IEEE Transactions on Quantum Engineering

Received September 23, 2021; revised April 26, 2022; accepted May 9, 2022; date of publication May 20, 2022; date of current version June 22, 2022.

Digital Object Identifier 10.1109/TQE.2022.3176375

A Connection-Oriented Entanglement Distribution Design in Quantum Networks

JIAN LI[1] (Member, IEEE), QIDONG JIA[1], KAIPING XUE[1] (Senior Member, IEEE), DAVID S. L. WEI[2] (Senior Member, IEEE), AND NENGHAI YU[1] (Member, IEEE)

[1]School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, China
[2]Department of Computer and Information Science, Fordham University, Bronx, NY 10458 USA

Corresponding author: Kaiping Xue (e-mail: kpxue@ustc.edu.cn)

This work was supported in part by the Anhui Initiative in Quantum Information Technologies under Grant AHY150300, in part by China Postdoctoral Science Foundation Funded Project under Grant 2021M703118, and in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant Y202093.

ABSTRACT Quantum networks create a completely new way for communication, and the most important function of a quantum network is to generate long-distance quantum entanglement to serve a number of quantum applications. As the scale of the network expands, in order to establish end-to-end entanglement between two remote nodes, entangled pairs need to be generated and distributed among multiple repeaters along the path from the source to the destination, which requires a specific protocol to negotiate resource allocations and quantum operations among the repeaters. Thus, to realize such remote entanglement distribution in a quantum network, designing a stable and reliable protocol becomes urgent and necessary. In this article, we focus on how to guarantee the generation rate of entangled pairs between any two quantum nodes to meet the requirements of various quantum applications in a large-scale quantum network. We present a connection-oriented entanglement distribution protocol inspired by the connection-oriented communication model in classical networks. Our protocol is located in the network layer of the quantum network to enable end-to-end quantum communication. Three main features are provided by the proposed protocol: 1) it is reliable and can guarantee the successful entanglement generation; 2) it can reduce the influence of quantum decoherence via reducing the latency caused by resource competition; and 3) it can guarantee the rate of generating entanglement between two quantum nodes according to the requirement of quantum applications.

INDEX TERMS Connection-oriented protocol, entanglement distribution, quantum networks.

I. INTRODUCTION

In entanglement-based quantum networks, the most important and challenging task is to generate quantum entanglement between any two far apart quantum nodes [1], [2]. By generating and distributing entangled pairs (Bell pairs), any two end nodes in quantum networks can achieve several entanglement-based applications, e.g., distributed quantum computing [3] and entanglement-based quantum key distribution [4], [5]. One of the most well-known application exploiting entanglement is quantum teleportation [6], as shown in Fig. 1(a). Although the point-to-point entanglement distribution over short distance has been implemented experimentally [7], the photon loss still hinders the entanglement distribution between any two distant quantum nodes. To diminish the photon loss over long-distance transmission, entanglement swapping, which can connect multiple short-distance entangled pairs, as shown in Fig. 1(b), is considered as a reliable solution [8]–[10]. Thus, by exploring the implementation of entanglement swapping, it becomes possible for entanglement-based quantum networks to establish long-distance quantum entanglement.

To communicate between two end nodes through quantum teleportation, remote entangled pairs should be generated between the sender and the receiver. In the existing work, inspired by the design of protocol stack in classic networks, Dahlberg *et al.* [11], Kozlowski *et al.* [12], and Li *et al.* [13]

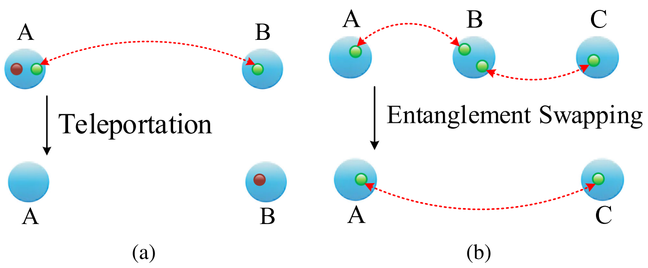


FIGURE 1. Illustration of teleportation and entanglement swapping. (a) Teleportation. (b) Entanglement swapping.

pioneered the design of quantum network protocol stack to facilitate the development of quantum networks, in which the network layer is responsible for establishing long-distance entanglement. However, since the development of quantum networks is still in its nascent stage, there is still no such specific protocol designed and developed for quantum networks. Thus, researchers from academia and practitioners from industry are highly expected to develop required protocols for each layer of quantum networks.

To design quantum network protocols, we need to consider not only the functions of quantum networks but also the physical characteristics of quantum networks. Tracing back to the classical networks, there are two different ways to transmit information: connection-oriented communication, such as multiprotocol label switching (MPLS) protocol in transmission control protocol/internet protocol (TCP/IP) stack [14], and connectionless communication, such as IP protocol in TCP/IP stack [15]. Although the connectionless protocol has become the standard and basic protocol in classical networks, it still has inherent imperfection. For example, in classical networks, the connectionless protocol can hardly provide quality-of-service (QoS) guarantee, e.g., latency, compared with connection-oriented protocol. However, in quantum networks, due to the characteristic of quantum decoherence, the information transmitted in the form of quantum bits, also called qubits, can be stored in quantum memory only for a short time considering the quantum decoherence [16]. Thus, providing a low latency of entanglement establishment becomes necessary and critical for quantum networks. On the other hand, the end-users and quantum applications may have various requirements of entanglement distribution rate. For example, due to the high requirement of secret key rate for one-time pad encryption, entanglement-based quantum key distribution [4], [5] may require relatively higher rate of entanglement than other quantum applications Thus, an eligible quantum network protocol design should provide QoS guarantee in terms of latency and entanglement distribution rate.

In this article, we propose a connection-oriented communication protocol for quantum networks to provide QoS guarantees in terms of entanglement distribution rate and end-to-end latency for quantum applications. Our proposed protocol is designed as a distributed one, which can work in arbitrary network topologies without the support of a central controller. The specific workflow of our protocol can be divided into three phases. The first phase is the link planning phase, in which a communication path is determined between two quantum end nodes and memory resources of intermediate quantum nodes are allocated for the path. By allocating the memory resources in advance, our protocol can guarantee entanglement distribution latency and generation rate when multiple flows exist simultaneously. The second phase is the entanglement establishing phase. In this phase, the routing nodes along the end-to-end path perform entanglement swapping operations to extend the entanglement distance, and entangled pairs are then generated between the two quantum end nodes. This process is repeated to realize the continuous distribution of entangled pairs between quantum end nodes to meet the requirements of the upper layer applications. The third phase is the connection releasing phase, in which the previously allocated memory resources are released so that they can be used by other communication paths.

We summarize the main contributions of this article as follows.

1) Taking the characteristics of quantum networks into account, we analyze the advantages and disadvantages of connectionless protocol and connection-oriented protocol in entanglement-based quantum networks and propose that the connection-oriented protocol is more suitable for entanglement-based quantum networks.
2) To promote the development and realistic application of quantum networks, we design a connection-oriented communication protocol. Through the design of specific signaling interaction process and reserving memory resources, the proposed protocol can reduce the latency caused by the resources competition and also provide QoS guarantees in terms of entanglement distribution rate and latency.
3) To verify the effectiveness and efficiency of the proposed protocol, a series of performance evaluations under different quantum network scenarios are conducted. Based on a purpose-built quantum network simulator, called NetSquid [17], we first develop a simulated quantum network and test the proposed protocol in the simulator. The results show that our connection-oriented communication protocol can guarantee QoS in terms of the rate and latency of generating entangled pairs.

The rest of this article is organized as follows. In Section II, we discuss the related work and the existing problems in current quantum network protocol design. In Section III, the difference between classical networks and quantum networks is analyzed, and the quantum network architecture and model are introduced in Section IV. After that, in Section V, we describe the details of our connection-oriented protocol design. Simulations to demonstrate the effectiveness of the proposed protocol and the results are discussed in Section VI. Finally, Section VII concludes this article.

II. RELATED WORK

To facilitate the development of quantum networks, some pioneering studies have been done for years. To name a few, in 2004, Lloyd *et al.* [18] first developed a prototype system for long-distance high-fidelity qubit teleportation, which can be foreseen to be required if future quantum computers are to be linked together into a quantum Internet. Based on the prototype system, Pirandola *et al.* [19], [20] further discussed the physical unit needed for building the quantum Internet and discussed the ultimate limits of point-to-point private communications.

Motivated by the proposed concept of quantum Internet and the corresponding architecture design, in recent years, it has become a trend that more and more studies focus on the specific scientific problem and protocol design, for example, routing algorithm design and link layer protocol design. Although the debate on the architecture of quantum network and the network protocol stack is ongoing and boiled, these existing studies still take an important step to boost the development of quantum networks. In summary, the existing work can be categorized into two types, i.e., the specific scheme/algorithm design for scientific problem in quantum networks, and the specific protocol design for the quantum network protocol stack.

For the former category, a typical research problem and also a critical problem is routing design, i.e., how to select the optimal route for generating end-to-end entanglement from the source node to the destination node. In 2013, Van Meter *et al.* [21] first discussed how to adapt the Dijkstra's algorithm for quantum repeater networks that generate entangled Bell pairs. A decentralized base-graph routing scheme is proposed in [22]. The proposed method allows an efficient routing to find the shortest paths in entangled quantum networks by using only local knowledge of the quantum nodes and the proposed scheme can be directly applied in practical quantum communications and quantum networking scenarios. After that, Gyongyosi and Imre [23] defined a method for routing space exploration and scalable routing in the quantum Internet and proved that scalable routing allows a compact and efficient routing in the entangled networks of the quantum Internet. Meanwhile, Gyongyosi [24] also proposed a mathematical model to quantify the dynamics of entangled network structures and entanglement flow in the quantum internet. By adopting a different mathematical model, Chakraborty *et al.* [25] constructed an efficient linear programming formulation that computes the maximum total achievable entanglement distribution rate, satisfying the end-to-end fidelity constraint in polynomial time. Considering the noisy quantum devices, Chakraborty *et al.* further proposed a routing algorithm for a quantum network such that each device can store a small number of qubits.

For the latter category, protocol designs for quantum repeater chain have received significant attentions [26]–[30]. However, these protocols cannot handle nonlinear topologies and do not have mechanisms for merging and splitting flows, which make these studies impractical especially for the scenario with arbitrary network topologies and multiple QoS-constraint requests. Thus, a universal protocol design has become more and more attractive in recent years. In 2019, Dahlberg *et al.* [11] proposed a novel quantum network stack architecture. In this architecture, the link layer is responsible for establishing entanglement between two adjacent quantum nodes, and the network layer is responsible for generating long-distance entanglement between two quantum nodes by entanglement swapping. To enable the practical system, Dahlberg *et al.* also proposed a link layer protocol, which generates entangled pairs between two adjacent nodes. However, the study still lacks the protocol design for the network layer, which aims to generate end-to-end entanglement. After that, a quantum data plane protocol of the network layer is proposed by Kozlowski *et al.* [12], but it does not perform any resource management, which will lead to quantum congestion collapse caused by the resources competition when multiple flows exist simultaneously and the quantum congestion collapse will result in an exceptionally low throughput.

Unlike the routing problem in quantum networks, which has been investigated in many studies [21], [31]–[33], the specific routing protocol design is still in the early stage. In 2017, Caleffi [34] modeled the entanglement generation through a stochastic framework and derived the closed-form expression of the end-to-end entanglement rate for an arbitrary path. Besides, this work designs a routing protocol and proves its optimality when used in conjunction with the entanglement rate as routing metric. After that, a "RuleSet"-based protocol is proposed in [35], but the authors only study two-node networks with a single link. Thus, the existing studies that attempt to design a specific routing protocol, which determines how to distribute end-to-end entanglement by negotiating resource allocations and quantum operations, still lack the applicability for large-scale quantum networks and ability for providing QoS-guaranteed service. In this case, a novel protocol design, which can provide efficient and QoS-guaranteed service, is urged to be designed.

III. MOTIVATION AND RESULT

A. DIFFERENCE BETWEEN QUANTUM NETWORKS AND CLASSICAL NETWORKS

In a classical network, information is encoded as electric signal or optical signal, called bit, whereas information is encoded as qubit in quantum networks. Qubits are fundamentally different from classical bits, which makes the difference between quantum networks and classical networks. The difference can be summarized as the following three aspects.

1) The first difference between classical networks and quantum networks is the way to transmit information between two nodes. The bits in classical networks can be transmitted through physical channel from one node to another node (the receiver) with the help of routers. In entanglement-based quantum networks, there are different ways to transmit qubits. When transmitting

qubits between two nodes that are close to each other or have a perfect channel between them, the qubits can be transmitted from one node to another through the channel directly. However, when transmitting qubits between two nodes that are far apart, as the decoherence eventually destroys information that cannot be replicated, the link entanglement[1] becomes a key resource for long-distance quantum communication. After the establishment of end-to-end entanglement, a qubit can be "teleported" from the source to the destination through the so-called quantum teleportation process. To realize quantum teleportation, a pair of parallel resources is needed. One of these resources is classical: two bits must be transmitted from the source to the destination. The other resource is quantum: an entangled pair of qubits must be generated and shared between the source and the destination [10]. Entangled pairs are the base of quantum teleportation, and it must guarantee that the entangled pairs have been generated between the sender and the receiver when they try to teleport quantum message.

2) The second difference is the way to communicate between two far apart nodes. In classical networks, the electric signal or optical signal can be copied and amplified by routers to extend transmissions. However, due to the quantum no-cloning theorem, unlike the electronic signals in classical networks, a qubit cannot be copied and amplified by a quantum repeater. To generate entangled pairs between two far apart quantum nodes, short-distance entangled pairs are generated first; then, entanglement swapping operations are performed. However, the entanglement swapping fails with a certain probability. If the entanglement swapping fails in one node, there is no point in continuing the entanglement swapping at the next node. Thus, the node failed in entanglement swapping should notify the sender to regenerate entanglement.
3) The third difference is that the binary bits in classical networks can be stored for a long time, whereas the qubits can just be stored for a short time in quantum networks due to the quantum decoherence [16]. Quantum decoherence is an important factor that impedes the development of quantum networks as it puts extremely stringent limits on how long qubits can be held in memory before they need to be used. In quantum networks, the qubit in the sender is stored in quantum memory until generating entanglement with the receiver successfully. Considering the influence of quantum decoherence, a network protocol should reduce the latency of generating entanglement between the sender and the receiver.

[1]In this article, Bell pairs shared between adjacent nodes are considered as link entanglement.

According to the characteristics of quantum networks, we propose that the quantum network protocol should possess the following functions:

1) reliable and stable procedure of generating end-to-end entanglement, which can deal with the failure of quantum operations;
2) low latency of generating end-to-end entanglement so as to reduce the storage time of qubits and the influence of quantum decoherence;
3) providing QoS-guaranteed service that meets the requirements of different quantum applications.

B. CONNECTION-ORIENTED PROTOCOL AND CONNECTIONLESS PROTOCOL IN ENTANGLEMENT-BASED QUANTUM NETWORKS

Owing to the different characteristics between quantum networks and classical networks, we cannot apply the protocol design principles adopted in classical networks to the quantum network protocol design directly. Nevertheless, the service models, e.g., connection-oriented versus connectionless, adopted by classical networks may still be useful in quantum protocol design. There are two kinds of communication protocols in the network layer of a classical network. One is connection-oriented communication protocol, such as MPLS protocol [14], and the other is connectionless communication protocol, such as IP protocol [15]. The connectionless communication protocol cannot guarantee that the messages arrive at the destinations successfully. Besides, a routing node may receive a large number of messages that exceed its capacity such that some messages suffer a high latency, while in the connection-oriented communication protocol, a virtual circuit is first established, and then, all the messages are transmitted to the destination along the virtual circuit. The connection-oriented communication protocol provides reliable message transmission and guarantees that each message arrives at the receiver successfully. Furthermore, by reserving resource for the circuit, the connection-oriented protocol can reduce the latency caused by resource competition. In quantum networks, in order to generate entanglement successfully and reduce the latency, a connection-oriented protocol is more suitable to generate entangled pairs between two quantum nodes. Furthermore, its ability to guarantee needed bandwidth enables the connection-oriented protocol to meet the QoS requirements of different quantum applications. Thus, we propose to develop a connection-oriented scheme for quantum networks.

C. ROUTINE OF CONNECTION-ORIENTED PROTOCOL IN ENTANGLEMENT-BASED QUANTUM NETWORKS

Our connection-oriented entanglement distribution protocol is divided into three phases. The first phase is the link planning phase, in which a communication path is determined between two quantum end nodes, and an appropriate amount of memory resources of the quantum nodes along the path are allocated for the path. By reserving sufficient resources, our

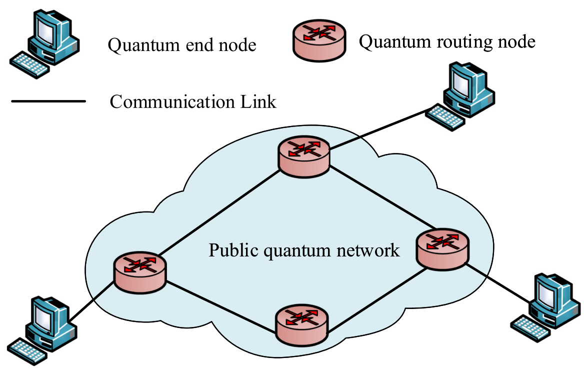


FIGURE 2. **Typical quantum network architecture.**

protocol can reduce the influence of quantum decoherence via reducing the latency caused by resource competition. The second phase is the entanglement establishing phase. In this phase, quantum routing nodes along the path perform entanglement swapping operations to extend the entanglement distance and entanglement are then generated between the two quantum end nodes. If the entanglement swapping fails, our protocol can notify the sender to regenerate entanglement. This process is repeated to realize the continuous distribution of entangled qubits between quantum end nodes to meet the requirements of the upper layer applications. The third phase is the connection releasing phase, in which the previously allocated memory resources are released so that they can be used by other communication paths.

IV. NETWORK ARCHITECTURE AND NETWORK MODEL

In this section, we first introduce the quantum network architecture, which concludes the definition of two different kinds of nodes and the function of different layers of a quantum network. After that, we describe the network model and define the speed of distributing entangled qubits between two adjacent nodes.

A. NETWORK ARCHITECTURE

In this part, we introduce the quantum network architecture (see Fig. 2), which is designed based on the existing physical experiments [15], [36]–[41]. According to different functions, we define two different kinds of network nodes, namely, *quantum end node* and *quantum routing node*. We assume that these two kinds of quantum nodes, i.e., quantum end nodes and quantum routing nodes, can transmit both classical messages and qubits if there is a communication link between them. Besides, we describe the functions located in different layers of a quantum network. The main components in our network architecture and the functions located in the link layer and the network layer of a quantum network are explained as follows.

1) QUANTUM ROUTING NODE

The quantum routing nodes are the devices equipped with necessary hardware to perform entanglement operations. They are designed to enable end-to-end entanglement and teleportation. A quantum routing node can process both classical messages and qubits and has two main functions. The first function is to select route for the establishment of end-to-end entanglement. In specific, arbitrary routing node works with other routing nodes to compute a suitable path that connects the source node and the destination node and store the path in a routing table, which is similar to the function of the routing nodes (i.e., router) in classical networks. The second function is performing entanglement swapping to generate end-to-end entanglement between source nodes and destination nodes. By utilizing a set of quantum routing nodes from the public quantum network, arbitrary two quantum end nodes can establish end-to-end entanglement through the public network and perform quantum teleportation.

2) QUANTUM END NODE

A quantum end node is similar to an end node in a classical network, and it can run quantum applications such as quantum key distribution [4], [5], quantum security directly communication [42]–[44], distributed quantum computing [3], [45], etc. Each quantum end node can process both qubits and classical messages. In our architecture, quantum end nodes are connected to the public quantum network by communication links. When a quantum end node wants to generate entanglement with another remote quantum end node, it should send a request to the public quantum network.

3) COMMUNICATION LINK

Communication link is responsible for transmitting classical messages and qubits. It includes a classical channel and a quantum channel. The classical channel is the medium, e.g., twisted pair and coaxial cable, to transmit classical messages, and the quantum channel is the medium, generally referring to optical fiber, that transmits qubits. If there is a link between two quantum nodes, the nodes can both transmit classical messages and qubits to each other. If there is a communication link between two quantum nodes, the nodes are called adjacent nodes.

4) LINK LAYER

The function of the link layer is to generate short-distance quantum entanglement between two adjacent nodes. A complete link layer protocol for quantum networks has been proposed in [11], in which the nodes can measure qubits directly, create and keep entangled pairs, prepare remote quantum states, and send qubits. As we focus on the communication protocol in the quantum network layer, we assume that the fidelity of the entangled pairs generated by the link layer is high enough for entanglement swapping. We only consider the rate of distributing entangled qubits with a high fidelity between two adjacent nodes, which affects the performance of network layer protocols.

5) NETWORK LAYER

The network layer aims at generating quantum entanglement between any two quantum end nodes. For two quantum nodes that are far apart, entanglement swapping will be performed to generate end-to-end entanglement between them. Entanglement swapping is a local joint measurement operation, i.e., a beam-splitter-based Bell state measurement (BSM) is performed on entangled pairs. After the operation of entanglement swapping, the final quantum state becomes one of the four Bell states with equal probability, and we can know the final state is which of the four Bell states through the result of the BSM. As practical devices used to perform quantum gate operation and quantum measurement operation are imperfect, these quantum operations may fail with probabilities [9]. In the network layer, we focus on the rate and latency of generating end-to-end entanglement between two quantum end nodes.

6) NODE CAPACITY AND CHANNEL CAPACITY

Each quantum node can have two types of qubits, i.e., memory qubits as a local memory and communication qubits with an optical interface, which can be entangled with a photon [10], [11]. When two quantum nodes generate entangled pair through the communication qubit, the communication qubit will be moved to a memory qubit, in order to free the communication qubit to produce the next entangled pair. Node capacity means the number of memory qubits; the more memory qubits the quantum node has, the larger the capacity of the quantum node is. The channel capacity means the rate of generating entangled pairs through communication qubits. In our architecture, each communication qubit can be regarded as a communication link, so the number of communication qubits in a quantum node equals the number of communication links connected to the quantum node.

B. NETWORK MODEL

In this part, we analyze the relationship between the entanglement generating rate of two adjacent nodes and the allocated memory resource and show that we can control the entanglement generating rate by allocating different memory resource. We denote the quantum network with the graph $G = (V, E, C)$, in which $V = \{v_i\}_{i=1}^{N}$, $E = \{e_{i,j}, v_i, v_j \in V\}$, and $C = \{c_i\}_{i=1}^{N}$ denote the set of nodes, communication links, and the available storage capacity of each node, respectively. For an arbitrary couple of nodes v_i and v_j, if there exist $e_{i,j} \in E$, then v_i and v_j are defined adjacent nodes. Furthermore, $d_{i,j}$ denotes the length of link $e_{i,j}$ between nodes v_i and v_j. We define that each node equips N quantum memory cells, and each cell can store one qubit for a certain amount of time. Thus, the maximal available storage capacity of one node is N.

An entanglement attempt to cross any one of the communication links $e_{i,j}$ succeeds with the probability $p(e_{i,j})$, which is proportional to $\eta(e_{i,j})$, where $\eta(e_{i,j}) = e^{-\alpha d_{i,j}}$ is the transmissivity of a lossy optical channel of length $d_{i,j}$ [32], [46], [47]. We denote that the rate of generating entangled pairs between two adjacent nodes is the same as R_0. Then, the rate of generating entangled qubits of communication link $e_{i,j}$ is $R_{e_{i,j}} = e^{-\alpha d_{i,j}} R_0$. Different from the bandwidth in classical networks, the available rate of each communication link is related not only to the link rate but also to the storage capacity currently available at the node. As the entangled qubits generated by adjacent quantum nodes are first stored in quantum memory before being consumed, the more storage capacity a communication path has, the higher rate of generating entangled qubits it has. Considering both the rate of generating entangled qubits of communication link and the storage capacity of the node, we denote the available rate of each communication link $e_{i,j}$ as $R'_{e_{i,j}} = \frac{\frac{1}{2}\cdot\min\{c_i, c_j\}}{N} \cdot R_{e_{i,j}} = \frac{\frac{1}{2}\cdot\min\{c_i, c_j\}}{N} \cdot e^{-\alpha d_{i,j}} \cdot R_0$, where c_i and c_j are the available storage capacities of node v_i and node v_j, respectively. In a communication connection, a quantum node has two communication links, and the coefficient $\frac{1}{2}$ means that each communication link can occupy half of available storage capacity.

V. CONNECTION-ORIENTED ENTANGLEMENT DISTRIBUTION PROTOCOL

The connection-oriented entanglement distribution protocol is divided into link planning phase, entanglement swapping phase, and link releasing phase. In link planning phase, a communication path is determined between two quantum end nodes, and a certain amount of memory space of the quantum nodes along the path is allocated to the path. By reserving sufficient resources, our protocol can reduce the latency caused by resource competition. In the entanglement swapping phase, the quantum routing nodes along the path perform the entanglement swapping operations to generate entanglement between two quantum end nodes. This process is performed repeatedly, and entangled qubits are generated between the two quantum end nodes until the communication process is over. For each established entanglement, the receiver sends an acknowledgment to the sender, which guarantees that the reliability of the communication and the memory resource allocated to the link planning phase are occupied until the communication finishes, which guarantees the stability of communication. In the link releasing phase, the nodes along the communication path release memory resources previously allocated to the path. The detailed description of each phase is given in what follows.

A. LINK PLANNING PHASE

The goal of the link planning phase is to find a suitable path between two quantum end nodes and inform quantum routing nodes along the path to reserve a certain amount of memory resource for the path. Finding a suitable path is an important issue in the quantum network, which has been studied in some studies [21], [32], [48]–[50]. In this article, instead of focusing on a specific routing algorithm, we assume that the routing table already exists. And our protocol aims to inform

TABLE 1. Structure of the Connection Table in Quantum Nodes

connection ID	left node	right node	memory resource
n	v_{prev}	v_{next}	C_n
...	...	...	...

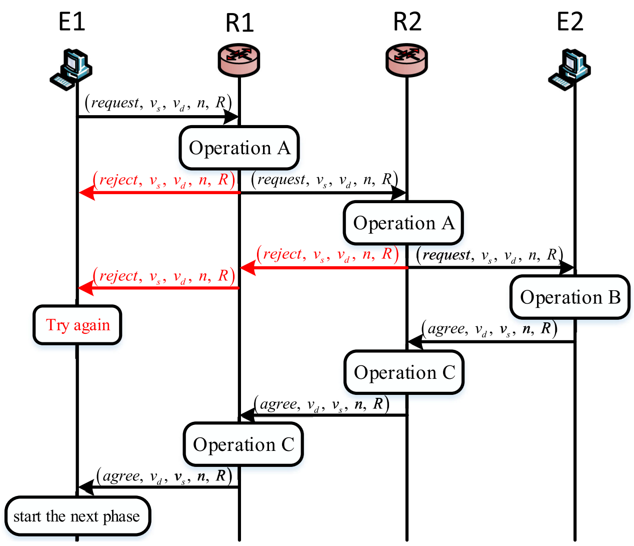


FIGURE 3. Procedure of the link planning phase.

quantum routing nodes to reserve appropriate memory space for each communication connection.

In this phase, we need to deal with how much memory resource should be allocated to a communication connection. We assume that the required rate of a communication path is R. According to the network model, for guaranteeing the rate of the previous communication link, the quantum routing node v_i needs to be allocated $C_{\text{prev}} = \frac{R}{R_0} \cdot e^{\alpha d_{\text{left},i}} \cdot N$ memory resource. Similarly, $C_{\text{next}} = \frac{R}{R_0} \cdot e^{\alpha d_{\text{next},i}} \cdot N$ memory resource is allocated to the next communication link. For each node v_i, the total memory resource allocated to the communication connection is $C = C_{\text{prev}} + C_{\text{next}} = \frac{R}{R_0} e^{\alpha d_{\text{prev},i}} N + \frac{R}{R_0} e^{\alpha d_{i,\text{next}}} N$, where C_{prev} and C_{next}, respectively, represent the number of storage units required for entanglement with the previous node and the number of storage units required for entanglement with the next node. The parameter N indicates the total number of storage units on the node. In this phase, the rate of generating entanglement is guaranteed by allocating appropriate amount of memory resource to the communication connection. Besides, a quantum routing node stores the status of memory resource allocations in the connection table, whose structure is given in Table 1.

The complete procedure of the link planning phase is shown in Fig. 3 and described as follows.

1) The sender sends a request message $(request, v_s, v_d, n, R)$ to the quantum network, where "*request*" is a flag that represents the request message, "v_s" is the source node, "v_d" is the destination node, and "R" is the rate of generating entangled qubits requested by the source node.
2) When quantum routing node v_i receives the request message from the previous quantum node v_{prev}, it performs the operation A, which includes three steps: First, it finds the next quantum routing node denoted as v_{next} by searching the routing table. Second, it allocates memory resource $C = C_{\text{prev}} + C_{\text{next}} = \frac{R}{R_0} \cdot e^{\alpha d_{\text{prev},i}} \cdot N + \frac{R}{R_0} \cdot e^{\alpha d_{i,\text{next}}} \cdot N$ to this communication connection. The C_{prev} memory resource is used to establish entanglement with the previous quantum node, and the C_{next} memory resource is used to establish entanglement with the next quantum node. Third, if there is not enough memory resource, node v_i sends a reject message $(reject, v_s, v_d, n, R)$ as acknowledgment to v_{prev}. Otherwise, quantum routing node v_i adds a term in the connection table to update the status of the local memory resource and transfers the request to the next quantum routing node v_{next}.
3) When the receiver receives the request message, it performs operation B, which includes two steps. First, it computes the memory resource that should be allocated for meeting the request $C = \frac{R}{R_0} \cdot e^{\alpha d_{\text{prev},i}} \cdot N$. Second, it judges whether the allocated memory resource C exceeds that it can provide. If not, the receiver adds the connection to the connection table, updates the local resources, and sends back an agree message along the original path. Otherwise, it replies a reject message along the original path.
4) When an intermediate node receives the reply message, it performs operation C. In this operation, if it receives an agree message, it transmits the message upward along the original path. However, if the intermediate node receives the reject message, it deletes the corresponding term from the connection table and update the local memory resource.
5) If the sender receives the agree message, it starts the entanglement swapping phase. Otherwise, if the sender receives the reject message, it tries to send a new request message or abandons this communication session.

B. ENTANGLEMENT SWAPPING PHASE

After the link planning phase, a communication path for generating entanglement between two quantum end nodes is determined. In the entanglement swapping phase, the routing nodes along the communication path will perform entanglement swapping operations to establish entanglement between the two quantum end nodes. To guarantee the reliability and the stability of the communication, the receiver replies by sending an acknowledgment to the sender for each established entanglement, and the memory resource allocated to the link planning phase will be occupied until the communication session finishes. The number of qubits consumed for per successful entanglement is related to the hop count in the path. Without considering the entanglement purification, each hop will consume one entangled qubit. As for the

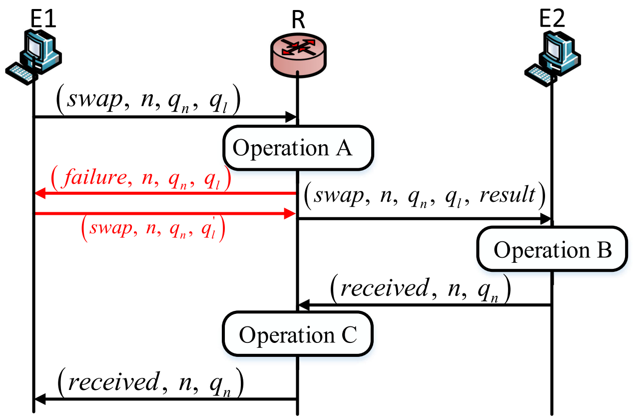


FIGURE 4. **Procedure of the entanglement swapping phase.**

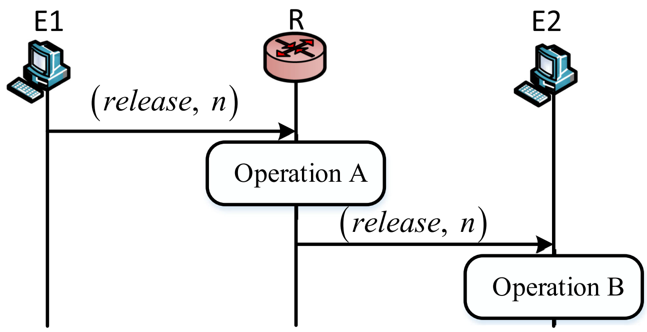


FIGURE 5. **Procedure of the link releasing phase.**

classical transmissions, in the entanglement swapping phase of our protocol, there are two classical transmissions for each successful entanglement. One is used for collecting the results of entanglement swapping in the quantum routing nodes and the other is for notifying the sender that the entangled pair has been generated successfully. However, in [11], three types of signaling are required, which include the FORWARD message, the TRACK message, and the COMPLETE message. If the length of the path is N, for generating one entangled pair, our protocol needs to transmit $2N$ messages, while $3N$ messages need to be transmitted in [11].

The procedure of entanglement swapping phase is shown in Fig. 4, and its details are given as follows.

1) The sender sends the swap message $(swap, n, q_n, q_l)$, which tells the quantum routing nodes on the communication path to perform the entanglement swapping operation. The message includes the flag of this message "*swap*," the connection number "n," the end-to-end entanglement identifier "q_n," and the link entanglement identifier "q_l."
2) When the first quantum routing node receives the "*swap*" message, it performs the operation A, which includes two steps. First, it performs entanglement swapping operation on the two specific qubits. One qubit is determined by the link entanglement identifier q_l and the other qubit is select randomly from the link entanglement between the node and the next node whose identifier is denoted as q_{l_r}. Second, if the entanglement swapping fails, the routing node sends the failure message $(swap, n, q_n)$ to the previous node according to the connection table. Otherwise, it adds the result of entanglement swapping and the identifier q_{l_r} into the start message and sends it to the next quantum node according to the connection table. In Fig. 4, we use q_l to represent q_{l_r} uniformly.
3) When the quantum end node receives the swap message, it performs the operation B, in which it performs some quantum gate operations according to the result in the message and replies with a received message including the connection number "n" and the end-end entanglement identifier "q_n" to indicate that the entanglement has been established.
4) When the quantum routing node receives the "*received*" message, it performs the operation C, in which the message is transmitted upward according to the connection table.
5) If the sender receives the "*received*" message, it starts to create another end-to-end entanglement until the communication finishes. If the sender receives the failure message, it attempts another link entanglement whose identifier is q_l' to re-establish this end-to-end entanglement.

C. LINK RELEASING PHASE

At the end of the communication, the resources allocated to this communication path during the link planning phase should be released. The link releasing phase is relatively simple. The nodes on the communication path just need to delete the corresponding entry in the connection table and release memory resources in turn. And the released memory resources can be used by other communications. The specific procedure of this phase is shown in Fig. 5, and the description is as follows.

1) The sender sends the release message $(release, n)$, which contains the connection number n, and deletes the term that has the connection number n from the connection table.
2) When the quantum routing node receives the release message, it performs the operation A, which includes two steps. First, it queries the connection table and transmit the message to the next node. Second, the quantum routing node updates the local memory resource and deletes the term that has the connection number n from the connection table.
3) When the receiver receives the release message, it performs the operation B, in which it updates the local memory resource and removes the term that has the connection number n from the connection table.

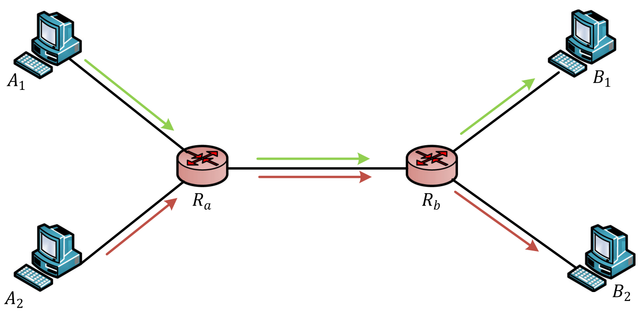


FIGURE 6. Evaluation topology. $R_a - R_b$ is a bottleneck link between the *A* nodes and the *B* nodes. Each communication link consists of a quantum channel and a classical channel.

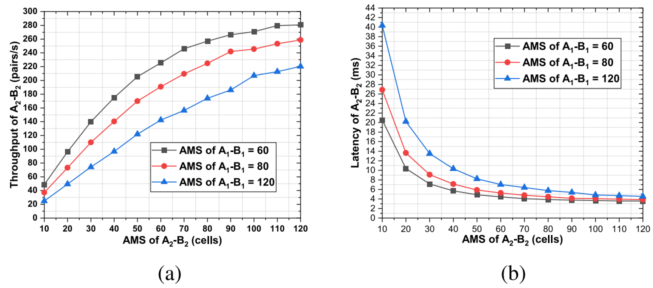


FIGURE 7. (a) Average throughput and (b) latency of communication $A_2 - B_2$ versus different allocated memory sizes with different background flows.

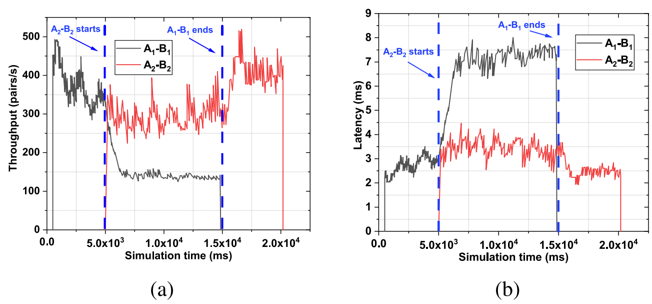


FIGURE 8. (a) Throughput and (b) latency of $A_1 - B_1$ and $A_2 - B_2$ changes over the simulation time.

VI. EVALUATION AND ANALYSIS

A. EVALUATION SETUP

For evaluating the performance of our protocol, we implement the proposed connection-oriented entanglement distribution protocol on a discrete-event simulator called NetSquid [17]. The simulator is responsible for the accurate representation of the physical hardware, including decoherence, propagation delay, optical fiber losses, quantum gate operations, and their time dependence. As our work is focused on the network layer [11] of quantum networks, we use a simple first-in first-out (FIFO) scheduling algorithm in the link layer. The link layer generates entangled pairs for each item in the connection table created in the link planning phase following the order of the creation time of the item.

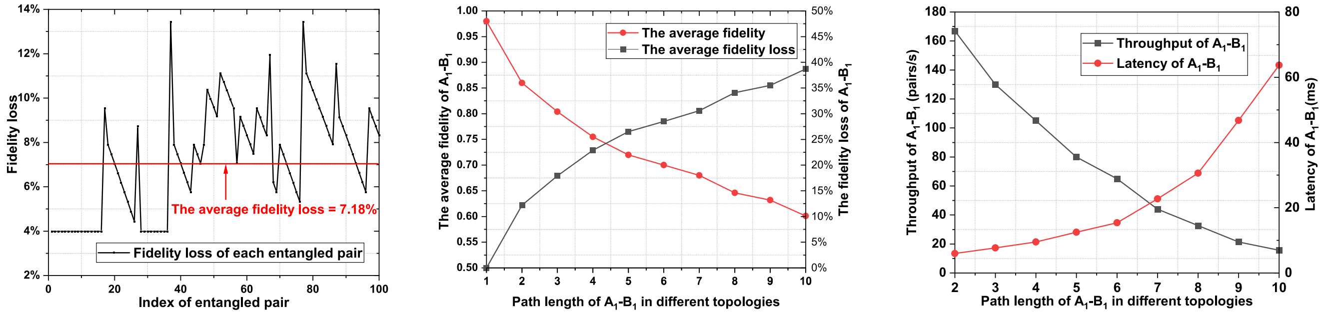
For the evaluation, we consider a classic topology called dumbbell topology shown in Fig. 6, which has six nodes in total, four of which are used as quantum end nodes (A_1, A_2, B_1, B_2), and two of which are used as quantum routing nodes (R_a, R_b) with one bottleneck link ($R_a - R_b$). In the evaluation, we set that each node can store up to 400 qubits, and the length of each communication link is 0.5 km. We set the maximum time that qubit can be stored is 1 s. Besides, the fidelity after entanglement swapping is $F_{\text{final}} = F_1 \times F_2 + \frac{(1-F_1)\times(1-F_2)}{3}$, in which F_1 and F_2 are the fidelities of two Bell pairs [51]. Meanwhile, we adopt the decoherence model provided by NetSquid to account the effect of decoherence on quantum memories. This model randomly dephase a qubit for a delay with a given dephasing rate. The dephasing probability is calculated as $p = 1 - \exp(-\Delta t * R)$, where R is the dephasing rate and Δt is the delay.

B. THROUGHPUT, LATENCY, AND FIDELITY

We define the throughput in a quantum network as the rate of generating entangled pairs between two quantum end nodes. At first, we investigate how the size of allocated memory space impacts the throughput and latency of end-to-end communication. Here, we attempt to observe the performance of our protocol in guaranteeing the throughput and latency when the background end-to-end communication is changing in the quantum network at different times. To reach such goal, scenarios with competition are considered: two end-to-end communications, i.e., ($A_1 - B_1, A_2 - B_2$), exist at the same time.

First, we set the size of allocated quantum memory of $A_1 - B_1$ to be 60, 80, and 120 cells and vary the allocated quantum memory size of $A_2 - B_2$. The rate of generating entangled pairs between the two adjacent nodes is set as 500 pairs/s. The average end-to-end throughput and latency of $A_2 - B_2$ are shown in Fig. 7(a) and (b), respectively. Note that "AMS" in these two figures means "allocated memory size."

The results presented in Fig. 7(a) and (b) show that our protocol can provide different throughput and latency of communication by adjusting the size of the quantum memory allocated to the communication. From Fig. 7(a), we can see that the throughput of $A_2 - B_2$ grows as the allocated quantum memory grows. The maximum throughput of $A_2 - B_2$ is related to the size of quantum memory allocated to another communication $A_1 - B_1$. From Fig. 7(b), we can also see that the end-to-end latency of $A_2 - B_2$ decreases as the allocated quantum memory grows. The latency of $A_2 - B_2$ is also affected by the size of quantum memory allocated to another communication $A_1 - B_1$. In the link layer, we use a simple FIFO algorithm: for each item in the connection table, entangled pairs are generated and stored in the allocated quantum memory between the adjacent nodes, and the items are processed in the order of creation timestamps. Thus, in the simulation, the rate of generating entangled pairs of $A_2 - B_2$ is $R_{A_2-B_2} = \frac{c_2}{c_1+c_2} R_0$, where c_1 and c_2 represent the allocated quantum memory of $A_1 - B_1$ and $A_2 - B_2$, respectively. This explains why the throughput is related to the size of quantum



(a) (b) (c)

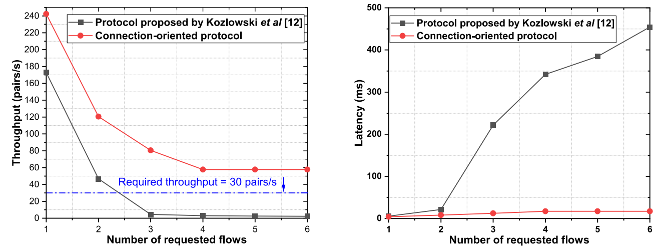
FIGURE 9. Average fidelity, fidelity loss, and throughput of $A_1 - B_1$ in different lengths of the communication path. (a) Fidelity of each entangled pair generated between A_1 and B_1. (b) Average fidelity and average fidelity loss of $A_1 - B_1$. (c) Throughput of $A_1 - B_1$ in different lengths of the communication path.

memory allocated to another communication $A_1 - B_1$ and the same principle for latency.

Second, we set the size of allocated quantum memory of $A_1 - B_1$ to be 30 cells and continuously generate entangled pairs at the beginning of the simulation. The communication of $A_2 - B_2$, whose allocated quantum memory is 60, starts at simulation time 5 s, and the rate of generating entangled pairs between the two adjacent nodes is set as 500 pairs/s. The throughput and the latency of $A_1 - B_1$ and $A_2 - B_2$ are shown in Fig. 8(a) and (b), respectively.

The results presented in Fig. 8(a) and (b) show that the throughput and the latency fluctuate in a small range, which is caused by the failure of entanglement swapping. Besides, we can also observe that the latency of $A_1 - B_1$ increases, and the throughput of it decreases when the communication $A_2 - B_2$ starts. Furthermore, we find that when communication $A_1 - B_1$ and communication $A_2 - B_2$ co-exist at the same time, the throughput and the latency are affected by the allocated memory size. The communication that allocated more memory space has a higher throughput and a lower latency. When the communication $A_1 - B_1$ finishes, the latency decreases and the throughput regains. The simulation results also show that our protocol can guarantee the throughput and the latency by allocating sufficient memory space.

Third, to explore the fidelity of the entangled pairs between two remote quantum nodes, we record the fidelity of 100 entangled pairs generated between A_1 and B_1. In this simulation, we assume that the original fidelity of each link follows a normal distribution with a mean of 0.98, and the fidelity loss of each entangled qubits in two remote quantum end nodes is shown in Fig. 9(a). The average fidelity loss is 7.18%. As our protocol is connection oriented, the communication path between two remote nodes remains unchanged during the end-to-end entanglement distribution. The length of the communication path is the key factor affecting fidelity and throughput. The average fidelity, throughput, and latency of $A_1 - B_1$ in different lengths of communication path are shown in Fig. 9(b) and (c). From Fig. 9(b) and (c), we can see that the average fidelity and throughput decrease significantly with the increase in the path length. Thus, owing to the fidelity loss, without considering the entanglement purification operation, the scale of the quantum network will be limited to a few nodes.



(a) (b)

FIGURE 10. Throughput and latency of $A_1 - B_1$ in different number of requested flows (requested throughput for each flow = 30 pairs/s). (a) Throughput of $A_1 - B_1$ in different number of requested flows. (b) Latency of generating 100 entangled pairs on $A_1 - B_1$ in different number of requested flows.

At last, to make performance comparison between our protocol design and the existing protocol design [12], we also conduct an evaluation and adopt the dumbbell topology, as shown in Fig. 6; the performance in terms of the throughput and the latency is evaluated by generating 100 entangled pairs of $A_1 - B_1$ in total for all requested flows, as shown in Fig. 10(a) and (b). Based on the simulation result, we can find that the gap between the protocol in [12] and our protocol continues to widen in terms of latency. As we know, the protocol in [12] adopts "lazy entanglement tracking." Thus, with the increase of the number of requested flows, different flows compete for the limited memory resources, which leads to a higher latency, and thus, the throughput of the protocol in [12] declines rapidly. In our protocol, the memory resources are reserved for these flows in the link planning phase, which avoids the competition of memory resources, and thus, our protocol provides higher throughput for multiple requested flows. Meanwhile, one important phenomenon shown in Fig. 10 is that, as the number of requested flows increases, the throughput and the latency of the proposed protocol are close to a fixed value eventually. This phenomenon is also caused by the fact that our protocol reserves memory resources for existing flows. Since the memory resources have been allocated

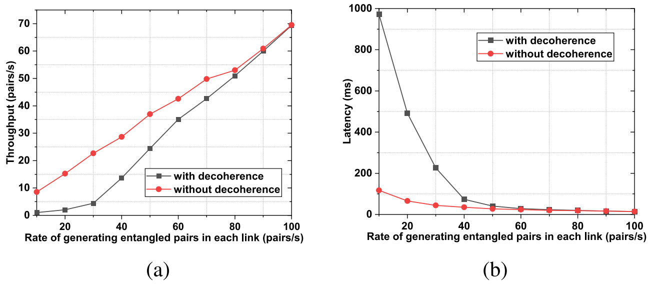


FIGURE 11. **Performance evaluation of the effect of decoherence in different rates of generating entangled pairs between adjacent nodes. (a) Throughput of $A_1 - B_1$ in different rates. (b) Latency of $A_1 - B_1$ in different rates.**

to $A_1 - B_1$ during the link planning phase, as the number of requested flows increases, the proposed protocol will reject to allocate memory resources for new requested flows, which guarantees the latency and the throughput of existing flows that have been previously allocated memory resources.

C. DECOHERENCE

To evaluate the influence of decoherence during the process of entanglement distribution, in this simulation, we set the size of allocated quantum memory of $A_1 - B_1$ to be 40 cells and make the performance comparison of the proposed protocol in two different situations, i.e., with decoherence model and without decoherence model. The results are shown in Fig. 11; the decoherence can significantly affect the performance of $A_1 - B_1$ both in terms of throughput and latency. The reason can be explained as follows. Since we adopt the FIFO scheduling algorithm in the link layer, the allocated quantum memory of $A_1 - B_1$ has to wait for generating entangled pair and then execute entanglement swapping operation. During the waiting time, the decoherence can destroy the entangled pair, and thus, it consumes more resource of entangled pairs and takes more time to successfully establish end-to-end entanglement. However, with the increase in the generation rate, the influence of decoherence becomes diminishing since the waiting time is decreasing. Inspired by this phenomenon, to facilitate the development of entanglement distribution in quantum networks, the ability of generating entangled pairs in each link is important.

VII. CONCLUSION

In this article, based on the different characteristics between classical networks and quantum networks, we indicated that compared to the connectionless protocol, the connection-oriented protocol is more suitable for quantum networks. Then, we designed a connection-oriented entanglement distribution protocol for generating entangled pairs between two end nodes. Our protocol is reliable as the receiver replies with a "received message" as an acknowledgment to the sender for each established entanglement. It avoids the latency caused by resource competition by reserving resources and also provides different throughput and latency by adjusting the size of quantum memory allocated to the communication, which can meet the requirements of different quantum applications. In the future, we will consider integrating routing path finding into our current work to further improve the throughput performance and lower the latency.

REFERENCES

[1] H. J. Kimble, "The quantum internet," *Nature*, vol. 453, no. 7198, pp. 1023–1030, 2008, doi: 10.1038/nature07127.

[2] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, 2018, Art. no. eaam9288, doi: 10.1126/science.aam9288.

[3] V. S. Denchev and G. Pandurangan, "Distributed quantum computing: A new frontier in distributed systems or science fiction," *ACM SIGACT News*, vol. 39, no. 3, pp. 77–95, 2008, doi: 10.1145/1412700.1412718.

[4] A. K. Ekert, "Quantum cryptography based on Bell's theorem," *Phys. Rev. Lett.*, vol. 67, pp. 661–663, 1991, doi: 10.1103/PhysRevLett.67.661.

[5] C. H. Bennett, G. Brassard, and N. D. Mermin, "Quantum cryptography without Bell's theorem," *Phys. Rev. Lett.*, vol. 68, no. 5, pp. 557–559, 1992, doi: 10.1103/PhysRevLett.68.557.

[6] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels," *Phys. Rev. Lett.*, vol. 70, no. 13, 1993, Art. no. 1895, doi: 10.1103/PhysRevLett.70.1895.

[7] J. Yin *et al.*, "Satellite-based entanglement distribution over 1200 kilometers," *Science*, vol. 356, no. 6343, pp. 1140–1144, 2017, doi: 10.1126/science.aan3211.

[8] M. Zukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert, "'Event-ready-detectors' Bell experiment via entanglement swapping," *Phys. Rev. Lett.*, vol. 71, pp. 4287–4290, 1993, doi: 10.1103/PhysRevLett.71.4287.

[9] R.-B. Jin, M. Takeoka, U. Takagi, R. Shimizu, and M. Sasaki, "Highly efficient entanglement swapping and teleportation at telecom wavelength," *Sci. Rep.*, vol. 5, no. 1, pp. 1–7, 2015, doi: 10.9734/JSRR/2015/14076.

[10] A. S. Cacciapuoti, M. Caleffi, R. Van Meter, and L. Hanzo, "When entanglement meets classical communications: Quantum teleportation for the quantum internet," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3808–3833, Jun. 2020, doi: 10.1109/TCOMM.2020.2978071.

[11] A. Dahlberg *et al.*, "A link layer protocol for quantum networks," in *Proc. ACM Spec. Interest Group Data Commun.*, ACM, 2019, pp. 159–173, doi: 10.1145/3341302.3342070.

[12] W. Kozlowski, A. Dahlberg, and S. Wehner, "Designing a quantum network protocol," in *Proc. 16th Int. Conf. Emerg. Netw. Exp. Technol.*, 2020, pp. 1–16, doi: 10.1145/3386367.3431293.

[13] Z. Li *et al.*, "Building a large-scale and wide-area quantum internet based on an OSI-alike model," *China Commun.*, vol. 18, no. 10, pp. 1–14, 2021, doi: 10.23919/JCC.2021.10.001.

[14] A. Viswanathan, E. C. Rosen, and R. Callon, "Multiprotocol label switching architecture," *RFC* 3031, IETF, 2001. Accessed: Feb. 2022. [Online]. Available: https://www.ietf.org/rfc/rfc3031.txt

[15] J. Postel, "Internet protocol," RFC 791, IETF, 1981. Accessed: Feb. 2022. [Online]. Available: https://www.ietf.org/rfc/rfc791.txt

[16] D. S. Golubev and A. D. Zaikin, "Quantum decoherence in disordered mesoscopic systems," *Phys. Rev. Lett.*, vol. 81, no. 5, pp. 1074–1077, 1998, doi: 10.1103/PhysRevLett.81.1074.

[17] T. Coopmans *et al.*, "NetSquid, A network simulator for quantum information using discrete events," *Commun. Phys.*, vol. 4, no. 1, pp. 1–15, 2021, doi: 10.1038/s42005-020-00504-0.

[18] S. Lloyd, J. H. Shapiro, F. N. Wong, P. Kumar, S. M. Shahriar, and H. P. Yuen, "Infrastructure for the quantum Internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 5, pp. 9–20, 2004, doi: 10.1145/1039111.1039118.

[19] S. Pirandola and S. L. Braunstein, "Physics: Unite to build a quantum internet," *Nature*, vol. 532, no. 7598, pp. 169–171, 2016, doi: 10.1038/532169a.

[20] S. Pirandola *et al.*, "Advances in quantum cryptography," *Adv. Opt. Photon.*, vol. 12, no. 4, pp. 1012–1236, 2020, doi: 10.1364/AOP.361502.

[21] R. Van Meter, T. Satoh, T. D. Ladd, W. J. Munro, and K. Nemoto, "Path selection for quantum repeater networks," *Netw. Sci.*, vol. 3, pp. 82–95, 2013, doi: 10.1007/s13119-013-0028-0.

[22] L. Gyongyosi and S. Imre, "Decentralized base-graph routing for the quantum internet," *Phys. Rev. A*, vol. 98, no. 2, 2018, Art. no. 022310, doi: 10.1103/PhysRevA.98.022310.

[23] L. Gyongyosi and S. Imre, "Routing space exploration for scalable routing in the quantum internet," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020, doi: 10.1038/s41598-020-68354-y.

[24] L. Gyongyosi, "Dynamics of entangled networks of the quantum internet," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 12909, doi: 10.1038/s41598-020-68498-x.

[25] K. Chakraborty, D. Elkouss, B. Rijsman, and S. Wehner, "Entanglement distribution in a quantum network: A multicommodity flow-based approach," *IEEE Trans. Quantum Eng.*, vol. 1, 2020, Art. no. 4101321, doi: 10.1109/TQE.2020.3028172.

[26] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, "Quantum repeaters: The role of imperfect local operations in quantum communication," *Phys. Rev. Lett.*, vol. 81, no. 26, pp. 5932–5935, 1998, doi: 10.1103/PhysRevLett.81.5932.

[27] S. Brand, T. Coopmans, and D. Elkouss, "Efficient computation of the waiting time and fidelity in quantum repeater chains," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 3, pp. 619–639, Mar. 2020, doi: 10.1109/JSAC.2020.2969037.

[28] A. M. Goebel *et al.*, "Multistage entanglement swapping," *Phys. Rev. Lett.*, vol. 101, no. 8, 2008, Art. no. 080403, doi: 10.1103/PhysRevLett.101.080403.

[29] S. Khatri, C. T. Matyas, A. U. Siddiqui, and J. P. Dowling, "Practical figures of merit and thresholds for entanglement distribution in quantum networks," *Phys. Rev. Res.*, vol. 1, no. 2, 2019, Art. no. 023032, doi: 10.1103/PhysRevResearch.1.023032.

[30] W. J. Munro, K. Azuma, K. Tamaki, and K. Nemoto, "Inside quantum repeaters," *IEEE J. Sel. Topics Quantum Electron.*, vol. 21, no. 3, pp. 78–90, May/Jun. 2015, doi: 10.1109/JSTQE.2015.2392076.

[31] K. Chakraborty, F. Rozpedek, A. Dahlberg, and S. Wehner, "Distributed routing in a quantum internet," 2019, *arXiv:1907.11630*, doi: 10.48550/arXiv.1907.11630.

[32] M. Pant *et al.*, "Routing entanglement in the quantum internet," *npj Quantum Inf.*, vol. 5, no. 1, 2019, Art. no. 25, doi: 10.1038/s41534-018-0113-z.

[33] J. Li *et al.*, "Fidelity-guaranteed entanglement routing in quantum networks," 2021, *arXiv:2111.07764*, doi: 10.48550/arXiv.2111.07764.

[34] M. Caleffi, "Optimal routing for quantum networks," *IEEE Access*, vol. 5, pp. 22299–22312, 2017, doi: 10.1109/ACCESS.2016.2641474.

[35] T. Matsuo, C. Durand, and R. V. Meter, "Quantum link bootstrapping using a ruleset-based communication protocol," *Phys. Rev. A*, vol. 100, no. 5, 2019, Art. no. 052320, doi: 10.1103/PhysRevA.100.052320.

[36] H. Bernien *et al.*, "Heralded entanglement between solid-state qubits separated by three metres," *Nature*, vol. 497, no. 7447, pp. 86–90, 2013, doi: 10.1038/nature12016.

[37] D. L. Moehring *et al.*, "Entanglement of single-atom quantum bits at a distance," *Nature*, vol. 449, no. 7158, pp. 68–71, 2007, doi: 10.1038/nature06118.

[38] S. Olmschenk, D. Matsukevich, P. Maunz, D. Hayes, L.-M. Duan, and C. Monroe, "Quantum teleportation between distant matter qubits," *Science*, vol. 323, no. 5913, pp. 486–489, 2009, doi: 10.1126/science.1167209.

[39] J.-W. Pan, D. Bouwmeester, H. Weinfurter, and A. Zeilinger, "Experimental entanglement swapping: Entangling photons that never interacted," *Phys. Rev. Lett.*, vol. 80, no. 18, 1998, Art. no. 3891, doi: 10.1103/PhysRevLett.80.3891.

[40] Y. Pu, N. Jiang, W. Chang, H. Yang, C. Li, and L. Duan, "Experimental realization of a multiplexed quantum memory with 225 individually accessible memory cells," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 15359, doi: 10.1038/s41467-016-0009-6.

[41] J. Illiano, M. Caleffi, A. Manzalini, and A. S. Cacciapuoti, "Quantum internet protocol stack: A comprehensive survey," 2022, *arXiv:2202.10894*, doi: 10.48550/arXiv.2202.10894.

[42] K. Boström and T. Felbinger, "Deterministic secure direct communication using entanglement," *Phys. Rev. Lett.*, vol. 89, no. 18, 2002, Art. no. 187902, doi: 10.1103/PhysRevLett.89.187902.

[43] F.-G. Deng, G. L. Long, and X.-S. Liu, "Two-step quantum direct communication protocol using the Einstein-Podolsky-Rosen pair block," *Phys. Rev. A*, vol. 68, no. 4, 2003, Art. no. 042317, doi: 10.1103/PhysRevA.68.042317.

[44] W. Zhang, D.-S. Ding, Y.-B. Sheng, L. Zhou, B.-S. Shi, and G.-C. Guo, "Quantum secure direct communication with quantum memory," *Phys. Rev. Lett.*, vol. 118, no. 22, 2017, Art. no. 220501, doi: 10.1103/PhysRevLett.118.220501.

[45] A. S. Cacciapuoti, M. Caleffi, F. Tafuri, F. S. Cataliotti, S. Gherardini, and G. Bianchi, "Quantum internet: Networking challenges in distributed quantum computing," *IEEE Netw.*, vol. 34, no. 1, pp. 137–143, Jan./Feb. 2019, doi: 10.1109/MNET.001.1900092.

[46] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, "Fundamental limits of repeaterless quantum communications," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 15043, doi: 10.1038/ncomms15043.

[47] S. Muralidharan, L. Li, J. Kim, N. Lütkenhaus, M. D. Lukin, and L. Jiang, "Optimal architectures for long distance quantum communication," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 20463, doi: 10.1038/srep20463.

[48] S. Shi and C. Qian, "Concurrent entanglement routing for quantum networks: Model and designs," in *Proc. Annu. Conf. ACM Spec. Int. Group Data Commun.*, 2020, pp. 62–75, doi: 10.1145/3387514.3405853.

[49] L. Gyongyosi and S. Imre, "Adaptive routing for quantum memory failures in the quantum internet," *Quantum Inf. Process.*, vol. 18, no. 2, pp. 1–21, 2019, doi: 10.1007/s11128-018-2153-x.

[50] F. Hahn, A. Pappa, and J. Eisert, "Quantum network routing and local complementation," *npj Quantum Inf.*, vol. 5, no. 1, 2019, Art. no. 76, doi: 10.1038/s41534-019-0191-6.

[51] R. F. Werner, "Quantum states with Einstein-Podolsky-Rosen correlations admitting a hidden-variable model," *Phys. Rev. A*, vol. 40, no. 8, pp. 4277–4281, 1989, doi: 10.1103/PhysRevA.40.4277.

Jian Li (Member, IEEE) received the bachelor's degree from Anhui University, Hefei, China, in 2015, and the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, in 2020.

From November 2019 to November 2020, he was a Visiting Scholar with the Department of Electronic and Computer Engineering, University of Florida, Gainesville, FL, USA. He is currently a Postdoctoral Researcher with the School of Cyber Science and Technology, USTC. His research interests include quantum networks, wireless networks, and next-generation Internet.

Qidong Jia received the bachelor's degree in information security in 2019 from the School of Cyber Science and Technology, University of Science and Technology of China, Hefei, China, where he is currently working toward the master's degree.

His research interests include quantum Internet architecture and quantum networking.

Kaiping Xue (Senior Member, IEEE) received the bachelor's degree from the University of Science and Technology of China (USTC), Hefei, China, in 2003, and the Ph.D. degree from USTC in 2007.

From May 2012 to May 2013, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA. He is currently a Professor with the School of Cyber Science and Technology, USTC. His research interests include next-generation Internet architecture design, transmission optimization, and network security.

Dr. Xue is on the editorial board of several journals, including IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT. He was a (Lead) Guest Editor for many reputed journals/magazines, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, *IEEE Communications Magazine*, and IEEE NETWORK. He is a Fellow of the Institution of Engineering and Technology.

David S. L. Wei (Senior Member, IEEE) received the Ph.D. degree in computer and information science from the University of Pennsylvania, Philadelphia, PA, USA, in 1991.

From May 1993 to August 1997, he was on the Faculty of Computer Science and Engineering, University of Aizu, Aizuwakamatsu, Japan, as an Associate Professor and then a Professor. He has authored or coauthored more than 120 technical papers in various archival journals and conference proceedings. He is currently a Professor with the Department of Computer and Information Science, Fordham University, New York, NY, USA. His research interests include cloud computing, big data, Internet of Things, and cognitive radio networks.

Dr. Wei was a Guest Editor or a Lead Guest Editor for several special issues in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON CLOUD COMPUTING, and IEEE TRANSACTIONS ON BIG DATA. He was an Associate Editor for IEEE TRANSACTIONS ON CLOUD COMPUTING from 2014 to 2018 and *Journal of Circuits, Systems and Computers*, from 2013 to 2018.

Nenghai Yu (Member, IEEE) received the B.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1987, the M.E. degree from Tsinghua University, Beijing, China, in 1992, and the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2004.

He is currently a Professor with the Department of Electronic Engineering and Information Science and the School of Cyber Science and Technology, USTC. He is also the Executive Dean of the School of Cyber Science and Technology, USTC, and the Director of the Information Processing Center, USTC. He has authored or coauthored more than 130 papers in journals and international conferences. His research interests include multimedia security, multimedia information retrieval, video processing, information hiding, and quantum networking.