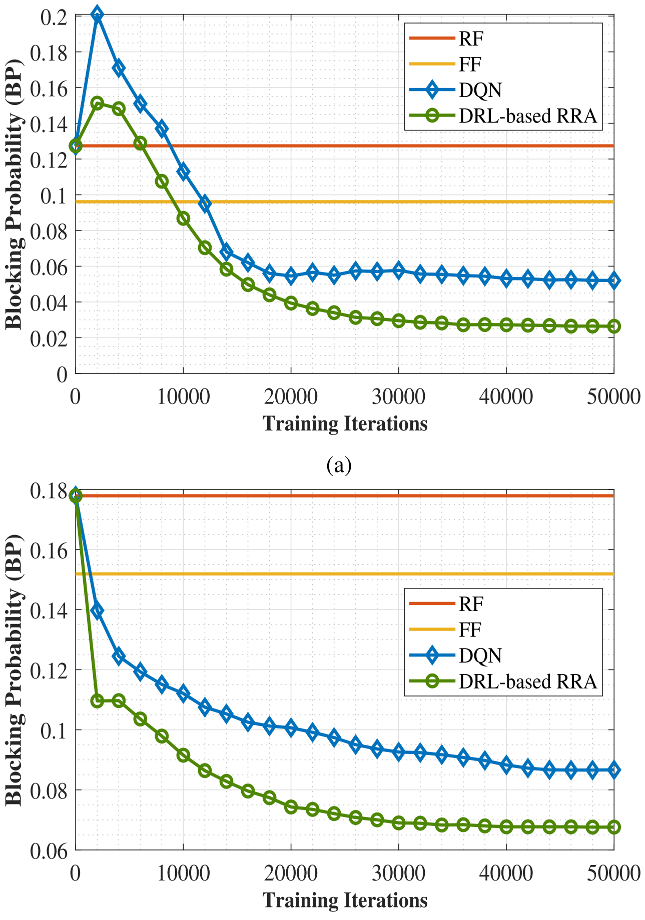
The training results of blocking probability (BP) and average reward (AR) versus training iterations for the proposed DRL-based RRA, DQN, FF, and RF for both the considered networks are illustrated in Figures 4 and 5, respectively.

BP is defined as the ratio of the total rejected QLRs to the total QLRs in the QKD-ONs. *AR* is the average of total rewards after each training iteration. At initial training iteration, the *BP* of the proposed DRL-based RRA and DQN is equal or higher than the two baseline schemes represented with the straight line, and the *AR* is minimum as shown in Figures 4 and 5, respectively. The performance metrics, namely, *BP* and *AR*, of DRL-based RRA and DQN improve with the number of training iterations during training. However, *BP* and *AR* of DRL-based RRA using PPO are much better than the DQN method because PPO has smaller policy updates, which help to obtain an optimal solution with better training stability, as shown in Figures 4 and 5, respectively.

The *BP* of the proposed DRL-based RRA surpassed RF and FF after 6000th and 8000th training iterations, respectively, whereas the *BP* of DQN surpassed RF and FF after 8000th and 10,000th training iterations for NSFNET, respectively. Similarly, for UBN24, the *BP* of the proposed DRL-based RRA and DQN surpassed RF and FF at 2000th training iterations. The *BP* of DRL-based RRA and DQN reaches its minimum value, and the training performance becomes stable after 40,000th and 42,000th training iterations for NSFNET and 38,000th and 42,000th training iterations for UBN24, respectively. The size of a network topology can have a significant impact on network performance as large size networks tend to be more complex. However, it has been observed that the proposed DRL-based RRA performs better than the DQN method and the two baseline schemes for the considered networks of different sizes and connectivity.

4.2.2 | Blocking probability

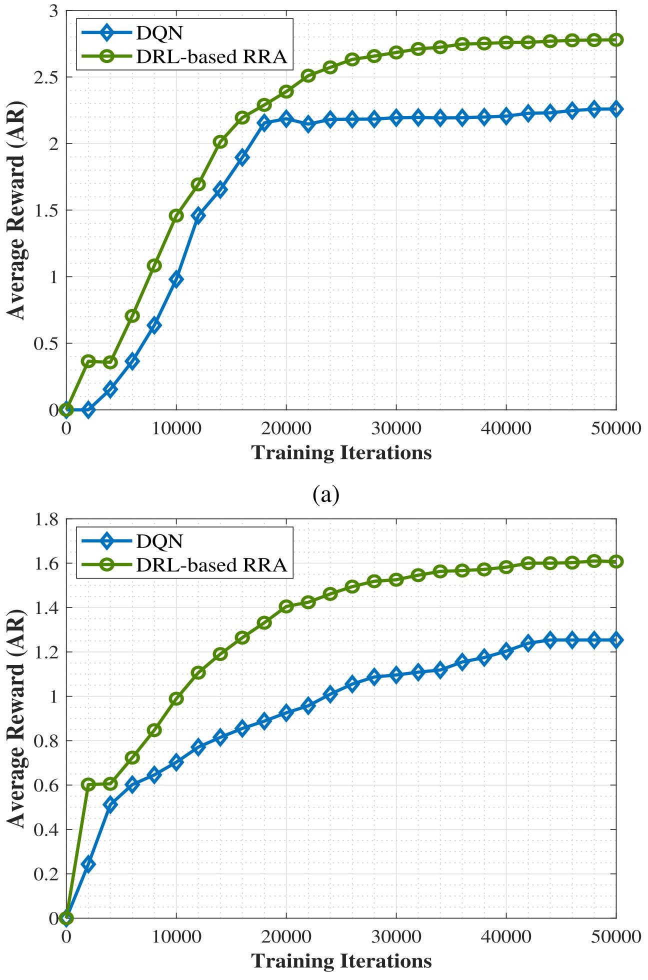
Figure 6 illustrates the performance of the DRL-based RRA compared to the DQN method and the two baseline schemes, namely, FF and RF, for NSFNET and UBN24 in terms of *BP* under different average arrival rates of traffic. The average arrival rate is the mean number of arrivals of QLRs per unit time. It can be observed from Figure 6 that the *BP* of QLRs increases with the rise in the traffic arrival rate for the



(a)

(b)

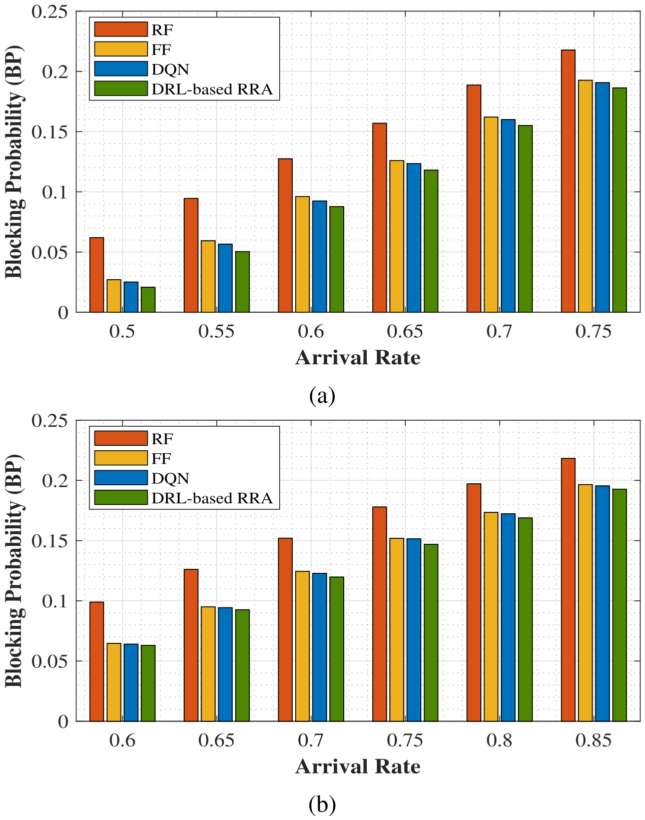
FIGURE 4 Training results of *BP* versus training iterations for (a) the National Science Foundation Network and (b) the UBN24.



(a)

(b)

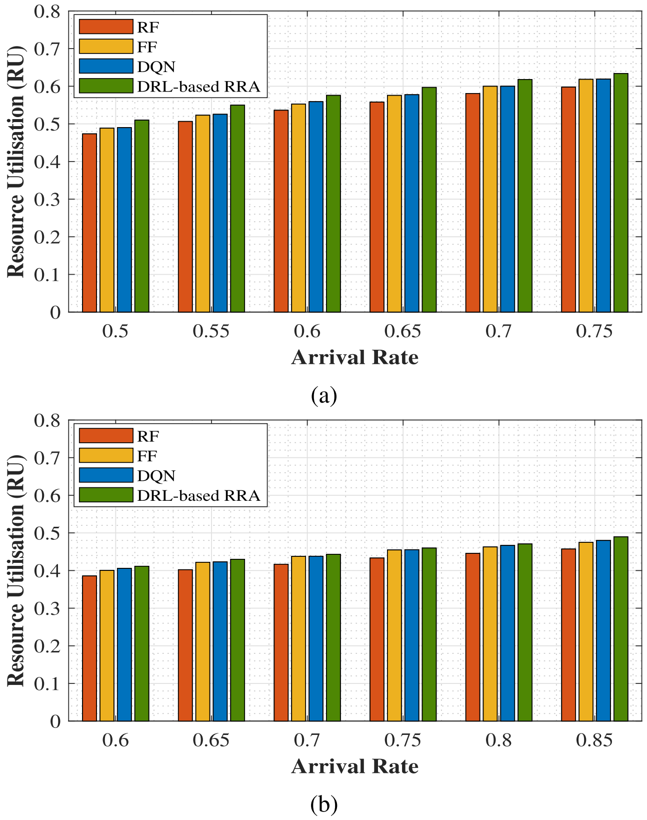
FIGURE 5 Training results of *AR* versus training iterations for (a) the National Science Foundation Network and (b) the UBN24.



(a)

(b)

FIGURE 6 Test results of *BP* versus average traffic arrival rate for (a) the National Science Foundation Network and (b) the UBN24.



(a)

(b)

FIGURE 7 Test results of *RU* versus average traffic arrival rate for (a) the National Science Foundation Network and (b) the UBN24.

proposed DRL-based RRA as well as for the DQN and the two baseline schemes because more resources get occupied during assignment and reassignment in the QKD-ONs. With the approximation capability of DNN, the DRL agent is able to build an optimal policy in order to minimise the blocking of QLRs, and the proposed scheme outperforms the DQN method and the two baseline schemes in terms of *BP*. Compared to the DQN, FF, and RF, the proposed DRL-based RRA achieves an average reduction in *BP* of 7.19%, 10.11%, and 33.50% for NSFNET, shown in Figures 6a, and 2.47%, 3.20%, and 19.60% for UBN24, shown in Figure 6b, respectively.

4.2.3 | Resource utilisation

RU for different average traffic arrival rate is an important metric and shown in Figure 7a,b for NSFNET and UBN24, respectively. *RU* is defined as the ratio of resources (time slots) utilised by the QLRs to the total resources available in QKD-ONs. It can be seen from Figure 7 that the *RU* for proposed DRL-based RRA, DQN, FF, and RF schemes increases with the increase in the arrival rate of traffic due to the accommodation of more QLRs in the QKD-ONs. The proposed DRL-based RRA achieved an average improvement in *RU* of 3.40%, 4.33%, and 7.18% for NSFNET and 1.34%, 1.96%, and 6.44% for UBN24 because of the acceptance of more QLRs (reduction in *BP* at each corresponding arrival rate, shown in Figure 6) compared to the DQN, FF, and RF, respectively.

5 | CONCLUSION

This work addressed the routing and resource assignment problem in the quantum signal channel of QKD-ONs by exploiting the deep reinforcement learning (DRL) technique. A deep reinforcement learning-based routing and resource assignment (DRL-based RRA) scheme using proximal policy optimisation to select an optimal route and efficient utilisation of network resources to satisfy the resource requirements of QLRs in the quantum signal channel of QKD-ONs is proposed. The simulation results indicate that the proposed DRL-based RRA scheme considerably outperforms the deep-Q network and the two baseline schemes, namely first-fit and random-fit, for the considered networks in terms of both blocking probability and resource utilisation. In future, methods to address various networking challenges of QKD-ONs based on DRL can be developed.

AUTHOR CONTRIBUTIONS

Purva Sharma: Conceptualisation; data curation; formal analysis; methodology; software; visualisation; writing—original draft preparation. **Shubham Gupta:** Conceptualisation;

data curation; formal analysis; methodology; software; visualisation. **Vimal Bhatia:** Conceptualisation; formal analysis; investigation; supervision; project administration; funding acquisition; validation; writing—review and editing. **Shashi Prakash:** Conceptualisation; formal analysis; investigation; supervision; project administration; validation; writing—review and editing.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education (MoE), Government of India (GoI) and Indo-U.S. Science and Technology Forum (IUSSTF) Project (ID: IUSSTF/JC-089/2019).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Purva Sharma https://orcid.org/0000-0001-7945-7289

REFERENCES

1. Debnath, S., et al.: Demonstration of a small programmable quantum computer with atomic qubits. Nature 536(7614), 63–66 (2016). https://doi.org/10.1038/nature18648
2. Cuomo, D., Caleffi, M., Cacciapuoti, A.S.: Towards a distributed quantum computing ecosystem. IET Quan. Commun 1(1), 3–8 (2020). https://doi.org/10.1049/iet-qtc.2020.0002
3. Yin, H.L., et al.: Experimental quantum secure network with digital signatures and encryption. Natl. Sci. Rev. 10(4), 1–11 (2022). https://doi.org/10.1093/nsr/nwac228
4. Lo, H., Marcos, C., Tamak, K.: Secure quantum key distribution. Nat. Photon. 8(8), 595–604 (2014). https://doi.org/10.1038/nphoton.2014.149
5. Sharma, P., et al.: Quantum key distribution secured optical networks: a survey. IEEE Open J. Commun. Soc. 2, 2049–2083 (2021). https://doi.org/10.1109/ojcoms.2021.3106659
6. Pirandola, S.: General upper bound for conferencing keys in arbitrary quantum networks. IET Quan. Commun. 1(1), 22–25 (2020). https://doi.org/10.1049/iet-qtc.2020.0006
7. Wu, J., Long, G.L., Hayashi, M.: Quantum secure direct communication with private dense coding using a general preshared quantum state. Phys. Rev. App 17(6), 064011 (2022). https://doi.org/10.1103/physrevapplied.17.064011
8. Liu, Z.P., et al.: Automated machine learning for secure key rate in discrete-modulated continuous-variable quantum key distribution. Opt Express 30(9), 15024–15036 (2022). https://doi.org/10.1364/oe.455762
9. Zhou, M.G., et al.: Neural network-based prediction of the secret-key rate of quantum key distribution. Sci. Rep. 12(1), 8879 (2022). https://doi.org/10.1038/s41598-022-12647-x
10. Heisenberg, W.: The Physical Principles of the Quantum Theory. Courier Corporation (1949)
11. Wootters, W.K., Zurek, Z.H.: A single quantum cannot be cloned. Nature 299(5886), 802–803 (1982). https://doi.org/10.1038/299802a0
12. Liu, R., et al.: Towards the industrialisation of quantum key distribution in communication networks: a short survey. IET Quan. Commun 3(3), 151–163 (2022). https://doi.org/10.1049/qtc2.12044
13. Bennett, C.H., Brassard, G.: Quantum cryptography: public key distribution and coin tossing. In: Proc. IEEE Int. Conf. Computers, Systems & Signal Processing, pp. 175–179. Bangalore (1984)
14. Sheng, Y.B., Zhou, L., Long, G.L.: One-step quantum secure direct communication. Sci. Bull. 67(4), 367–374 (2022). https://doi.org/10.1016/j.scib.2021.11.002
15. Ekert, A.K.: Quantum cryptography based on Bell's theorem. Phys. Rev. Lett. 67(6), 661–663 (1991). https://doi.org/10.1103/physrevlett.67.661
16. Lo, H.K., Curty, M., Qi, B.: Measurement-device-independent quantum key distribution. Phys. Rev. Lett. 108(13), 1–5 (2012). https://doi.org/10.1103/physrevlett.108.130503
17. Xie, Y.M., et al.: Breaking the rate-loss bound of quantum key distribution with asynchronous two-photon interference. PRX Quan. 3(2), 020315 (2022). https://doi.org/10.1103/prxquantum.3.020315
18. Zhao, Y., et al.: Resource allocation in optical networks secured by quantum key distribution. IEEE Commun. Mag. 56(8), 130–137 (2018). https://doi.org/10.1109/mcom.2018.1700656
19. Cao, Y., et al.: Resource assignment strategy in optical networks integrated with quantum key distribution. Ieee/osa J. Opt. Commun. Netw. 9(11), 995–1004 (2017). https://doi.org/10.1364/jocn.9.000995
20. Inoue, K.: Quantum key distribution technologies. IEEE J. Sel Top Quan. Electron 12(4), 888–896 (2006). https://doi.org/10.1109/jstqe.2006.876606
21. Bahrami, A., Lord, A., Spiller, T.: Quantum key distribution integration with optical dense wavelength division multiplexing: a review. IET Quan. Commun 1(1), 9–15 (2020). https://doi.org/10.1049/iet-qtc.2019.0005
22. Bahrani, S., Razavi, M., Salehi, J.A.: Optimal wavelength allocation in hybrid quantum-classical networks. In: Proc. IEEE EUSIPCO, pp. 483–487. Budapest (2016)
23. Peters, N., et al.: Dense wavelength multiplexing of 1550 nm QKD with strong classical channels in reconfigurable networking environments. New J. Phys. 11(4), 1–17 (2009). https://doi.org/10.1088/1367-2630/11/4/045012
24. Dong, K., et al.: Auxiliary graph based routing, wavelength, and time-slot assignment in metro quantum optical networks with a novel node structure. Opt Express 28(5), 5936–5952 (2020). https://doi.org/10.1364/oe.380329
25. Hua, W., et al.: Resilient quantum key distribution (QKD)-integrated optical networks with secret-key recovery strategy. IEEE Access 7, 60079–60090 (2019). https://doi.org/10.1109/access.2019.2915378
26. Li, X., et al.: Key-recycling strategies in quantum-key-distribution networks. Appl. Sci. 10(11), 3734 (2020). https://doi.org/10.3390/app10113734
27. Dong, K., et al.: Tree-topology-based quantum-key-relay strategy for secure multicast services. Ieee/osa J. Opt. Commun. Netw. 12(5), 120–132 (2020). https://doi.org/10.1364/jocn.385554
28. Cao, Y., et al.: Key on demand (KoD) for software-defined optical networks secured by quantum key distribution (QKD). Opt Express 25(22), 26453–26467 (2017). https://doi.org/10.1364/oe.25.026453
29. Cao, Y., et al.: Time-scheduled quantum key distribution (QKD) over WDM networks. Ieee/osa J. Lightw Technol. 36(16), 3382–3395 (2018). https://doi.org/10.1109/jlt.2018.2834949
30. Dong, K., et al.: Distributed subkey-relay-tree-based secure multicast scheme in quantum data center networks. Opt. Eng. 59(6), 065102 (2020). https://doi.org/10.1117/1.oe.59.6.065102
31. Sutton, R.S., et al.: Reinforcement Learning: An Introduction. MIT press (2018)
32. Arulkumaran, K., et al.: Deep reinforcement learning: a brief survey. IEEE Signal Process. Mag. 34(6), 26–38 (2017). https://doi.org/10.1109/msp.2017.2743240
33. Luong, N.C., et al.: Applications of deep reinforcement learning in communications and networking: a survey. IEEE Commun. Surv. Tutor 21(4), 3133–3174 (2019). https://doi.org/10.1109/comst.2019.2916583
34. Correa, I., et al.: Simultaneous beam selection and users scheduling evaluation in a virtual world with reinforcement learning. ITU J-FET 3(2), 202–213 (2022). https://doi.org/10.52953/chuz8770
35. Zuo, Y., et al.: Reinforcement learning-based resource allocation in quantum key distribution networks. In: ACP, pp. T3C–6. Beijing (2020)

36. Cao, Y., et al.: Multi-tenant provisioning for quantum key distribution networks with heuristics and reinforcement learning: a comparative study. IEEE Trans. Netw. Serv. Manag 17(2), 946–957 (2020). https://doi.org/10.1109/tnsm.2020.2964003
37. Schulman, J., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:170706347, 1–12 (2017)
38. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature 518(7540), 529–533 (2015). https://doi.org/10.1038/nature14236
39. Agrawal, A., Bhatia, V., Prakash, S.: Spectrum efficient distance-adaptive paths for fixed and fixed-alternate routing in elastic optical networks. Opt. Fiber Technol. 40, 36–45 (2018). https://doi.org/10.1016/j.yofte.2017.11.001
40. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980, 1–15 (2014)

How to cite this article: Sharma, P., et al.: Deep reinforcement learning-based routing and resource assignment in quantum key distribution-secured optical networks. IET Quant. Comm. 4(3), 136–145 (2023). https://doi.org/10.1049/qtc2.12063