Swapping-Based Entanglement Routing Design for Congestion Mitigation in Quantum Networks

Zhonghui Li, *Graduate Student Member, IEEE*, Jian Li, *Member, IEEE*, Kaiping Xue, *Senior Member, IEEE*, David S. L. Wei, *Life Senior Member, IEEE*, Ruidong Li, *Senior Member, IEEE*, Nenghai Yu, Qibin Sun, *Fellow, IEEE*, and Jun Lu

***Abstract*—The quantum network is designed to connect numerous quantum nodes and support various ground-breaking quantum applications. Most of these applications require communicating parties to share entangled pairs. Therefore, entanglement routing, a technology distributing entangled pairs between distant quantum nodes, plays a vital role in realizing quantum networks' capability. However, due to the limitation of quantum memory size and quantum decoherence, the entangled pairs shared by adjacent quantum nodes can hardly satisfy concurrent entanglement routing requests, thus leading to severe network congestion. In this paper, we propose a novel congestion mitigation (CM) scheme to tackle such bottleneck problems. The basic idea of CM is to "recycle" idle link-level entanglement resources from well-resourced links to bottleneck links utilizing a unique enabling technology of quantum networks, called entanglement swapping. CM can increase the capacity of each bottleneck link, thus overcoming resource limitations to improve resource utilization and network throughput. To complete our work, we also propose a swapping-based entanglement routing design, including path selection and resource allocation algorithms. Extensive simulations show that our design can significantly alleviate network congestion and improve the request service rate of quantum networks compared to the traditional entanglement routing designs.**

***Index Terms*—Quantum networks, entanglement routing, entanglement swapping, congestion mitigation.**

Manuscript received 18 June 2022; revised 1 November 2022; accepted 8 May 2023. Date of publication 12 May 2023; date of current version 12 December 2023. This work is supported in part by National Scientific and Technological Innovation 2030 Major Project of Quantum Communications and Quantum Computers under grant No. 2021ZD0301301, Anhui Initiative in Quantum Information Technologies under grant No. AHY150300, Youth Innovation Promotion Association Chinese Academy of Science (CAS) under grant No. Y202093, and JSPS KAKENHI under Grant No. 23H03380. The associate editor coordinating the review of this article and approving it for publication was C. Avin. *(Corresponding authors: Jian Li; Kaiping Xue.)*

Zhonghui Li, Jian Li, Kaiping Xue, Nenghai Yu, and Qibin Sun are with the School of Cyber Science and Technology and the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, Anhui, China (e-mail: lijian9@ustc.edu.cn; kpxue@ustc.edu.cn).

David S. L. Wei is with the Department of Computer and Information Science, Fordham University, Bronx, NY 10458 USA.

Ruidong Li is with the College of Science and Engineering, Kanazawa University, Kanazawa 920-1192, Japan.

Jun Lu is with the School of Cyber Science and Technology, the School of Information Science and Technology, and the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, Anhui, China.

Digital Object Identifier 10.1109/TNSM.2023.3275815

I. Introduction

QUANTUM networks [1], [2], [3], [4] are distributed systems to connect numerous quantum nodes to support various quantum applications, such as quantum communication [5], distributed quantum computing [6], [7], [8], and improved sensing [9], [10]. As an essential function in a quantum network, the distribution of entangled pairs (also known as EPR pairs) between two direct-linked quantum nodes has been experimentally validated over short-distance quantum channels [11], [12]. However, due to quantum decoherence [13], [14] and the "no-cloning" theorem [15], the "store-and-forward" communication approach used in classical networks becomes unavailable for remote entanglement distribution in quantum networks. Fortunately, entanglement swapping [16] - a quantum technology that can "couple" multiple EPR pairs shared by adjacent quantum nodes into end-to-end entanglement - can be considered a reliable relay solution to enable EPR pairs to be shared by a pair of distant quantum end nodes. Thus, for a remote entanglement distribution request from a source-destination (SD) pair in quantum networks, end-to-end entanglement can be established by performing entanglement swapping along a path (hereafter referred to as a swapping path) consisting of multiple quantum repeaters [17], [18].

In this paper, we investigate the entanglement routing problem, namely, how to generate end-to-end entanglement by performing entanglement swapping in quantum networks [19]. The design of entanglement routing poses several challenges: 1) *The swapping operation might fail*. Due to the imperfection of quantum hardware [20], the implementation of end-to-end entanglement distribution is probabilistic. Hence, an entanglement routing design should minimize the effects of imperfect swapping operations. 2) *An EPR pair cannot be shared by multiple requests*. Due to the collapse-after-measurement theory, an EPR pair can only serve one SD pair for end-to-end entanglement distribution. Besides, it is hard to distribute EPR pairs between adjacent quantum nodes. As a result, the limited EPR pairs will block concurrent entanglement routing requests. Therefore, we expect that entanglement routing design can effectively mitigate network congestion. 3) *Quantum decoherence limits the life span of entangled states*. Since entangled states decay during storing in quantum memory [21], each EPR pair can only be valid for a very short time, e.g., a typical lifetime is 1.46s [22].

To fully synchronize and utilize the entanglement resources, a time-division network model is preferred to guarantee an appropriate entanglement routing duration. These challenges, with no counterpart in classical networks, make it impossible to apply classical designs directly to quantum networks. Therefore, designing an effective entanglement routing scheme becomes an urgent yet challenging problem.

A. Related Work

Several enlightening entanglement routing designs have been proposed in quantum networks. Research [23] is one of the first works to study the entanglement routing problem in quantum networks using the traditional Dijkstra's algorithm. In the study of [19], the author proposes a more complicated routing metrics taking into account some physical factors, such as decoherence time and the success probability of entanglement swapping. Both [23] and [19] assume that there is only one entanglement routing request at a specific time, which is not practical in quantum networks. Pirandola [24] proposed an entanglement routing protocol in a diamond topology. However, their protocol relies on the assumption that entanglement swapping can be performed flawlessly, which is obviously not applicable to the practical system. Pant et al. [25] introduced a greedy algorithm for path selection in a grid topology. However, this algorithm only works well on the networks having one shared EPR pair between two adjacent quantum nodes. Shi and Qian [26] proposed the Q-CAST algorithm, using the EPR pairs shared by adjacent nodes in the sub-optimal paths as a backup resource to remedy the failure of entanglement routing. Although the Q-CAST algorithm was shown to achieve higher throughput than some existing algorithms, the performance of Q-CAST is sub-optimal. In addition, Schoute et al. [27] investigated the entanglement routing problem but limited for ring and sphere topologies. Das et al. [28] assessed entanglement routing designs in different particular topologies.

The existing research works mentioned above mainly focus on path selection in the entanglement routing design. However, the problem of limited entanglement resources unable to meet the demand of numerous requests in a concurrent entanglement routing request scenario, i.e., network congestion problem, is also one of the essential issues in entanglement routing design. To the best of our limited knowledge, how to mitigate network congestion caused by the gap between concurrent entanglement routing requests and the limited capacity of quantum networks remains an open problem in entanglement routing design. In this work, we focus on how to alleviate network congestion in entanglement routing design.

B. Our Contributions

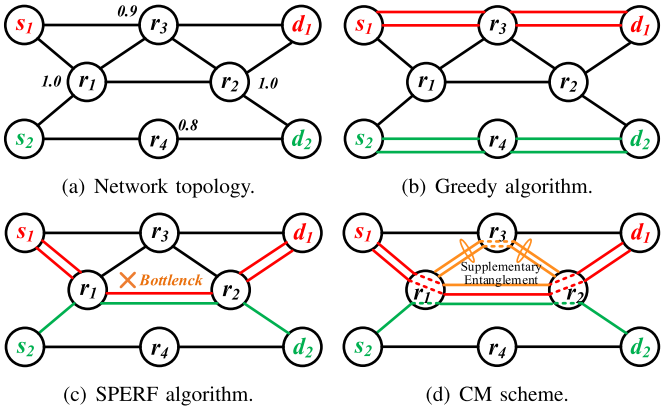
To satisfy the demands posed by concurrent entanglement routing requests, in this paper, we propose a swapping-based entanglement routing design, which can effectively reduce the number of blocked requests in a resource-limited quantum network. We adopt a time-division model, in which each entanglement routing task in a time slot is completed by two core parts, i.e., path selection and resource allocation. For the path selection problem, we propose a success probability of entanglement routing first algorithm, named SPERF, to select the swapping path with hops limit. Furthermore, we propose a novel congestion mitigation (CM) scheme for tackling the resource competition problem, which contains two steps, i.e., entanglement resource allocation (ERA) and entanglement resources transformation (ERT). At the beginning, ERA allocates the entanglement resources to entanglement routing requests along swapping paths selected by SPERF. When there is insufficient entanglement resource to meet concurrent entanglement routing requests, the bottleneck links might exist in quantum networks. Then, ERT "recycles" the idle EPR pairs from the well-resourced quantum links to improve the capacities of bottleneck links based on such a finding that swapping operation can transform entanglement relationships between quantum nodes. In this way, the performance of resource utilization and the number of satisfied SD pairs can be significantly improved.

We summarize the main contributions of this paper as follows:

- We propose the SPERF algorithm for path selection to mitigate the imperfection of entanglement swapping. Based on a centralized network model, we consider a time-synchronous network operation model, and a complicated entanglement routing problem can be decomposed into two essential sub-problems. For the first sub-problem, i.e., path selection, we adopt the success probability of establishing end-to-end entanglement as a new routing metric to design SPERF.
- We propose a novel CM scheme including ERA and ERT algorithms to tackle the second sub-problem, i.e., resource allocation. The ERA algorithm adopts a priority-based solution to pre-allocate the entanglement resources on the bottleneck links to improve the number of satisfied SD pairs. To improve the capacity of each bottleneck link, the ERT algorithm presents a novel idea to "recycle" the idle entanglement resources based on the unique properties of entanglement swapping.
- We conduct extensive simulations to verify the effectiveness of the proposed algorithms. Compared to the existing greedy and random allocation approaches, the performance evaluations demonstrate the superiority of our scheme in terms of the number of satisfied SD pairs.

C. Paper Organization

The remainder of this paper is organized as follows. In Section II, we first briefly describe the overview of our work and network components. Moreover, we present the system model considered in this paper and describe two important problems, i.e., path selection and resource allocation, in quantum networks. Furthermore, the design of the path selection algorithm and CM scheme consisting of ERA and ERT algorithms are introduced in Section III. At last, the performance evaluation is conducted in Section IV, and the conclusions are drawn in Section V.



(a) Network topology. (b) Greedy algorithm. (c) SPERF algorithm. (d) CM scheme.

Fig. 1. Work examples (Black solid lines are physical edges, and the existence of an edge between two nodes means that they can share entangled pairs. Solid lines in other colors represent link-level entanglement. Dashed lines represent the swapping operation.

II. Background and System Model

This section first provides an overview to show the whole picture of our work. Then we present the network components and system model considered in this paper. Furthermore, the entanglement routing problem in quantum networks is defined and transformed into two sub-problems, i.e., path selection and entanglement resource allocation.

A. Overview

Entanglement routing works by selecting a swapping path connecting each SD pair [29], [30], [31] to "couple" multiple EPR pairs shared by adjacent quantum nodes into an end-to-end entanglement. Generally, link-level entanglement is established by distributing EPR pairs between adjacent quantum nodes [32], [33], [34], and multiple parallel EPR pairs can be distributed over a quantum link. Considering that entanglement distribution is probabilistic and the layered design is feasible in quantum networks [35], [36], [37], we consider an "continuous model" [38], [39] (also known as connectionless strategy [40]) in this paper, which means a given number of EPR pairs are pre-shared by two adjacent quantum nodes before path selection. The transition from link-level entanglement to end-to-end entanglement is realized by performing entanglement swapping- essentially a local measurement operation assisted by classical communication.

We focus on the implementation of concurrent entanglement routing requests between multiple SD pairs in a general quantum network topology, such as the one shown in Fig. 1(a). In this example, we assume that there are two SD pairs, $\langle s_1, d_1\rangle$ and $\langle s_2, d_2\rangle$, simultaneously request to share one and two EPR pairs in a time slot, respectively. Considering the limited quantum memory resources of each quantum node, we assume that each edge shares two EPR pairs. Besides, the success probability of entanglement swapping for the four quantum repeaters, i.e., r_1, r_2, r_3, and r_4, is set to 1.0, 1.0, 0.9 and 0.8, respectively.

The recent work in [25] introduces a greedy routing algorithm, which selects the swapping path with the fewest hops in quantum networks to generate end-to-end entanglement between SD pairs. As a result, the greedy algorithm will find swapping path $s_1 \rightarrow r_3 \rightarrow d_1$ and $s_2 \rightarrow r_4 \rightarrow d_2$ for two SD pairs as shown in Fig 1(b). However, it is worth noting that entanglement swapping is an imperfect operation, which means that the swapping path with the minimum hop counts does not represent the path with the maximum probability of successful end-to-end entanglement distribution. Consequently, there are other swapping paths that would be better for the SD pair $\langle s_1, d_1\rangle$ to successfully build end-to-end entanglement, e.g., the path going through quantum repeaters r_1 and r_2.

Fig. 1(c) illustrates the SPERF routing algorithm proposed in this paper for path selection. SPERF algorithm defines a new routing metric, i.e., the success probability of establishing an end-to-end entanglement, to value the quality of a swapping path. The new routing metric mainly takes the success probability of entanglement swapping and hops into account simultaneously. SPERF is realized based on two core ideas, i.e., the quantum node with a high success probability of entanglement swapping is preferred, and the path connecting each SD pair with a smaller number of hops is preferred. As a result, the SPERF routing algorithm finds swapping paths $s_1 \rightarrow r_1 \rightarrow r_2 \rightarrow d_1$ and $s_2 \rightarrow r_1 \rightarrow r_2 \rightarrow d_2$ for two SD pairs, respectively, which can improve the success probability of entanglement routing.

Due to the mismatch between the total resource demands of concurrent entanglement routing requests and the limited EPR pairs shared by each pair of adjacent quantum nodes, both the greedy algorithm and the SPERF algorithm result in network congestion. As shown in Fig. 1(c), the capacity of the edge (r_1, r_2) is less than the entanglement resources required by two SD pairs. Consequently, the bottleneck edge can only allow one entanglement routing request ($\langle s_1, d_1\rangle$ or $\langle s_2, d_2\rangle$) to be served in a time slot. To tackle such a bottleneck problem, we present the CM scheme to effectively realize the generation of end-to-end entanglement between multiple SD pairs simultaneously. As shown in Fig. 1(d), the idle link-level entanglement resources possessed by edges (r_1, r_3) and (r_2, r_3) can be "recycled" to increase the capacity of the bottleneck edge (r_1, r_2). Concretely, (r_1, r_3) and (r_2, r_3) allocate two EPR pairs, respectively, to supplement one EPR pair to (r_1, r_2) by performing entanglement swapping on quantum repeater r_3. As a result, the number of EPR pairs possessed by edge (r_1, r_2) equals the number of EPR pairs required by the two SD pairs, i.e., concurrent entanglement routing requests can be satisfied at the same time.

B. Network Components

Centralized Processor: Generally, a quantum network works in synergy with classical networks because quantum operations are inseparable from classical communication [41]. Notably, the frequent and complex classical information interactions when multiple SD pairs attempt to establish end-to-end entanglement cannot be ignored [42]. Local classical controllers adopted for entanglement routing would result in classical message flooding since each quantum node does not have

the global information (e.g., the success probability of entanglement swapping and entanglement resources) of quantum networks [25]. Besides, the centralized processing design is beneficial for tracking entanglement relationships during the generation of end-to-end entanglement, thus facilitating entanglement relationship synchronization. Hence, we introduce a classical centralized processor as an auxiliary tool in quantum networks, the function of which is to perform the routing calculations and broadcast the scheduling information to quantum nodes.

Quantum Nodes: Each quantum node is a quantum information processing device. In this paper, we assume that each node can generate, store, transmit, and manipulate quantum states. Any pair of adjacent quantum nodes can establish link-level entanglement by entanglement generation. Each quantum node is equipped with a finite number of quantum memory units [43] and the necessary hardware to perform quantum operations, such as entanglement swapping. Here, quantum nodes mainly include quantum end nodes, quantum repeaters, and quantum routers. Quantum end nodes are responsible for teleporting unknown quantum bits and processing quantum information to support various quantum applications. Quantum repeaters aim at extending the distance of entanglement distribution. A quantum router is a networking device that connects numerous quantum end nodes together and routes each entanglement routing request to the destination node. Quantum routers and quantum repeaters work together to support the interconnection of numerous quantum end nodes in a quantum network.

Quantum Link: In quantum networks, a quantum link connecting two adjacent quantum nodes supports the distribution of EPR pairs, and it is essentially a quantum channel. There are two types of quantum links: optical fiber and free space. Both types of quantum links are inherently lossy and decoherence [44], which leads to the fact that the success probability of entanglement distribution between adjacent quantum nodes exponentially decays with the physical length of a quantum link [45], [46]. Hence, two adjacent quantum nodes must make multiple entanglement distribution attempts to generate entanglement over a quantum link. Besides, multiple parallel EPR pairs can be possessed by a quantum link with the aid of quantum memory.

Quantum Memory: In the past decades, quantum memory has been studied in a variety of storage schemes [47], [48], [49], [50]. Quantum memories are becoming practical in terms of coherence time, fidelity, and efficiency. Considering that the success probability of entanglement distribution between adjacent quantum nodes is extremely low and the storage fidelity of quantum memory is up to 99.5% [51], we adopt a "continuous model", i.e., EPR pairs are shared by adjacent quantum nodes before path selection, to serve entanglement routing requests. Besides, since quantum memory can be designed as the combination of multiple independent accessible memory units [52], we can assign a unique identity for each entanglement stored in memory in the form of the shared EPR pairs to distinguish each other, which can guarantee that end-to-end entanglement is correctly established between each SD pair.

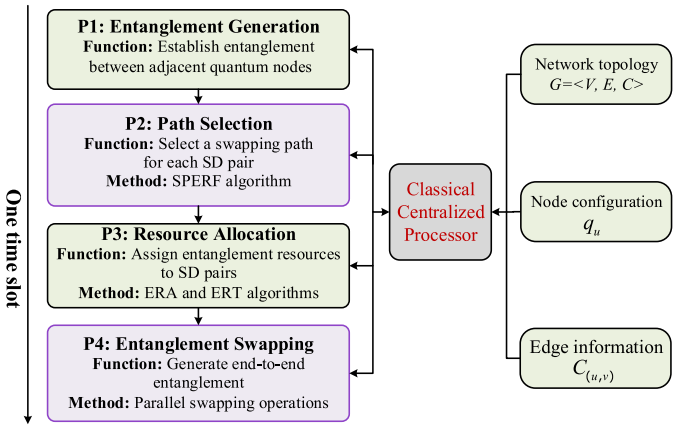


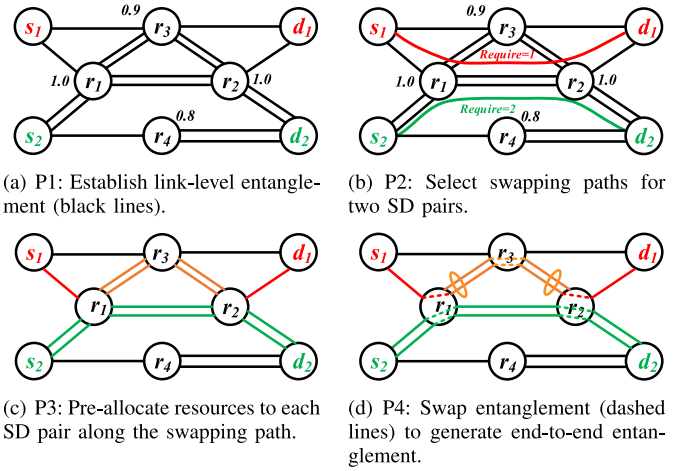
Fig. 2. Four phases assisted by the centralized processor in one time slot.

C. System Model

We abstract a quantum network as an undirected graph $G = \langle V, E, C \rangle$, where V is the set of $|V|$ nodes, E is the set of $|E|$ edges in the graph, and $|C|$ is the set of all edge capacities. Each node $u \in V$ represents a quantum node (including the end node and networking node), and the edge $(u, v) \in E$ presents the physical quantum link between adjacent quantum nodes u and v. The success probability of entanglement swapping of the quantum node u is denoted as q_u. Besides, we define the edge capacity, $C_{(u,v)}$, as the number of the EPR pairs shared by a pair of adjacent quantum nodes u and v.

To reduce the delay in the generation of end-to-end entanglement, we adopt a parallel strategy to perform quantum operations, including entanglement generation and entanglement swapping. For parallel swapping operations, all quantum nodes on a swapping path must simultaneously be entangled with their predecessor and successor. Hence, time synchronization among all nodes is necessary [26]. Here, we introduce the concept of time slots in quantum networks. By dividing each time slot into four phases, a complicated entanglement routing problem can be decomposed into several sub-problems that are easier to solve. Concretely, a time slot consists of four phases, i.e., entanglement generation, path selection, resource allocation, and entanglement swapping. As shown in Fig. 2, entanglement routing is realized under the control of a classical centralized processor, which collects the global information of quantum networks, e.g., network topology, node configuration, and edge information, via the classical Internet. We elaborate on the implementation of each phase in what follows.

Phase One (**P1**) is called the entanglement generation phase, responsible for generating link-level entanglement resources. Each quantum node allocates a non-uniform number of quantum memory units to each edge to store EPR pairs. At the beginning of each time slot, the light source of entangled photons makes several attempts to distribute EPR pairs to two adjacent quantum nodes through quantum channels until all the allocated quantum memory units are occupied or timeout [22]. As a result, each edge possesses multiple EPR pairs (black line in Fig. 3(a)). For example, the capacity of the edge (r_1, r_2) is two, i.e., $C_{(r_1,r_2)} = 2$.



(a) P1: Establish link-level entanglement (black lines).

(b) P2: Select swapping paths for two SD pairs.

(c) P3: Pre-allocate resources to each SD pair along the swapping path.

(d) P4: Swap entanglement (dashed lines) to generate end-to-end entanglement.

Fig. 3. The main work of four phases: entanglement generation, path selection, resource allocation, and entanglement swapping.

Phase Two (**P2**), called the path selection phase, aims to select swapping paths for multiple SD pairs according to the results of the routing algorithm running on the classical central processor. At the beginning of **P2**, the classical centralized processor receives the requests from multiple SD pairs that need to establish end-to-end entanglement and collects the configuration information of quantum nodes and edge states via the Internet.[1] Then, the centralized processor executes the SPERF algorithm to select a swapping path for each SD pair. A swapping path is identified by the sequence of the quantum nodes along the swapping path. As shown in Fig. 3(b), $s_1 \rightarrow r_1 \rightarrow r_2 \rightarrow d_1$ and $s_2 \rightarrow r_1 \rightarrow r_2 \rightarrow d_2$ are the swapping paths selected for two SD pairs $\langle s_1, d_1 \rangle$ and $\langle s_2, d_2 \rangle$, respectively.

Phase Three (**P3**), also called the resource allocation phase, is responsible for assigning link-level entanglement resources to each SD pair under the control of the centralized processor. In this phase, the centralized processor runs the CM scheme (including ERA and ERT algorithms) and broadcasts entanglement resource allocation scheduling to all quantum nodes on the swapping path. We assume two SD pairs expect to share one and two EPR pairs, respectively, in a time slot. As shown in Fig. 3(c), the edge (r_1, r_2) allocates two EPR pairs to SD pair $\langle s_2, d_2 \rangle$, and the link-level entanglement resources that SD pair $\langle s_1, d_1 \rangle$ lacks can be supplemented by utilizing idle link-level entanglement resources (yellow line in Fig. 3(c)) on edges (r_1, r_3) and (r_2, r_3). As a result, no entanglement routing requests are blocked in this time slot.

Phase Four (**P4**) is the entanglement swapping phase, the primary function of which is to establish end-to-end entanglement between SD pairs. In this phase, quantum nodes on the swapping path perform entanglement swapping to patch multiple link-level entanglement resources allocated on each edge together to form long-distance end-to-end entanglement, as shown in Fig. 3(d). In order to reduce the delay in entanglement routing, a parallel entanglement swapping strategy [53] can be adopted instead of performing swapping operations hop-by-hop. In other words, the quantum nodes on a swapping path simultaneously perform a local measurement to swap entanglement, and the destination node manipulates the entangled quantum bits it possesses according to the aggregated measurement results to establish end-to-end entanglement with the source node.

[1]In general, a quantum network is designed to assist the Internet to realize unconditional secure communication.

D. The Entanglement Routing Problem

After link-level entanglement between adjacent quantum nodes is created, the main work of entanglement routing is to select swapping paths and to establish end-to-end entanglement by performing swapping operations to "couple" the allocated EPR pairs [54]. Considering some practical situations, e.g., concurrent entanglement routing requests, the imperfection of entanglement swapping, and the limited entanglement resources, there are mainly two problems, i.e., path selection and entanglement resource allocation, that need to be solved for the design of entanglement routing.

Path Selection: The path selection problem in entanglement routing design is defined as follows: Given a quantum network with non-uniform edge capacity and random topology, design a routing algorithm that can provide an effective solution to select a swapping path for arbitrary SD pair.

Path selection significantly affects the distribution rate of end-to-end entanglement in quantum networks since entanglement swapping is an imperfect operation. Although the basic structure of quantum networks is analogous to classical networks, the existing routing technologies adopted in classical networks are insufficient to solve the path selection problem in quantum networks. The reasons are described as follows: 1) Classical packets can be buffered in any node for a long time for future transmission. However, entangled states only have a short lifespan due to the phenomenon (known as quantum decoherence) that the quality of the entangled system decays gradually during the interaction with the noisy environment; 2) An EPR pair can only be used to establish end-to-end entanglement for one SD pair due to the "collapse after measurement" phenomenon. That is, EPR pairs cannot be shared by different SD pairs. However, a classical link can serve multiple data flows. In summary, we need to present a new path selection algorithm considering the unique characteristics of quantum networks. In this paper, we introduce the SPERF algorithm to select swapping paths for entanglement routing, and the specific design of the SPERF algorithm will be discussed in Section III-A.

Resource Allocation: Considering entanglement resource competition and bottleneck problem, the resource allocation problem in this work is defined as follows: Given an arbitrary network topology, design a scheme to handle resource competition between multiple SD pairs to maximize the number of satisfied entanglement routing requests and determine which idle link-level entanglement resources are used to increase the capacity of the bottleneck edge.

It is worth noting that one EPR pair can not be shared by multiple SD pairs. Generally, each SD pair will request to share more than one EPR pair. However, there are limited entanglement resources over each edge. Consequently, an edge has two states. We call the first state is *satisfied state*. In this state, the edge possesses sufficient link-level entanglement resources to meet the requirements of multiple entanglement routing requests, i.e., $\sum_{i=1}^{k} D_i \leq C_{(u,v)}$, where D_i is the number of the required EPR pairs of *i*-th SD pair among *k* SD pairs on edge (u,v). In this state, the edge directly allocates EPR pairs to SD pairs as needed. The other state of the edge is *scarce state*. That is, the required EPR pairs exceed the capacity of the edge, i.e., $\sum_{i=1}^{k} D_i \geq C_{(u,v)}$. The edge in the *scarce state* is a bottleneck edge. The resource allocation on the bottleneck edge limits the number of SD pairs that can realize entanglement routing in a time slot. Here, we propose a CM scheme to complete the resource allocation and alleviate network congestion in Section III-B.

III. Algorithm Designs

The proposed entanglement routing design includes three algorithms, i.e., SPERF, ERA, and ERT. These algorithms are designed under the consideration of realistic quantum network models: arbitrary topology, concurrent entanglement routing requests, different numbers of entanglement resources required, and limited quantum memory resources. This section elaborates on how these three algorithms select a swapping path, allocate link-level entanglement resources, and tackle the bottleneck problem to achieve entanglement routing in quantum networks.

A. Path Selection

It is challenging to avoid network congestion by path selection in quantum networks where link-level EPR pairs have been pre-distributed between adjacent quantum nodes. Besides, the ERT algorithm designed in this paper can "recycle" idle entanglement resources to increase the capacities of bottleneck edges. In other words, the ERT algorithm can be used as a remedy to alleviate the network congestion on the selected swapping path. Hence, we do not focus on the design of routing algorithms for selecting the optimal swapping path. In this paper, our goal is to satisfy as many entanglement routing requests as possible. Hence, the selected path needs to perform well in the capability of end-to-end entanglement distribution, i.e., the success probability of entanglement routing. We design a routing algorithm to select the path with the higher success probability of entanglement routing as the swapping path. Concretely, the routing algorithm is designed based on the following two principles:

(1) The quantum node with a high success probability of entanglement swapping is preferred: Entanglement swapping is an imperfect operation, the success probability of which indicates the capability of a quantum repeater to extend the distance of entanglement distribution. In a nutshell, the higher the success probability of entanglement swapping, the easier it is to establish entanglement between the two distant quantum nodes. We note that the success probability of entanglement routing between an SD pair is proportional to the product of the success probability of entanglement swapping of all the intermediate quantum nodes on the swapping path. Hence, to create end-to-end entanglement more efficiently, the quantum node with the high-quality operation of entanglement swapping is generally preferred in entanglement routing.

(2) The swapping path with a smaller number of hops is preferred: The larger the hops of the swapping path, the more entanglement swapping is performed. As a result, the success probability of entanglement routing will decrease with the number of hops since the imperfection of entanglement swapping. Besides, more swapping operations will result in the additional consumption of entanglement resources and end-to-end entanglement distribution delay, which is not conducive to improving the performance of quantum networks. Hence, the path selection algorithm prefers to select the swapping path with fewer hops for each SD pair.

We define a new metric, called the success probability of end-to-end entanglement distribution, to quantify a swapping path. For a swapping path that spans *n* hops, the success probability of entanglement swapping of quantum node *u* is q_u, where $u \in (1, 2, \ldots, n)$. We get the success probability of entanglement routing between the source nodes *src* and the destination node *dst*:

$$Q_{\langle src,dst \rangle} = \prod_{u=1}^{n-1} q_u.$$

Obviously, two metrics, i.e., the number of hops and the success probability of entanglement swapping of each hop, together affect the success probability of end-to-end entanglement distribution. The path selection algorithm should maximize $Q_{\langle src,dst \rangle}$ as much as possible for all SD pairs. Here, we introduce the success probability of the entanglement routing first algorithm (SPERF) to select a swapping path for performing entanglement swapping.

In order to determine the optimal path with the highest success probability of entanglement routing, all the paths connecting each SD pair are required. However, traversing the network topology to obtain the set of paths connecting each SD pair results in extremely poor algorithm convergence. Moreover, when the hops of a path exceed a specific value, the success probability of entanglement routing is extremely low. As such, the path is not conducive to end-to-end entanglement distribution, and we can discard this path. Therefore, to reduce computational time, the SPERF algorithm first finds paths with a hops constraint, i.e., the hop of the swapping path is less than or equal to *k*. If we can find multiple paths with a hop number less than *k* between the source node and the destination node, the swapping path is the path with the highest success probability of entanglement routing in the path set. Otherwise, we take the end node of each path in this path set as the new "source" node to find the path connecting each SD pair using the minimum hop scheme. In this way, we can get a new path set, and the path with the highest success probability of entanglement routing in this path set is selected as the swapping path.

Algorithm 1: SPERF Algorithm

Input: $G = \langle V, E, C\rangle$, $\langle src, dst\rangle$;
Output: The swapping path $P^{swap}_{\langle src,dst\rangle}$;

1 $P_{\langle src,dst\rangle} \leftarrow$ the set of paths from *src* to *dst*;
2 $q, p \leftarrow$ two stacks of *n* elements, all set to null;
3 $visited \leftarrow$ an array of *n* elements, all set to false;
4 q.pushstack(*src*);
5 Push all neighbor nodes of *src* onto p;
6 **while** q *is not empty* **do**
7 $hops = 0$;
8 $i \leftarrow p$.popstack();
9 **if** *visited[i]* **then**
10 **continue**;
11 **end**
12 **else**
13 $visited[i] \leftarrow$ true;
14 $hops = hops + 1$;
15 **end**
16 $List \leftarrow$ construct the list of adjacent nodes of *i*;
17 **if** *List is not empty* **then**
18 $u \leftarrow$ the first element of *List*;
19 q.pushstack (u);
20 $visited[u] \leftarrow$ true;
21 $hops = hops + 1$;
22 **if** $hops \leq k$ **then**
23 Push the rest elements of *List* in p;
24 **end**
25 **end**
26 **else**
27 q and p push back until the top of p is not null;
28 **end**
29 **if** *the top of q is dst* **then**
30 $P_{\langle src,dst\rangle} \leftarrow$construct a path from *src* to *dst*;
31 **end**
32 **else**
33 set the top of q as "*src*" node;
34 find path using the minimum hop scheme;
35 $P_{\langle src,dst\rangle} \leftarrow$ update the path set;
36 **end**
37 **end**
38 $P^{swap}_{\langle src,dst\rangle} \leftarrow$ the path with the highest probability;

Formally, the SPERF algorithm includes two steps. The first step aims to find paths for each SD pair. When the centralized processor receives the entanglement routing requests from SD pairs, it adopts a deep-first search to traverse the network topology $G = \langle V, E, C\rangle$ (Lines 6-36). In this process, Lines 6-30 attempt to find paths with hops less than *k*. If no path can connect each SD pair with hops less than *k*, the minimum hop scheme is adopted to update the path set (Lines 32-36). As a result, we get a set of paths from the source node to the destination node, $P_{\langle src,dst\rangle}$. The second step aims to select a swapping path from $P_{\langle src,dst\rangle}$ with the highest success probability of entanglement routing instead of the minimum hop counts (Line 33). Assisted by the centralized processor, the network system completes the computation of $Q_{\langle src,dst\rangle}$ on each selected path according to the configuration of each quantum node. Then, the results are sorted, and the path with the highest success probability of end-to-end entanglement distribution is selected as the swapping path.

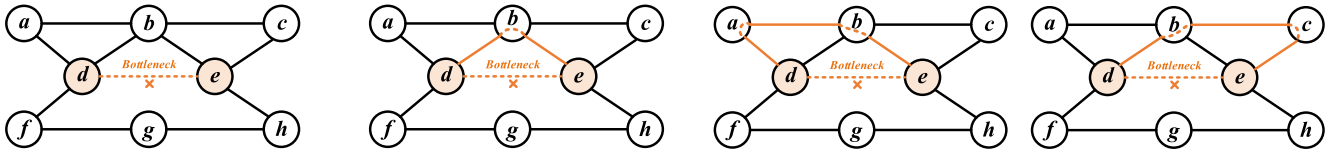
B. Resource Allocation and Congestion Mitigation

Resource allocation aims to mitigate congestion caused by entanglement resource competition and satisfy as many SD pairs' demands as possible in one time slot. Here, we propose the CM scheme, a two-step entanglement resource allocation scheme. The first step of CM is to complete the entanglement resource pre-allocation to meet the demands of SD pairs as much as possible, which is implemented by the ERA algorithm. Then the ERT algorithm is designed to tackle the bottleneck problem by "recycling" the idle entanglement resources to improve the capacity of the bottleneck edge.

(1) Entanglement Resource Allocation: There would be a mismatch between the resource demands of concurrent entanglement routing requests and the limited entanglement resources over some edges, and we call such edges in the scarce state, i.e., each edge is a bottleneck edge. It is challenging for a bottleneck edge to assign insufficient entanglement resources to multiple requests. A poor allocation scheme will further block a larger number of entanglement routing requests. Hence, we need to design an effective algorithm to tackle the problem of resource allocation on the bottleneck edge to improve the number of satisfied entanglement routing requests.

There may be a greedy algorithm: each bottleneck edge preferentially attempts to assign link-level entanglement resources to the SD pair with a low requirement of entanglement resources. In this way, the number, s_i, of the SD pair whose requirement is satisfied on the *i*-th bottleneck edge is maximized. Most notably, the request of each SD pair is satisfied only if all edges on the selected swapping path allocate sufficient entanglement resources to it simultaneously. Hence, we cannot equate the sum of s_i on all bottleneck edges with the number of satisfied SD pairs in the network topology. The reason is that there is a situation where a selected path has multiple bottleneck edges, and not all bottleneck edges can meet the demand for entanglement resources. The greedy algorithm mistakenly assumes that the SD pair satisfied on one bottleneck edge is contented on all edges. Therefore, local optimality on one bottleneck edge cannot be regarded as global optimality in the network topology for the greedy resource allocation algorithm.

To tackle the problem of resource allocation on the bottleneck edge, we introduce the ERA algorithm as shown in Algorithm 2. ERA aims to maximize the total number of satisfied entanglement routing requests. The core idea of ERA is to preferentially allocate entanglement resources to the SD pairs with the least demand for entanglement resources. The ERA algorithm first orders all unsatisfied SD pairs in increasing order of their demands (Line 1). Then, a judgment is performed on all bottleneck edges to determine if the



(a) Bottleneck edge (d, e) lacks one shared EPR pair. (b) Entanglement resource transformation via path $\{d, b, e\}$. (c) Entanglement resource transformation via path $\{d, a, b, e\}$. (d) Entanglement resource transformation via path $\{d, b, c, e\}$.

Fig. 4. In the general network topology, swapping operations can be performed on different paths to increase the capacity of the bottleneck edge.

Algorithm 2: Entanglement Resource Allocation

Input: L : the set of all bottleneck edge
R : the set of all unsatisfied SD pairs;
D : the set of all SD pairs demands
Output: Allocation scheme on the bottleneck edge;

```
1  Sort R in ascending order based on each SD pair's demand;
2  satisfied ← an array of n elements, all set to false;
3  for All request R_j do
4      for All bottleneck edge L_(u,v) do
5          if R_j in edge L_(u,v) then
6              if C_(u,v) > D_j then
7                  satisfied[R_j] ← true;
8              end
9              else
10                 satisfied[R_j] ← false;
11             end
12         end
13     end
14     if satisfied[R_j] then
15         Allocate entanglements to R_j;
16         Update C_(u,v) of the bottleneck edge;
17     end
18 end
```

resource meets the demands of SD pairs (Lines 3-13). The ERA algorithm only attempts to assign entanglement resources to the SD pair whose demand can be satisfied on all selected edges (Lines 14-17). As long as one bottleneck edge fails to meet the demand of an SD pair, all bottleneck edges will not allocate entanglement resources to the SD pair.

(2) Solve the Bottleneck Problem: The bottleneck problem significantly affects the performance of quantum networks. Fortunately, we can perform swapping operations to increase the capacity of the bottleneck edge since entanglement swapping can transform the entanglement relationship between quantum nodes. In a nutshell, the state of an edge can be switched from the scarce state to the satisfied state with the assistance of entanglement swapping. Consequently, the idle entanglement resources can be "recycled" to solve the bottleneck problem, thus improving the request service capability of quantum networks.

The basic idea of the ERT algorithm is to find paths with sufficient idle entanglement resources to perform entanglement swapping. For example, as shown in Fig. 4, three paths can be selected to supplement entanglement resources for the bottleneck edge (d, e) who lacks one EPR pair (Fig. 4(a)). For the first path, entanglement swapping is performed on intermediate node b to generate an additional link-level entanglement between d and e (Fig. 4(b)). For other two paths, $d \rightarrow a \rightarrow b \rightarrow e$ shown in Fig. 4(c) and $d \rightarrow b \rightarrow c \rightarrow e$ shown in Fig. 4(d), swapping operations need to be performed two times. Considering the success probability of entanglement swapping and the variability in the number of idle EPR pairs on different paths, there are different performances in the number of the satisfied entanglement routing requests by generating supplementary entanglement through these three paths. Here, we introduce the ERT algorithm to construct a path efficiently tackling the bottleneck problem.

Suppose that there is a bottleneck edge $L_{(u,v)}$, and the number of EPR pairs that $L_{(u,v)}$ lacks is denoted as $Miss_{(u,v)}$. The ERT algorithm aims to construct a path (called ERT path) connecting u and v in the revised topology $G' = \langle V, E', C' \rangle$, where E' is the set of remaining edges after cutting off the bottleneck edges, and C' is the set of the remaining capacities of all edges after ERA. The minimum number of remaining EPR pairs on the selected path must be greater than $Miss_{(u,v)}$. Note that the original path we select may include the edge (i, j) whose entanglement resources cannot meet the requirement of the bottleneck edge, i.e., $C_{(i,j)} < Miss_{(u,v)}$. For example, we assume that $Miss_{(d,e)}$ is equal to 2, and the original ERT path $d \rightarrow b \rightarrow e$ is selected to increase $C_{(d,e)}$ in Fig. 4(a). However, the capacity of the edge (b, e) is less than $Miss_{(d,e)}$. Here, we consider (b, e) as the new bottleneck edge to discover the appended path $b \rightarrow c \rightarrow e$. Hence, the construction of the ERT path is an iterative process. To avoid endless searches, we limit the hops of the ERT path. Besides, the ERT algorithm selects the path with the highest success probability of entanglement routing as the optimal ERT path.

The ERT algorithm is shown in Algorithm 3. The first step of the ERT algorithm is to build the set of all paths connecting u and v by the improved DFS. If the path is only one hop and the link-level entanglement resources are sufficient, the bottleneck edge is resolvable (Lines 6-9). For each multi-hop path, it is necessary to judge whether the entanglement resources of all edges meet the demand. Lines 11-13 are used to count the number of edges that are in the satisfied state. If the capacity of each edge $L_{(j,k)}$ on the path is greater than the requirement of the bottleneck edge, i.e., $C_{(j,k)} \geq Miss_{(u,v)}$, the path can be used to tackle the bottleneck problem (Lines 11-13, Lines 21-23). On the contrary, the edges that cannot satisfy the requirement of entanglement

Algorithm 3: Entanglement Resource Transformation

Input: $G' = \langle V, E', C' \rangle, L_{(u,v)}$: the bottleneck edge;
Output: The set of optimal path for $P^{opt}_{(u,v)}$;

1 $satisfied \leftarrow$ an array of n elements, all set to false;
2 $P_{(u,v)} \leftarrow$ paths connecting u and v within h hops;
3 **if** $P_{(u,v)}$ *is not empty* **then**
4 **for** *All path* $P^i_{((u,v)}$ **do**
5 $m \leftarrow$ the number of hops of $P^i_{(u,v)}$;
6 **if** $m = 1$ *and* $C_{(n,v)} \geq Miss_{(u,v)}$ **then**
7 $satisifed[L_{(u,v)}] \leftarrow$ true;
8 **end**
9 **for** *All edge* $L_{(j,k)}$ **do**
10 **if** $C_{(j,k)} \geq Miss_{(u,v)}$ **then**
11 $l \leftarrow l + 1$;
12 **end**
13 **else**
14 Delete the edges in the topology;
15 $P_{(j,k)} \leftarrow$ paths with $(h - m)$ hops;
16 Add $P_{(j,k)}$ into $P_{(u,v)}$;
17 $Miss_{(j,k)} \leftarrow Miss_{(u,v)} - C_{(j,k)}$;
18 **end**
19 **end**
20 **if** $l = m + 1$ **then**
21 $satisifed[L_{(u,v)}] \leftarrow$ true;
22 **end**
23 **else**
24 Remove $P^i_{(u,v)}$ from $P_{(u,v)}$;
25 **end**
26 **end**
27 **end**
28 **else**
29 $satisfied[L_{(u,v)}] \leftarrow$ false;
30 **end**
31 **if** $satisfied[L_{(u,v)}]$ **then**
32 $P^{opt}_{(u,v)} \leftarrow$ the path with highest probability;
33 **end**

are treated as new bottleneck edges. In this situation, we need to update the network topology and the requirement of entanglement resources of the new bottleneck edge (Lines 14-18). Besides, the path that does not possess sufficient entanglement resources needs to be removed from the path set (Lines 24-26). Finally, the path with the highest success probability of entanglement routing in the path set meeting the entanglement requirement of the bottleneck edge $L_{(u,v)}$ is chosen as the optimal ERT path (Lines 32-34).

C. Implementation

The quantum internet uses the fundamental concepts of quantum mechanics for networking numerous quantum processors to support ground-breaking quantum applications [55]. The main function of the quantum Internet is to distribute EPR pairs between distant quantum end nodes, which is also the focus of this paper. Here, we describe the implementation of the swapping-based entanglement routing design in a practical scenario of quantum networks, i.e., multiple SD pairs request to share EPR pairs for teleportation-based quantum communication in each time slot [56], [57]. After the first and second phases, the centralized processor receives the entanglement routing requests from the SD pairs and then executes the SPERF algorithm to select the swapping path for each SD pair in P3. Generally, the topology of a quantum network is stable [58], i.e., the set of nodes and edges of a quantum network does not change over several consecutive time slots. Hence, the results of path selection obtained in each time slot can be cached in the centralized processor. In this way, the cached paths can be used directly when some SD pairs initiate entanglement routing requests again in the subsequent time slots, thus reducing the computational overhead of path selection. Besides, some classical techniques, e.g., machine learning, can realize accurate predictions based on historical information. Hence, the SPERF algorithm can also be executed in the first phase of a time slot to pre-select the swapping path for the possible SD pairs in advance to reduce the computational overhead.

ERA and ERT algorithms cooperate to realize resource allocation in P3. After path selection, the ERA algorithm is first performed to allocate link-level entanglement sources for requests that can be satisfied along the initial swapping path. After ERA, the initial topology G is revised to G' where only the set of nodes is the same as G. The difference between E in G and E' in G' is the set of bottleneck edges, and each edge capacity in C' equals the difference between its initial capacity in C and the number of EPR pairs allocated to the request. Then the ERT algorithm is performed to select the ERT path for boosting the capacity of the bottleneck edge in the revised topology G'. After ERT, the centralized processor reruns the ERA algorithm to allocate the supplementary entanglement resources for the unsatisfied SD pairs.

After resource allocation and congestion mitigation, it is clear which entanglement routing requests can be satisfied. For the satisfied requests, all the intermediate quantum nodes on the selected swapping path simultaneously perform swapping operations to generate end-to-end entanglement, and then each SD pair performs teleportation to realize the transmission of quantum information. Notably, although the ERT algorithm can boost the capacity of the bottleneck edge by "recycling" the idle entanglement resources, there are still some entanglement routing requests that will be blocked in a time slot. For the blocked requests, we queue them to a request pool with a processing priority at the next time slot. The priority of each blocked request can be determined by its demand for entanglement resources and waiting time. In this way, the blocked requests have a higher priority to be processed in the next slot. Hence, the problem of "starvation", i.e., some requests will wait indefinitely before being processed, can be effectively avoided.

D. Discussion and Complexity Analysis

In this work, we define the path with the highest success probability of entanglement routing as the optimal swapping path for end-to-end entanglement distribution. To gain the

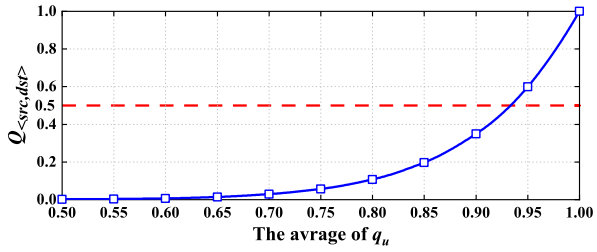


Fig. 5. $Q_{\langle src,dst\rangle}$ vs the average q_u when the hop count equals 10.

optimal swapping path, we must obtain the set of all paths that connect each SD pair, which does harm to the convergence of the path selection algorithm, especially in a large-scale quantum network. Fortunately, we find that the success probability of entanglement routing will be less than 0.5 when the hops of the selected path are greater than 10, even though the average success probability of entanglement swapping of each quantum node is as high as 0.9. In other words, when the hops exceed 10, the success probability of entanglement swapping has little effect on the success probability of entanglement routing. Therefore, to quickly obtain the swapping path, we adopt a cut-off solution, i.e., k in the SPERF algorithm can be set to 10. Moreover, the hops of the ERT path are limited to less than 5, i.e., $h \leq 5$ in the ERT algorithm. In this, the central processor can select the swapping path for each SD pair in a relatively short time.

For the exhaustive search algorithm, the time cost of finding a swapping path with the highest success probability of entanglement routing under the limited condition that the number of hops of the path is less than 10 in $G = \langle V, E, C\rangle$ is $O(|V|!)$. However, the time cost of the DFS-based SPERF algorithm is $O(|V|^2)$. When the scale of quantum networks is large, the SPERF algorithm can find the optimal path faster than the exhaustive search algorithm. Considering n SD pairs and m bottleneck edges in a time slot, each SD pair needs to check whether a bottleneck edge meets the requirement of entanglement connections. Hence, the time cost of the ERA algorithm is $O(mn)$. Although the ERA algorithm has the same time cost as the greedy allocation algorithm, it outperforms the greedy algorithm in the number of satisfied SD pairs. The time cost of the ERT algorithm is $O(|V|^2)$ because of utilizing the DFS-based strategy. ERT algorithm also outperforms the exhaustive search algorithm on computational complexity.

IV. Performance Evaluation

In this section, we perform extensive simulations to evaluate the performance of our entanglement routing design. Simulations involve randomly generated networks with a certain number of node and edge resources, a set of SD pairs, and a series of requests for generating a different number of EPR pairs. We show the averaged results of multiple simulations based on a given set of parameters.

A. Evaluation Methodology

Comparison Schemes: We compare the SPERF algorithm with the other two general schemes under different network scales and with a different success probability of entanglement swapping. One is the minimum hop algorithm (referred to as Min-Hops). The other is a greedy algorithm (referred to as Greedy) which always selects the quantum nodes with the highest success probability of entanglement swapping to construct a swapping path between each SD pair. Besides, we compare the ERA algorithm with two strategies, i.e., the random allocation scheme (referred to as Random) and the greedy algorithm that preferentially assigns entanglement resources to SD pairs with the lower requirement of entanglement resources on each edge. Finally, the scheme using the ERT algorithm and the scheme without ERT are compared based on the ERA algorithm.

Performance Metrics: We compare the performance of different path selection algorithms concerning one metric, i.e., the success probability of entanglement routing. The success probability indirectly represents the number of EPR pairs shared between SD pairs in a time slot. Besides, we compare the performances of different resource allocation schemes for the number of satisfied SD pairs in a time slot. The number of satisfied SD pairs is defined as the entanglement resource allocation scheme's capability to tackle the bottleneck problem. A larger value means better performance in mitigating network congestion.

B. Evaluation Results

(1) Main Observations: From our simulations, we observe that the path selection algorithm we proposed outperforms Min-hops and Greedy in terms of the success probability of entanglement routing. Nonetheless, the superiority of the SPERF algorithm is not apparent when the network scale and the success probability of entanglement swapping are high since the number of hops of the swapping path is too larger. Besides, the ERA algorithm performs better than Random and Greedy in terms of the number of satisfied SD pairs in one time slot, especially in scenarios with high concurrent requests and high demand for entanglement resources. The ERT algorithm can effectively mitigate congestion to improve the number of satisfied SD pairs.

(2) Path Selection Algorithm: To evaluate the performance of three different path selection algorithms, we vary the number of quantum nodes from 100 to 500 in the network topology, and the success probability of entanglement swapping for each quantum node is randomly generated from a range from 0.5 to 1.0. The results of repeated simulation show that the SPERF algorithm is conducive to selecting the swapping path with a high success probability of entanglement routing in quantum networks.

Effect of network scale: Fig. 6 shows the success probability of establishing end-to-end entanglement decreases as the number of network nodes increases for three routing algorithms. The success probability of the swapping path selected by the Greedy algorithm gradually approaches 0 when the number of quantum nodes exceeds 400. The degradation rate of the SPERF and the Min-Hops algorithms increases, and the difference between the Min-Hops algorithm and the SPERF algorithm in the performance of the success probability

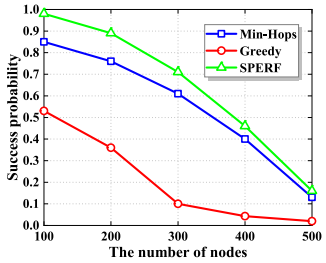


Fig. 6. Effect of the number of nodes in a quantum network.

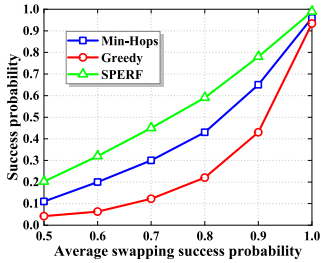


Fig. 7. Effect of the success probability of entanglement swapping.

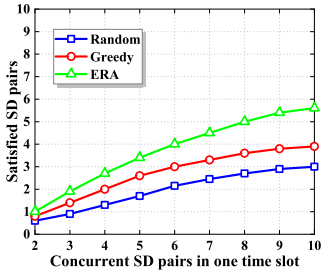


Fig. 8. Effect of the number of concurrent SD pairs in one time slot.

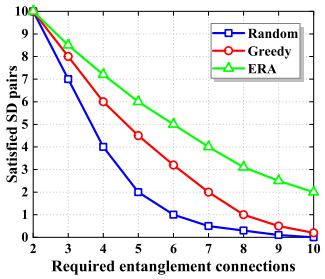


Fig. 9. Effect of the number of the required entanglement resources.

becomes smaller as the network size continues to increase. The reason is that the number of quantum nodes in the selected path increases as the network sizes increase and the SPERF tends to choose the path with a smaller hop count to guarantee the success probability. Generally, the SPERF algorithm performs well on the success probability of entanglement routing than the Greedy and the Min-Hops as the network grows in size.

Effect of entanglement swapping success probability: We evaluate the performance of three algorithms with different average swapping success probability when the number of nodes is fixed at 100. In Fig. 7, we can see that the success probability of entanglement routing increases with the average swapping success probability, and the SPERF is consistently better than the Min-Hops and the Greedy in the success probability. When the swapping success probability is about equal to 1.0, the performance of the three path selection algorithms sharply increases and tends to be the same. This is because we can regard entanglement swapping as a perfect operation in this case, i.e., end-to-end entanglement can be built with a 100 percent success probability through any path.

(3) Entanglement Resource Allocation: We compare the performance of three different resource allocation schemes in the scenario: the number of concurrent SD pairs is 10, the number of required entanglement resources of each SD pair is randomly generated between 2 and 10, multiple bottleneck edges, and different edge capacities. Simulation results show that the ERA algorithm can significantly improve request service capability in a quantum network with limited quantum memory resources.

Effect of concurrent SD pairs: When the number of SD pairs in the network varies from 2 to 10, the change in the number of the satisfied SD pairs under different resource allocation schemes is shown in Fig. 8. We can see that the number of satisfied SD pairs first increases with the number of SD pairs. When the number of SD pairs continues to increase, the increase rate slightly decreases due to the resource contention among different SD pairs. Besides, we can also observe in Fig. 8 that the advantage of the ERA algorithm over the Random and the Greedy is gradually expanded as the concurrent SD pairs increase.

Effect of required entanglement resources: To investigate how the average number of required entanglement connections how affects the performance of the ERA, we pick a value from the set {2, 3, 4, 5, 6, 7, 8, 9, 10} to be the independent variable and set the edge capacity to 20. The greater the number of required entanglement resources, the more intense the resource competition, i.e., the more bottleneck edges in a time slot. Accordingly, the number of the satisfied SD pairs decreases with the number of the average required entanglement resources for all three schemes, as shown in Fig. 9. We can also observe that the ERA algorithm's performance degrades at a slower pace than the Random and the Greedy. This is because both Random and Greedy would cause an SD pair to fail to possess sufficient entanglement resources among all selected edges, and the probability of failure increases as the number of the required entanglement resources increases.

Effect of bottleneck edge: We vary the number of bottleneck edges on each selected path from 1 to 10 to compare three schemes. The more bottleneck edges, the lower the probability that Random and Greedy enable the requirement of each SD pair to be satisfied among all selected edges. Hence, the number of satisfied SD pairs decreases with the number of bottleneck edges, as shown in Fig. 10, and the performance of the ERA decreases at a slower rate than the other two schemes. When the number of bottleneck edges surpasses a certain value, the Random and the Greedy tend to stabilize at an extremely low number of satisfied SD pairs.

Effect of edge capacity: We evaluate the performance of three resource allocation schemes by varying the average capacity of each edge from 5 to 30. Fig. 11 shows how the number of satisfied SD pairs changes with the capacity of each

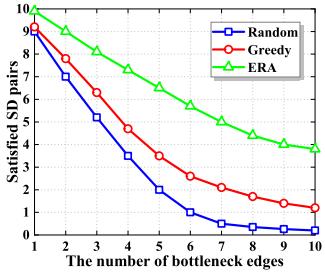


Fig. 10. Effect of the number of bottleneck edges.

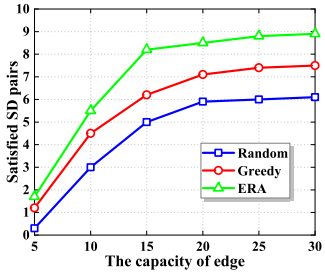
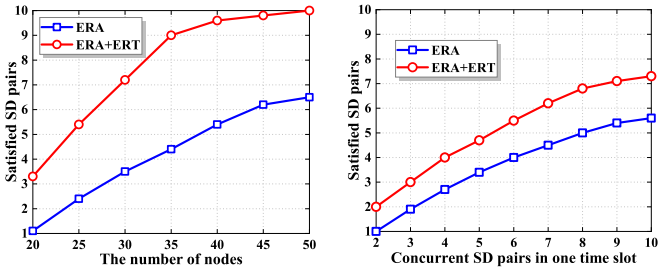


Fig. 11. Effect of the capacity of each edge.



(a) Satisfied SD pairs VS the number of concurrent SD pairs.

(b) Satisfied SD pairs VS network sizes.

Fig. 12. The validity of the ERT algorithm.

edge. Larger edge capacity in quantum networks means less resource contention between concurrent entanglement routing in a time slot. Accordingly, the number of satisfied SD pairs would increase as edge capacity increases for all three schemes. We can also see that the increase rate of the ERA is greater than the other two schemes due to the Random and the Greedy cannot ensure that all selected edges assign sufficient entanglement resources to each SD pair. When this capacity exceeds a threshold set to 20 in our simulations, the number of satisfied SD pairs cannot be further improved. This is because the size of a quantum memory becomes the bottleneck.

(4) Entanglement Resource Transformation: To evaluate the ERT algorithm, we compare the performance of the scheme adopting both ERA and ERT algorithms and the scheme using the ERA algorithm only. Fig. 12 shows the performance of two schemes to solve the bottleneck problem with different network scales and concurrent SD pairs. As expected, the ERT algorithm can solve the bottleneck problem in the scenario of high concurrent entanglement routing between multiple SD pairs (Fig. 12). The advantages of the ERT algorithm increase with network scale since more idle entanglement resources can be utilized to improve the capacity of the bottleneck edge (Fig 12(a)). Moreover, the effect of the ERT algorithm on mitigating congestion increases with the number of concurrent SD pairs due to the competition for entanglement resources intensifying (Fig 12(b)).

V. Conclusion

In this paper, we proposed a swapping-based entanglement routing design for establishing end-to-end entanglement between multiple SD pairs simultaneously in quantum networks. This design fully considers two unique properties of the swapping operation, i.e., entanglement swapping is an imperfect operation, and the swapping operation can transform entanglement relationships between quantum nodes. To reduce the negative effect of the imperfect swapping operations, we introduced the SPERF algorithm, which presents a new metric, the success probability of entanglement routing, to evaluate the selected path. In order to mitigate congestion, the two-step scheme CM is proposed. The ERA algorithm first achieves entanglement resource allocation on the bottleneck edges, and then the ERT algorithm solves the bottleneck problem by "recycling" idle entanglement resources. Extensive simulations have shown that our entanglement routing design can significantly improve the rate of end-to-end entanglement distribution than the min-hop and greedy routing schemes. In addition, CM can effectively mitigate network congestion and increase the number of satisfied SD pairs in a time slot. In this work, we mainly consider the impact of the hops of the selected path and the success probability of entanglement swapping on the remote entanglement distribution rate. However, due to quantum decoherence, entanglement fidelity is also a vital factor affecting the performance of entanglement routing design in quantum networks. In future works, we plan to improve the entanglement routing design by developing fidelity-based algorithms.

References

[1] H. J. Kimble, "The quantum Internet," *Nature*, vol. 453, no. 7198, pp. 1023–1030, 2008.
[2] S. Pirandola and S. L. Braunstein, "Physics: Unite to build a quantum Internet," *Nature*, vol. 532, no. 7598, pp. 169–171, 2016.
[3] S. Wehner, D. Elkouss, and R. Hanson, "Quantum Internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, p. 9288, 2018.
[4] D. Castelvecchi, "The quantum Internet has arrived (and it hasn't)," *Nature*, vol. 554, no. 7690, pp. 289–292, 2018.
[5] H.-K. Lo and H. F. Chau, "Unconditional security of quantum key distribution over arbitrarily long distances," *Science*, vol. 283, no. 5410, pp. 2050–2056, 1999.
[6] J. I. Cirac, A. K. Ekert, S. F. Huelga, and C. Macchiavello, "Distributed quantum computation over noisy channels," *Phys. Rev. A*, vol. 59, pp. 4249–4254, Jun. 1999.
[7] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, Aug. 2018.
[8] L. Gyongyosi and S. Imre, "A survey on quantum computing technology," *Comput. Sci. Rev.*, vol. 31, pp. 51–71, Feb. 2019.
[9] C. L. Degen, F. Reinhard, and P. Cappellaro, "Quantum sensing," *Rev. Mod. Phys.*, vol. 89, Jul. 2017, Art. no. 35002.
[10] S. Pirandola, B. R. Bardhan, T. Gehring, C. Weedbrook, and S. Lloyd, "Advances in photonic quantum sensing," *Nat. Photon.*, vol. 12, pp. 724–733, Nov. 2018.
[11] H. Takesue and K. Inoue, "Generation of polarization-entangled photon pairs and violation of Bell's inequality using spontaneous four-wave mixing in a fiber loop," *Phys. Rev. A*, vol. 70, no. 3, 2004, Art. no. 31802.

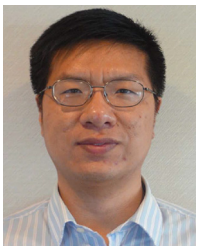
[12] P. C. Humphreys et al., "Deterministic delivery of remote entanglement on a quantum network," *Nature*, vol. 558, no. 7709, pp. 268–273, 2018.
[13] A. Albrecht, "Investigating decoherence in a simple system," *Phys. Rev. D*, vol. 46, pp. 5504–5520, Dec. 1992.
[14] W. H. Zurek, "Decoherence, einselection, and the quantum origins of the classical," *Rev. Mod. Phys.*, vol. 75, no. 3, p. 715, 2003.
[15] W. K. Wootters and W. H. Zurek, "A single quantum cannot be cloned," *Nature*, vol. 299, pp. 802–803, Oct. 1982.
[16] J.-W. Pan, D. Bouwmeester, H. Weinfurter, and A. Zeilinger, "Experimental entanglement swapping: Entangling photons that never interacted," *Phys. Rev. Lett.*, vol. 80, pp. 3891–3894, May 1998.
[17] 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, p. 5932, Dec. 1998.
[18] D. Gottesman, T. Jennewein, and S. Croke, "Longer-baseline telescopes using quantum repeaters," *Phys. Rev. Lett.*, vol. 109, Aug. 2012, Art. no. 70503.
[19] M. Caleffi, "Optimal routing for quantum networks," *IEEE Access*, vol. 5, pp. 22299–22312, 2017.
[20] E. Shchukin, F. Schmidt, and P. van Loock, "Waiting time in quantum repeaters with probabilistic entanglement swapping," *Phys. Rev. A*, vol. 100, Sep. 2019, Art. no. 32322.
[21] T. Gorin, C. Pineda, and T. H. Seligman, "Decoherence of an *n*-qubit quantum memory," *Phys. Rev. Lett.*, vol. 99, no. 24, 2007, Art. no. 240405.
[22] A. Dahlberg et al., "A link layer protocol for quantum networks," in *Proc. Spec. Interest Group Data Commun. Appl. Technol. Archit. Protocols Comput. Commun. (SIGCOMM)*, 2019, pp. 159–173.
[23] R. Van Meter, T. Satoh, T. D. Ladd, W. J. Munro, and K. Nemoto, "Path selection for quantum repeater networks," *Netw. Sci.*, vol. 3, no. 1, pp. 82–95, 2013.
[24] S. Pirandola, "End-to-end capacities of a quantum communication network," *Commun. Phys.*, vol. 2, no. 1, pp. 1–10, 2019.
[25] M. Pant et al., "Routing entanglement in the quantum Internet," *npj Quantum Inf.*, vol. 5, pp. 25–34, Mar. 2019.
[26] S. Shi and C. Qian, "Concurrent entanglement routing for quantum networks: Model and designs," in *Proc. Special Interest Group Data Commun. Appl. Technol. Archit. Protocols Comput. Commun. (SIGCOMM)*, 2020, pp. 62–75.
[27] E. Schoute, L. Mancinska, T. Islam, I. Kerenidis, and S. Wehner, "Shortcuts to quantum network routing," 2016, *arXiv.1610.05238*.
[28] S. Das, S. Khatri, and J. P. Dowling, "Robust quantum network architectures and topologies for entanglement distribution," *Phys. Rev. A*, vol. 97, no. 1, 2018, Art. no. 12335.
[29] 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, pp. 1–10, 2016.
[30] K. Azuma and G. Kato, "Aggregating quantum repeaters for the quantum Internet," *Phys. Rev. A*, vol. 96, Sep. 2017, Art. no. 32332.
[31] J. Li, Q. Jia, K. Xue, D. S. L. Wei, and N. Yu, "A connection-oriented entanglement distribution design in quantum networks," *IEEE Trans. Quantum Eng.*, vol. 3, pp. 1–13, May 2022.
[32] H.-K. Lo, X. Ma, and K. Chen, "Decoy state quantum key distribution," *Phys. Rev. A*, vol. 94, Jun. 2005, Art. no. 230504.
[33] H. Bernien et al., "Heralded entanglement between solid-state qubits separated by three metres," *Nature*, vol. 497, pp. 86–90, Apr. 2013.
[34] Y. Yu et al., "Entanglement of two quantum memories via fibres over dozens of kilometres," *Nature*, vol. 578, no. 7794, pp. 240–245, 2020.
[35] C. Jones, D. Kim, M. T. Rakher, P. G. Kwiat, and T. D. Ladd, "Design and analysis of communication protocols for quantum repeater networks," *New J. Phys.*, vol. 18, no. 8, 2016, Art. no. 83015.
[36] 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, Oct. 2021.
[37] J. Illiano, M. Caleffi, A. Manzalini, and A. S. Cacciapuoti, "Quantum Internet protocol stack: A comprehensive survey," *Comput. Netw.*, vol. 213, Aug. 2022, Art. no. 109092.
[38] C. Li, T. Li, Y.-X. Liu, and P. Cappellaro, "Effective routing design for remote entanglement generation on quantum networks," *npj Quantum Inf.*, vol. 7, no. 1, pp. 1–12, 2021.
[39] K. Chakraborty, F. Rozpedek, A. Dahlberg, and S. Wehner, "Distributed routing in a quantum Internet," 2019, *arXiv:1907.11630*.
[40] Z. Li, K. Xue, J. Li, N. Yu, D. S. L. Wei, and R. Li, "Connection-oriented and connectionless remote entanglement distribution strategies in quantum networks," *IEEE Netw.*, vol. 36, no. 6, pp. 150–156, Nov./Dec. 2022.
[41] M. Caleffi, D. Chandra, D. Cuomo, S. Hassanpour, and A. S. Cacciapuoti, "The rise of the quantum Internet," *Computer*, vol. 53, no. 6, pp. 67–72, Jun. 2020.
[42] W. Dai, T. Peng, and M. Z. Win, "Quantum queuing delay," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 3, pp. 605–618, Mar. 2020.
[43] P.-Y. Li et al., "Hyperfine structure and coherent dynamics of rare-Earth spins explored with electron-nuclear double resonance at subkelvin temperatures," *Phys. Rev. A*, vol. 13, Feb. 2020, Art. no. 24080.
[44] L. Gyongyosi, S. Imre, and H. V. Nguyen, "A survey on quantum channel capacities," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1149–1205, 2nd Quart., 2018.
[45] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, "Fundamental limits of repeaterless quantum communications," *Nat. Commun.*, vol. 8, no. 1, pp. 1–15, 2017.
[46] R. V. Meter and J. Touch, "Designing quantum repeater networks," *IEEE Commun. Mag.*, vol. 51, no. 8, pp. 64–71, Aug. 2013.
[47] A. I. Lvovsky, B. C. Sanders, and W. Tittel, "Optical quantum memory," *Nat. Photon.*, vol. 3, no. 12, pp. 706–714, 2009.
[48] J.-S. Tang et al., "Storage of multiple single-photon pulses emitted from a quantum dot in a solid-state quantum memory," *Nat. Commun.*, vol. 6, no. 1, pp. 1–7, 2015.
[49] J.-P. Dou et al., "A broadband DLCZ quantum memory in room-temperature atoms," *Commun. Phys.*, vol. 1, no. 1, pp. 1–7, 2018.
[50] Y.-L. Hua, Z.-Q. Zhou, C.-F. Li, and G.-C. Guo, "Quantum light storage in rare-Earth-ion-doped solids," *Chin. Phys. B*, vol. 27, no. 2, 2018, Art. no. 20303.
[51] C. Liu et al., "On-demand quantum storage of photonic qubits in an on-chip waveguide," *Phys. Rev. Lett.*, vol. 125, Dec. 2020, Art. no. 260504.
[52] Y.-F. Pu, N. Jiang, W. Chang, H.-X. Yang, C. Li, and L.-M. Duan, "Experimental realization of a multiplexed quantum memory with 225 individually accessible memory cells," *Nat. Commun.*, vol. 8, pp. 1–6, May 2017.
[53] G. Yang, L. Xing, M. Nie, Y.-H. Liu, and M.-L. Zhang, "Hierarchical simultaneous entanglement swapping for multi-hop quantum communication based on multi-particle entangled states," *Chin. Phys. B*, vol. 30, no. 3, 2021, Art. no. 30301.
[54] L. Gyongyosi and S. Imre, "Decentralized base-graph routing for the quantum Internet," *Phys. Rev. A*, vol. 98, no. 2, 2018, Art. no. 22310.
[55] L. Gyongyosi and S. Imre, "Advances in the quantum Internet," *Commun. ACM*, vol. 65, no. 8, pp. 52–63, 2022.
[56] L. Vaidman, "Teleportation of quantum states," *Phys. Rev. A*, vol. 49, no. 2, p. 1473, 1994.
[57] S. Pirandola, J. Eisert, C. Weedbrook, A. Furusawa, and S. L. Braunstein, "Advances in quantum teleportation," *Nat. Photon.*, vol. 9, no. 10, pp. 641–652, 2015.
[58] S. Perseguers, G. J. Lapeyre Jr., D. Cavalcanti, M. Lewenstein, and A. Acín, "Distribution of entanglement in large-scale quantum networks," *Rep. Progr. Phys.*, vol. 76, no. 9, 2013, Art. no. 96001.

Zhonghui Li (Graduate Student Member, IEEE) received the B.E. degree in software engineering from the University of Electronic Science and Technology of China in 2018. He is currently pursuing the Ph.D. degree in information security with the School of Cyber Science and Technology, University of Science and Technology of China. His current research interests include quantum Internet and quantum networks.

Jian Li (Member, IEEE) received the B.S. degree from the Department of Electronics and Information Engineering, Anhui University in 2015, and the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2020. From November 2019 to November 2020, he was a Visiting Scholar with the Department of Electronic and Computer Engineering, University of Florida. He is currently a Postdoctoral Fellow with the School of Cyber Science and Technology, USTC. His research interests include wireless networks, next-generation Internet architecture, and quantum networks.

Kaiping Xue (Senior Member, IEEE) received the bachelor's degree from the Department of Information Security and the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2003 and 2007, respectively. From May 2012 to May 2013, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, University of Florida. He is currently a Professor with the School of Cyber Science and Technology, USTC. He has authored and coauthored more than 150 technical papers in various archival journals and conference proceedings. His research interests include next-generation Internet architecture design, transmission optimization, and network security. He serves on the Editorial Board for several journals, including the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT. He has also served as the (Lead) Guest Editor for many reputed journals/magazines, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, *IEEE Communications Magazine*, and IEEE NETWORK. He is an IET Fellow.

David S. L. Wei (Life Senior Member, IEEE) received the Ph.D. degree in computer and information science from the University of Pennsylvania in 1991. From May 1993 to August 1997, he was on the Faculty of Computer Science and Engineering with the University of Aizu, Japan (as an Associate Professor and then a Professor). He is currently a Professor with the Computer and Information Science Department, Fordham University. He has authored and coauthored more than 140 technical papers in various archival journals and conference proceedings. He currently focuses his research efforts on cloud and edge computing, cybersecurity, and quantum computing and communications. Due to his research achievements in information security, he is the recipient of IEEE Region 1 Technological Innovation Award (Academic), 2020, for contributions to information security in wireless and satellite communications and cyber–physical systems. He was the Lead Guest Editor or the Guest Editor for several special issues in the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, and the IEEE TRANSACTIONS ON BIG DATA. He also served as an Associate Editor for IEEE TRANSACTIONS ON CLOUD COMPUTING from 2014 to 2018, an Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the Series on Network Softwarization & Enablers from 2018 to 2020, and an Associate Editor for *Journal of Circuits, Systems and Computers* from 2013 to 2018. He is a member of ACM and AAAS, and a Life Senior Member of IEEE Computer Society and IEEE Communications Society.



Ruidong Li (Senior Member, IEEE) received the Bachelor in Engineering degree from Zhejiang University, China, in 2001, and the Doctoral degree in engineering from the University of Tsukuba in 2008. He is an Associate Professor with the College of Science and Engineering, Kanazawa University, Japan. Before joining Kanazawa University, he was a Senior Researcher with the Network System Research Institute, National Institute of Information and Communications Technology. His current research interests include future networks, big data networking, blockchain, information-centric network, Internet of Things, network security, wireless networks, and quantum Internet. He is the Founder and the Chair for the IEEE SIG on Big Data Intelligent Networking and IEEE SIG on Intelligent Internet Edge, and the Secretary of IEEE Internet Technical Committee. He also serves as the Chair for conferences and workshops, such as IWQoS 2021, MSN 2020, BRAINS 2020, ICC 2021 NMIC Symposium, ICCN 2019/2020, NMIC 2019/2020, and organized the special issues for the leading magazines and journals, such as *IEEE Communications Magazine*, IEEE NETWORK, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING. He is a member of IEICE.

Nenghai Yu 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 Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004, where he is currently a Professor with the School of Cyber Security and the School of Information Science and Technology. He is the Executive Dean of the School of Cyber Science and Technology 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, and information hiding.

Qibin Sun (Fellow, IEEE) received the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China in 1997, where he is currently a Professor with the School of Cyber Security. He has published more than 120 papers in international journals and conferences. His research interests include multimedia security and network intelligence and security.

Jun Lu received the bachelor's degree from Southeast University in 1985, and the master's degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China in 1988, where he is currently a Professor. His research interests include theoretical research and system development in the field of integrated electronic information systems. He is an Academician of the Chinese Academy of Engineering.