Entanglement Routing in Quantum Networks: A Comprehensive Survey

AMAR ABANE[1], MICHAEL CUBEDDU[2], VAN SY MAI[1], AND ABDELLA BATTOU[1]

[1]National Institute of Standards and Technology, Gaithersburg, MD 20899 USA
[2]Aliro Technologies, Inc., Brighton, MA 02135 USA

Corresponding author: Amar Abane (email: amar.abane@nist.gov).

ABSTRACT Entanglement routing in near-term quantum networks consists of choosing the optimal sequence of short-range entanglements to combine through swapping operations to establish end-to-end entanglement between two distant nodes. Similar to traditional routing technologies, a quantum routing protocol uses network information to choose the best paths to satisfy a set of end-to-end entanglement requests. However, in addition to network state information, a quantum routing protocol must also take into account the requested entanglement fidelity, the probabilistic nature of swapping operations, and the short lifetime of entangled states.

In this work, we formulate a practical entanglement routing problem and analyze and categorize the main approaches to address it, drawing comparisons to, and inspiration from, classical network routing strategies where applicable. We classify and discuss the studied quantum routing schemes into reactive, proactive, and hybrid routing.

INDEX TERMS Entanglement Routing, Optical Quantum, Quantum Internet, Quantum Networks

I. INTRODUCTION

IN the emerging field of quantum networking, establishing efficient routing mechanisms emerges as a fundamental challenge to enable reliable quantum information transfer across distant quantum devices [1]–[3]. Quantum communication is made possible through entanglement [4], a phenomenon where distant particles (e.g., photons) exhibit strongly correlated behaviors in a way that defies classical physics. Two independent pairs of entangled particles can be combined to produce a single long-distance entangled pair in which the local particle of the first entanglement is entangled with the remote particle of the second one, via a process known as entanglement swapping [5], [6]. Swapping is implemented by quantum repeaters that are responsible for extending the range of quantum communication to longer distances. Entanglement routing, analogous to traffic management in classical networking, includes mechanisms for coordinating quantum data flow and operations across a quantum repeater network.

A. BACKGROUND

In recent years, several protocol stack abstractions have been proposed for quantum networks [7], all identifying entanglement as the main resource for quantum communication. Thus, quantum routing must deal with certain properties unique to quantum physics – properties with no counterpart in classical routing. For example, quantum signals are fragile and cannot be amplified, copied, or indefinitely stored as with classical communication signals. This distinction has major implications for the architecture of quantum networks, as these properties prohibit near-term quantum networks from being direct-transmission networks, or networks where quantum information seamlessly flows from source to destination.

Given the maturity of current quantum hardware technologies, the most viable way to use entanglement for transmitting quantum data between two distant nodes is to establish end-to-end (E2E) entangled pairs between the two end nodes. Consequently, routing in near-term quantum networks consists of choosing the best sequence of adjacent entanglements to stitch together using swapping to establish E2E entanglements between two end nodes. Similarly to classical networks, routing in quantum networks is supported by a routing protocol to collect network information, and a path computation algorithm to select the best paths to satisfy E2E entanglement requests.

The generated E2E entangled pairs can be consumed in a variety of ways in support of a multitude of applications. In

addition to quantum data transfer via teleportation, other well-known quantum networking applications include quantum cryptography, quantum computing, and quantum sensing. Each of these application instances may impose different requirements and, as such, may place different constraints on the E2E entanglement requests for the application to be successful. For example, many applications typically consume a stream of individual E2E entangled pairs and may specify a desired E2E entanglement generation rate to achieve performance. Other applications may place requirements on the quality of the E2E entanglement, known as fidelity. More concretely, applications like quantum key distribution (QKD) may desire a high-rate stream of E2E entangled pairs, where all the produced pairs fall above a minimum fidelity threshold, whereas a distributed quantum computing application may prioritize a high-fidelity stream of E2E entanglement for more reliable teleportation of quantum data amongst processors. Such requirements may be specified as quality-of-service (QoS) parameters in the E2E entanglement requests — it is the responsibility of the routing protocol to determine ways to optimally provision paths and resources to service these requests.

Due to the inherent quantum physical properties, routing in quantum networks is more challenging than in classical routing. For example, the path computation for an E2E entanglement request does not depend only on the number of hops, the link costs, or throughput. It must also consider the requested fidelity of the E2E entanglement, where higher fidelity entanglements may take more time to be generated. Path computation must also take into account the probabilistic nature of entanglement generation and swapping, as well as the short lifetime of entangled particles. If a swapping operation fails, the two entanglements involved are destroyed and must be regenerated, while the other entanglements along the path decay. Furthermore, establishing entanglements and performing swapping relies on Local Operations and Classical Communication (LOCC) for carrying and processing measurements and control messages, which in turn adds latency, complicates signaling, and may impact the overall success probability of the routing process.

The complex challenge of entanglement routing in quantum networks has gained substantial interest recently. Researchers have considered network modeling, path computation algorithms, protocol design, and theoretical and experimental studies to explore various aspects of entanglement routing through diverse approaches.

In this paper, we review research efforts on entanglement routing in near-term quantum networks, including algorithms, protocol designs, and studies. The survey aims to provide a comprehensive description of the quantum routing problem, categorize and discuss the main approaches proposed to address it, and draw comparisons to analogous classical networking technologies. Our objective is to bridge the gap between theoretical advancements and practical implementations, shedding light on the most promising and realistic solutions that can be adapted for emerging intermediate-scale quantum networks.

Intermediate-scale quantum networks are currently being deployed at a metropolitan scale around the world [8]. They may consist of a few tens of nodes arranged in a nontrivial topology (relative to the trivial topology of a linear quantum repeater chain) with point-to-point distances of a few 100s of kilometers. These networks are intended to serve multiple users and applications simultaneously. The repeaters expected to be deployed in these networks are termed first-generation quantum repeaters (see Appendix A), with future versions including quantum error correction (QEC) capabilities. This aligns with recent research efforts, where most proposed entanglement routing approaches assume first-generation quantum repeaters or first-generation quantum routers[1].

With quantum technologies rapidly evolving, the absence of a reference architecture for intermediate-scale quantum networks means that practical deployments can vary significantly, depending on the network's purpose and the target applications. Given this variability, a one-size-fits-all quantum network architecture comparable to classical networks has yet to emerge. In light of this context, we adopt a modular approach to categorize and discuss the diverse entanglement routing strategies documented in the literature to accommodate the varying architectural needs and future developments in quantum networks. Therefore, many concepts covered in this survey, such as route computation and fidelity support, can also be independently applied to repeaterless quantum local-area networks (QLAN) [9]–[11]. In such QLANs, optical switches and wavelength division multiplexing (WDM) may be deployed to support reconfigurable topologies and optical paths, but, for example, quantum routing will not need to take into account entanglement swapping operations when servicing E2E entanglement requests. Similarly, as long-distance connections between QLANs are achieved through a linear chain of repeaters that perform swapping operations without involving path computation, aspects of these architectures (e.g., swapping, fidelity support) are also covered by this survey.

B. SCOPE

This survey reviews the literature that presents or studies routing schemes to compute paths and establish E2E entanglements. Before 2017, only a few studies had been published on routing entanglements and considered adaptations of the Dijkstra algorithm for entanglement path selection [12], [13]. Several innovative approaches have since been proposed, moving beyond Dijkstra's algorithm.

While quantum channels can be implemented over free-space optical links using satellites and optical ground stations, most of the works reviewed consider quantum channels implemented on optical fiber. Hence, this survey focuses on the entanglement distribution in optical fiber networks.

[1]A quantum router is a quantum repeater that includes routing decisions capabilities with more than two quantum network interfaces.

Moreover, we consider only routing for bipartite entanglement. This choice is motivated by several reasons. First, two of the three protocol stack models proposed for quantum networks are designed for bipartite entanglements [14], [15]. Second, the problem is well-defined mathematically and is the most studied to date. Third, because of their technological feasibility, bipartite entanglement networks are more likely to be implemented sooner than multipartite entanglement networks.

C. KEY CONTRIBUTIONS

We present several key contributions that collectively enhance the understanding and future development of routing in quantum networks. These include:

1) A discussion of quantum communications concepts through a lens of realistic quantum network design, with an alignment of these concepts to relevant classical terminology, providing a concrete reference for those familiar with traditional networking concepts.
2) The definition of a taxonomy for entanglement routing concepts, based on a distinction between routing and forwarding phases and their respective functions to provide a modular approach to quantum routing and network design.
3) A formal definition of the entanglement routing problem that comprehensively covers the major aspects necessary for effective routing.
4) A detailed discussion that encompasses entanglement routing schemes, swapping strategies, and path computation algorithms and metrics to offer a holistic perspective on the current state of entanglement routing.
5) An exploration of practical aspects in protocol design and network operation, with an identification of the main challenges and open questions for future research and development.

D. RELATED WORK

Recent literature has contributed significantly to understanding and addressing the unique challenges of entanglement routing in quantum networks. The research in [16] considers the transition from point-to-point quantum communications to wide-area quantum networks, or the quantum Internet. The authors identify the main challenges in this transition including the handling of longer distances, entanglement routing, and multi-commodity support. Their proposed framework categorizes the tasks of a quantum network into four distinct phases, each defined by its timescale (ranging from months to nanoseconds). These phases encompass network design, management, path selection, and swapping. The study focuses primarily on two types of routing: on-demand entanglement generation and proactive advance entanglement generation, providing a simple categorization of routing strategies. Our review goes beyond this by providing a more complete and detailed organization of entanglement routing functions. Moreover, various discussions are provided to bridge quantum networking concepts with general networking terminology, offering a concrete and realistic perspective. In [17], the authors explore the design challenges and opportunities in routing for quantum networks. They classify existing routing techniques into two main categories: simple routing and routing with link purification. Simple routing encompasses strategies like redundant routing (alternative paths with redundant links), concurrent routing (multiple paths provisioned simultaneously), multi-user routing (disjoint paths using intermediate nodes), and opportunistic routing. On the other hand, routing with fidelity involves limiting hops or utilizing purification techniques [18] to maintain entanglement quality. Our survey offers a more extensive review of the literature on entanglement routing, presenting a broader classification and a more comprehensive examination of the various challenges and solutions involved in quantum routing.

E. STRUCTURE

The structure of this paper is designed as follows. In the remainder of this section, we briefly introduce the key concepts in quantum communication along with the main characteristics of a first-generation quantum repeater network. For readers unfamiliar with these fundamental operations of quantum networking, Appendix A offers a more detailed description.

Section II presents a detailed mathematical formulation of a quantum routing model that highlights key challenges and differences with classical routing. In Section III, we explore the approaches for addressing the entanglement routing problem and propose a taxonomy to classify and examine them. The section also covers path computation algorithms, along with strategies for entanglement swapping, fidelity and purification support, and path reliability. We end Section III with a discussion of the various approaches introduced, focusing on their interplay and their practical limitations. In Section IV, we discuss practical approaches for implementing entanglement routing protocols by drawing inspiration and analogies from classical networking architectures and protocols.

Section V highlights the main challenges to address in the current landscape of entanglement routing and introduces key open questions. Finally, Section VI synthesizes the findings, discusses the implications of our review, and outlines potential directions for future research.

F. QUANTUM PRELIMINARIES

Figure 1 summarizes the key components, processes, and characteristics of quantum communication through an illustration of an E2E entanglement distribution between two quantum nodes (Alice and Bob) connected via a quantum repeater (router). Alice and Bob use quantum and classical channels to produce multiple entangled pairs with an intermediate router node. The router performs swapping to produce the E2E entangled pairs, which may be further purified by Alice and Bob to meet the fidelity requirements

of an application. The probabilistic nature of these processes must be taken into account in the entanglement routing process.

A brief glossary of the main quantum communication concepts is presented below, while Appendix A provides a detailed background of the primary hardware components and protocols that make up an operational quantum network.

a: Qubits and Quantum States

Quantum information is represented by quantum bits (qubits) and is usually encoded in the quantum state of particles such as photons, electrons, and atoms.

b: Elementary Entanglement

Bipartite entanglement is a special connection between two quantum states, where their properties are linked together regardless of the distance between them. These pairs of entangled qubits are known as *Bell pairs* or Einstein-Podolsky-Rosen (EPR) pairs (terms that we will use interchangeably throughout this paper). Elementary entanglement is defined as bipartite entanglement shared between two neighboring nodes (i.e., directly connected through an optical channel).

c: Entanglement Generation and Heralding

As the generation of entangled pairs is nondeterministic, near-term quantum communications will rely on a process called heralding in which two nodes mutually acknowledge the confirmed presence or absence of an entangled pair between them when attempting an entanglement creation. As such, a Heralded Entanglement Generation (HEG) protocol requires classical signaling messages between the two nodes, which incur additional communication overhead and thus have critical implications for quantum routing.

d: Quantum Memory

Qubits may be stored in quantum memory using a variety of technologies. Memories are characterized by key parameters, such as their storage time, which represents the time interval beyond which the stored quantum state is irreversibly degraded and can no longer be used. This results from entanglement decoherence, where the entangled pair of particles degrades over time because of interactions with their surrounding environment. The quantum state may incur a variety of errors as it is received through the quantum channel, stored in memory, operated on by quantum gates, emitted from the memory, and coupled into an output quantum channel. Some of these errors may be corrected by error correction codes, whereas other errors are uncorrectable and result in a lower fidelity entangled state [19]. Some quantum memory platforms may be able to store a plurality of qubits simultaneously, in an individually-addressable manner, resulting in more robust multiplexed entanglement generation and storage capabilities, albeit with their own set of errors to consider due to crosstalk and scattering [20]–[23]. Depending on the underlying technology, the memory may operate in a wavelength range different from that of the input or output quantum channel. In such cases, a quantum frequency conversion step may be needed to integrate the memory platform into the quantum network channel and preserve the quantum state [24]–[26]. Other memory parameters may include storage efficiency, retrieval efficiency, and the supported wavelength range.

e: Entanglement Swapping

Entanglement swapping is to perform a Bell state measurement on two independent pairs of entangled qubits. Given the inherently lossy nature of quantum channels, entanglement swapping via repeaters becomes key to distributing long-distance entanglements. Specifically, a repeater is placed between Alice and Bob to split their distance into two smaller distances. Two elementary entanglements are generated; one between Alice and the repeater and one between the repeater and Bob. The repeater then interferes and measures its two local (and ideally indistinguishable) qubits, thereby destroying the two elementary entanglements in the process, and sends the measurement outcome to Bob. If successful, this procedure establishes entanglement between the two remote qubits on Alice and Bob. The success of the swapping operation is also probabilistic.

f: Fidelity and Purification

The probability that a pair of entangled qubits is in a specific, desired state is quantified by a value known as the fidelity of an entanglement. Decoherence, fiber loss, or environmental disturbances can prevent a pair of qubits from achieving or maintaining a maximally entangled state. Purification techniques such as Heralded Entanglement Purification (HEP) have been developed to improve fidelity [18] by converting two or more low-fidelity Bell pairs into a single pair with higher fidelity. This process can sometimes fail and requires classical communication to notify both end nodes about the outcome of the purification effort [27].

g: Quantum Repeaters

The main function of a repeater is to capture qubits, store them in quantum memory, and implement the purification and swapping processes described above. Entanglement distribution over quantum repeaters is subject to two main types of errors [28]: loss errors arising from the attenuation of fibers and operational errors resulting from inaccuracies in manipulating and measuring quantum states. The evolution of quantum repeaters can be delineated into three distinct generations based on their error mitigation strategies [28]; see Appendix A for further details.

h: Teleportation

Quantum teleportation is a process in which a qubit is recreated at a distant destination node by utilizing an entangled link, while simultaneously destroying the original qubit at the source node. As long as the source and destination nodes share a Bell pair, teleportation can occur over any distance without physically transferring the quantum particle that

encodes the qubit. Teleportation serves as an exemplary application of utilizing E2E entanglements, highlighting just one of the potential ways to leverage these entangled states in quantum communication and computation systems.

II. A QUANTUM ROUTING PROBLEM

In this section, we describe a general entanglement routing model in a centralized, offline, and synchronous setting adapted from previous work in [29]–[32]. Time is loosely synchronized and slotted, and within each time slot, there are two main phases, namely the *external phase* for generating elementary entanglements among neighboring nodes and the *internal phase* for swapping pairs of entangled qubits inside each node in order to establish longer-distance entanglements. Here, we assume that the duration of a time slot is chosen appropriately depending on hardware components so that established entanglements do not decohere before being used within one time slot.[2] Additionally, while purification is important for improving entanglement fidelity (and thus critical for the robustness and scalability of future quantum networks), we will not include it in the following formulation. This is mainly to simplify the presentation of the main concepts and key elements of the entanglement routing problem. Specifically, we will focus on entanglement generation, swapping and routing, and provide a detailed discussion on purification in subsection III.B.2 below.

a: Topology

Consider a quantum network described by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Here, each node $u \in \mathcal{V}$ is a quantum node, equipped with a limited number of qubits to create Bell pairs. All nodes are connected via a classical network. An edge $(u, v) \in \mathcal{E}$ existing between two nodes u and v means that they share one or more quantum channels allowing for qubit transmission. We refer to the graph $\mathcal{G}$ formed by the nodes and physical channels as the physical topology of the quantum network.

b: Quantum Link

Since quantum channels are inherently lossy, each attempt to create an entanglement through a channel only succeeds with a certain probability. For an edge (u, v), this probability is proportional to $e^{-\alpha L_{uv}}$, where L_{uv} is the physical length of the channel and α is a constant depending on the physical media (e.g., optical fiber, free space). If an attempt succeeds, the two nodes u and v share an entangled pair, i.e., there is a *quantum link* between u and v. Let us denote by p_{uv} the overall success probability of a quantum link, taking into account the efficiencies of entanglement sources and photon detectors, and the number of attempts allowed in one phase within a time slot. Here, for simplicity, we can assume that the physical media are the same for all channels and the success probability p_{uv} is the same for all the channels connecting u and v.

Each edge $(u, v) \in \mathcal{E}$ is then characterized by a capacity C_{uv} representing the maximum number of quantum links that can be established before swapping within a time slot. For simplicity, one can consider C_{uv} as the number of parallel channels between u and v. In general, however, the capacity C_{uv} can be different from the number of channels, taking into account different multiplexing modes (including time, space, and wavelength multiplexing) and possibly the limited numbers of qubits at u and v.[3] We refer to the (multi-)graph of nodes and edges associated with quantum links as the *logical topology* or *virtual topology*, which could be time-varying. It is straightforward to consider the number of successful quantum links between u and v in each time slot as a random variable following a Binomial distribution with parameters C_{uv} and p_{uv}. In this model, the probability of having exactly k entanglements on the edge (u, v) in each time slot is given as follows for $k = 0, 1, \ldots, C_{uv}$

$$p_k(u, v) = \binom{C_{uv}}{k} p_{uv}^k (1 - p_{uv})^{C_{uv}-k}. \tag{1}$$

c: Swapping in a Path

A (quantum) path between two nodes is simply a concatenation of contiguous edges with positive capacities, where an E2E entanglement can be established by creating quantum links on all the edges and performing quantum swapping at the intermediate nodes. In particular, to combine two adjacent entanglements, say (u, v) and (v, w), node v attempts a swapping on its corresponding local qubits which may succeed with a probability q_v resulting in an (u, w) entanglement.[4] In a more general case, suppose that node v has l (u, v)-entanglements and r (v, w)-entanglements. Then, assuming that swapping operations can be performed between any pair of qubits in memory of a node,[5] node v can attempt at most $\min\{l, r\}$ swaps to establish multiple entanglements between u and w. As a result, by swapping at all the intermediate nodes, elementary entanglements are consumed to generate end-to-end entanglements.

In the following, we will refer to a set of rules for performing swapping in a path as a *swapping policy*. There are different swapping policies resulting in different swapping

[2]The duration of a time slot is actually an important factor that affects the quantity/quality of entanglements since it is related to the entanglement generation rates as well as coherence times of quantum memories. It is also crucial for practical implementations of routing algorithms because near-term quantum communication relies on heralded entanglement generation that requires classical signaling between quantum nodes. For simplicity, we assume in the this section that the duration of a time slot is sufficient for carrying out necessary operations of a routing algorithm before entanglements decohere, including processing time of routing protocols, entanglement generation/swapping and signaling, and their consumption by applications.

[3]The special case with $C_{uv} = 1$ has been studied extensively in the literature.

[4]Using linear optics platforms, the swap success probability is bounded by $q_v \leq 0.5$. However, other platforms have achieved higher success probabilities [33].

[5]i.e., a BSM can be performed between any pair of locally held qubits so that a quantum node can also act as a quantum switch.

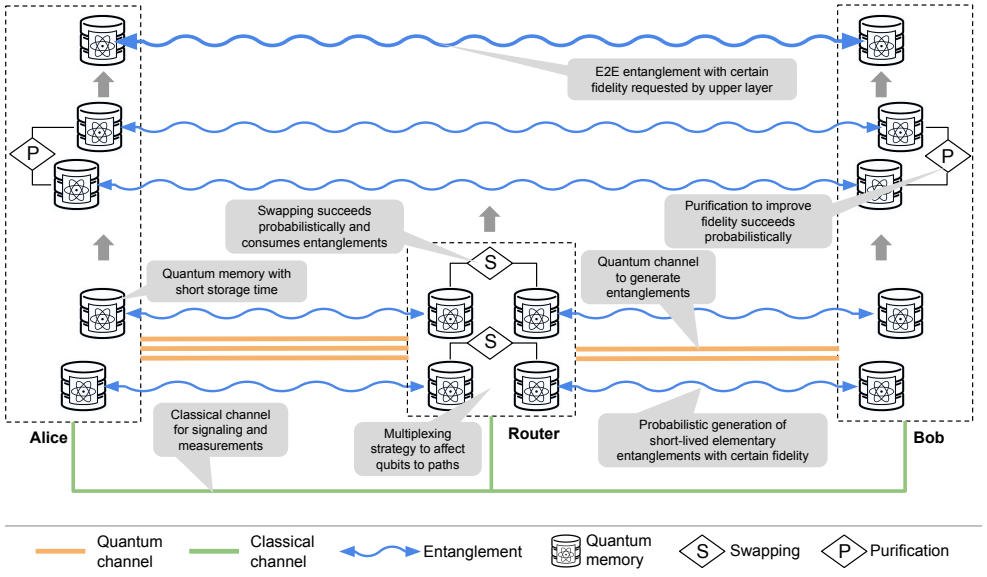

FIGURE 1. Key quantum communication protocols for servicing E2E entanglement requests.

schemes as we will detail in subsection III-B1 below. In this section, we present two groups of policies to accommodate the routing model described here, namely heralded and unheralded swapping.

- **Unheralded swapping**: In this scenario, routers locally perform swapping without awaiting the swapping outcomes of other routers. This allows all the nodes to carry out swapping operations independently and in parallel once all adjacent quantum links are available. Consider a general path of length n given without loss of generality by $\{(0,1),(1,2),(2,3),\dots(n-1,n)\}$ with corresponding capacities $C_{i-1,i}$ for $i=1,\dots,n$. For any $i<j\leq n$, we can define the capacity of the subpath from i to j as

$$C_{ij}=\min_{i<l\leq j} C_{l-1,l}. \tag{2}$$

Let $\tilde{p}_k(0,i)$ denote the probability that among the first i hops, at least one hop has exactly k quantum links while other hops have at least k quantum links, where $1\leq k\leq C_{0i}$. This probability can be computed recursively as follows for $i=2,\dots,n$

$$\begin{aligned}\tilde{p}_k(0,i)=\tilde{p}_k(0,i-1)\sum_{l=k}^{C_{i-1,i}} p_l(i-1,i)\\ +p_k(i-1,i)\sum_{l=k+1}^{C_{0,i-1}}\tilde{p}_l(0,i-1)\end{aligned} \tag{3}$$

with $\tilde{p}_k(0,1)=p_k(0,1)$ and $p_l(i-1,i)$ is computed as in (1).[6] Since edges can have different numbers of successful quantum links in each time slot, all the nodes will need to agree on how to perform internal swapping effectively instead of randomly pairing internal qubits. An easy way to do this is to assign IDs to quantum links and pair them in a common order (e.g., ascending, descending). Under such qubit binding and swapping, the probability of having exactly k end-to-end entanglements, denoted by $p_k(0,n)$, is computed as follows for $k=1,\dots,C_{0n}$

$$p_k(0,n)=\sum_{l=k}^{C_{0n}}\tilde{p}_l(0,n)\binom{l}{k}\bar{q}^k(1-\bar{q})^{l-k} \tag{4}$$

with $\bar{q}:=\prod_{i=1}^{n}q_i$, $p_0(0,n)=1-\sum_{k=1}^{C_{0n}}p_k(0,n)$ and $p_k(0,n)=0$ for all $k>C_{0n}$. Here, note that, because swapping operations are independent, the probability of successfully connecting all elementary entanglements is simply the product of individual swapping probabilities, which does not depend on any swapping order.

- **Heralded swapping**: In this scenario, routers carry out swapping based on the swapping outcomes of other routers, giving rise to a swapping order for the path in each time slot. Compared to unheralded swapping, heralded swapping aims to make better swapping choices at each node at the expense of increasing the amount of heralding signals. Again, let us consider a path of length n given by $\{(0,1),(1,2),(2,3),\dots(n-1,n)\}$ where the capacity C_{ij} of the subpath from i to j is defined as in (2). With abuse of notation, let $p_i(x,y)$ and $p_j(y,z)$ denote the probabilities of having exactly i (x,y)-entanglements and j (y,z)-entanglements at node y, where $i\leq C_{xy}$, $j\leq C_{yz}$, and x, y and z need not be adjacent nodes. Thus, the probability of having exactly k (x,z)-entanglements after at most

[6]Note that [30] used a similar expression to (3) but with C_{0n} as the upper limit for both summations on the right-hand side, thereby leading to a lower success probability.

C_{xz} swapping attempts at node y can be computed as follows [32] for $k = 1, \ldots, C_{xz}$

$$p_k(x,z) = \sum_{i=k}^{C_{xy}} \sum_{j=k}^{C_{yz}} p_i(x,y) p_j(y,z) \times \binom{\min\{i,j\}}{k} q_y^k (1-q_y)^{\min\{i,j\}-k} \tag{5}$$

with $p_0(x,z) = 1 - \sum_{k=1}^{C_{xz}} p_k(x,z)$. This allows us to find probabilities of long-distance entanglements $p_k(0,n)$ for any swapping order. The probability distribution of end-to-end entanglements in this case is more complicated to compute than the unheralded case.

d: Path Throughput

As shown above, given a path π in the graph $\mathcal{G}$ with a certain edge capacity $\mathcal{C}_\pi := \{C_{uv}^\pi : (u,v) \in \pi\}$, end-to-end entanglements can be established by swapping at all intermediate nodes, which results in different success probabilities depending on how swapping operations are carried out. Let $\mathcal{Q}$ be a swapping policy and $p_k^{\mathcal{Q}}(\pi)$ denote the probability of having exactly k end-to-end entanglements on path π under policy $\mathcal{Q}$. Then we can define the expected throughput as follows

$$\mathrm{EXT}(\pi; \mathcal{Q}) := \sum_{k=1}^{C_\pi} k \times p_k^{\mathcal{Q}}(\pi), \quad \text{with } C_\pi = \min \mathcal{C}_\pi, \tag{6}$$

where C_π is also known as the (minimum) width of the path. Note that the throughput can be computed in $O(|\pi| \max \mathcal{C}_\pi)$ time for unheralded swapping and in $O\big(|\pi|(\max \mathcal{C}_\pi)^2\big)$ time for the case of heralded swapping. In both cases, the time complexity scales linearly with the path length, but unlike the former, the latter is quadratic in the maximum width of the path.

e: Demands/Requests

A demand can be defined as a tuple $r = (s, d, \delta, \underline{F})$, representing a request to deliver δ E2E entanglements with a minimum fidelity $\underline{F}$ per time unit between two end-nodes s and d [31]. A request may also include other requirements, such as a desired latency $\bar{l}$, but we do not consider them here for simplicity.[7]

Let $\mathcal{P}_r$ denote the set of feasible paths serving request r and $\mathcal{Q}_{\mathcal{P}_r}$ the set of corresponding swapping policies, i.e., $\mathcal{Q}_{\mathcal{P}_r} = \{\mathcal{Q}_\pi : \pi \in \mathcal{P}_r\}$, where a swapping policy $\mathcal{Q}_\pi$ can depend on the path itself and thus can also be a design parameter. The expected end-to-end entanglements delivered for this request in a time slot can then be defined as

$$R(\mathcal{P}_r, \mathcal{Q}_{\mathcal{P}_r}) := \sum_{\pi \in \mathcal{P}_r} \mathrm{EXT}(\pi; \mathcal{Q}_\pi) \tag{7}$$

which can be considered as the raw throughput. The expected number of entangled qubit pairs delivered per time unit for a request that satisfies its fidelity (and possibly latency) constraints is referred to as the end-to-end (expected) throughput of the request, which is denoted by $\tilde{R}(\mathcal{P}_r, \mathcal{Q}_{\mathcal{P}_r})$ and upper bounded by $R(\mathcal{P}_r, \mathcal{Q}_{\mathcal{P}_r})$.

f: Objective

Consider a quantum network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that is required to serve a set of requests $\mathcal{R} = \{r_i\}_{i=1}^m$.[8] A common objective of routing algorithms is to maximize the total number of end-to-end entanglements delivered for all requests. The goal of a routing algorithm is then to efficiently determine a set of paths $\{\mathcal{P}_r : r \in \mathcal{R}\}$ together with allocated capacity $\{\mathcal{C}_\pi : \pi \in \mathcal{P}_r\}$ and corresponding swapping policies $\{\mathcal{Q}_\pi : \pi \in \mathcal{P}_r\}$ to either satisfy all requests or to maximize the overall throughput in the network [30], [31], [34], [35]. To this end, we can then consider a utility function $U_r\big(\delta_r, R(\mathcal{P}_r, \mathcal{Q}_{\mathcal{P}_r})\big)$ for each request $r \in \mathcal{R}$ with a desired flow rate of δ_r. The problem can be formulated as maximizing the following aggregated utility function[9]

$$\sum_{r \in \mathcal{R}} U_r\big(\delta_r, R(\mathcal{P}_r, \mathcal{Q}_{\mathcal{P}_r})\big) \tag{8}$$

with decision variables: $\{\mathcal{P}_r, \mathcal{C}_\pi, \mathcal{Q}_\pi, \forall \pi \in \mathcal{P}_r, \forall r \in \mathcal{R}\}$ and subject to the following:

- Feasible paths: each $\pi \in \mathcal{P}_r$ has no loops.
- Capacity: $C_{uv}^\pi \geq 0, \forall (u,v) \in \mathcal{E}$ and

$$\sum_{\pi \in \cup_{r \in \mathcal{R}} \mathcal{P}_r} C_{uv}^\pi \leq C_{uv}, \quad \forall (u,v) \in \mathcal{E}. \tag{9}$$

- Fidelity: Since the fidelity of an entanglement drops with each swapping, a minimum fidelity requirement $\underline{F}_r$ can be replaced with a path length constraint for each path $\pi \in \mathcal{P}_r$ without considering entanglement distillations [31]. In particular, assuming all the mixed entangled states are Werner states, the following hop constraint can be used instead

$$|\pi| \leq h_r := \frac{\log(\frac{4\underline{F}_r - 1}{3})}{\log(\frac{4F_0 - 1}{3})}, \quad \forall \pi \in \mathcal{P}_r, \forall r \in \mathcal{R}, \tag{10}$$

  where F_0 is the minimum fidelity of each elementary Bell pair.

Figure 2 illustrates the routing concepts in a quantum network, where multiple quantum links can be created over an edge (multi-channel) and the routing algorithm can accept one or multiple requests at a time (multi-request). Here, entanglement generation between S1 (resp. S2) and T1 (resp. T2) is requested and one or multiple paths are provisioned for each request (multi-path). We emphasize that while not

[7] In fact, due to the stochastic nature of entanglement generation and swapping operations, the generation latency of a request can be considered as the inverse of its expected E2E entanglement rate. As a result, a requirement on the generation latency can simply be transformed into a constraint on its minimum expected throughput.

[8] For simplicity, we assume here that $\mathcal{G}$ and $\mathcal{R}$ are fixed within a period of interest. In general, one can consider time-varying graph $\mathcal{G}$ due to failures as well as time-varying requests with possibly estimated or unknown arrival patterns.

[9] To provide certain levels of fairness, the overall objective function can be defined as a weighted sum of all utilities or throughputs.

exhaustive, the problem formulation presented here captures key challenges in quantum routing. Previous research has often focused on simplified or approximate versions of this problem as we will explain below and in the next sections.

g: Routing Decisions and Complexity

Given inputs to the routing problem, including network graph $G = (\mathcal{V}, \mathcal{E})$, edge capacity C_{uv} with corresponding quantum link success probability p_{uv} for all $(u, v) \in \mathcal{E}$, swapping probability q_v for all $v \in \mathcal{V}$ and the set of requests $\mathcal{R} = \{r_i = (s_i, d_i, \delta_i, \underline{F}_i)\}_{i=1}^{m}$, the outputs of a routing algorithm after solving the above problem are the routing decisions, namely,

- A set of paths $\mathcal{P}_r$ for serving each request $r \in \mathcal{R}$, together with allocated capacity C_{uv}^{π} for each edge (u, v) along any path $\pi \in \mathcal{R}_r$. These decision variables will be used for resource allocation and signaling protocols in each node. We will discuss this step further in subsection III-A below. Finding these decision variables corresponds to similar tasks in classical networking, namely path computation and path installation.
- A swapping policy $\mathcal{Q}_\pi$ for any path $\pi \in \mathcal{P}_r$. This will determine the local signaling needed for swapping at each node in support of multiple paths and multiple requests, which somewhat resembles the forwarding table in classical networking. Given that performing an entanglement swap on two imperfect Bell states results in a long-distance entanglement with reduced fidelity, if such reduction is significant, one might need to take into account entanglement purification or distillation steps to probabilistically convert multiple low-quality entangled pairs into a single high-quality entangled pair. In this case, one needs to design not just a swapping policy but a *forwarding policy* that includes swapping and purification steps, where purification can be done on elementary or distant entanglements to improve fidelity. Further discussions on swapping and purification will be given in subsection III-B below.

Here, let us briefly remark on the challenges in solving the routing problem described above that do not have a counterpart in classical networking. In general, it is difficult to solve the above optimization problem exactly and efficiently even in the offline and centralized setting because of the following challenges:

- The combinatoric nature of the solution space for the decision variables, namely paths in graph $\mathcal{G}$ with integer capacity C_{uv}^{π} and swapping policy/order, where the latter has no counterpart in classical routing and forwarding. Specifically, in the case of heralded swapping, the number of possible orders for a path of length n scales as $O(4^n)$.[10] In addition, we note that different orders will also have different levels of parallelism and signaling mechanisms, which in turn can also affect the practical throughput. For simplicity, we do not consider such effects in the formulation above (which is reasonable when the duration of the internal phase is negligible compared to that of the external phase for generating all elementary entanglements).
- Lack of efficient path metrics: As shown in (1)–(6), the expected throughput of a path is rather complicated, involving the characteristics of each node and every edge on the path as well as the swapping policy employed for the path. As a result, comparing two paths becomes nontrivial, especially when they have different edge capacities, different lengths, and different swapping policies, unless computation for the whole path is finished. However, the computational complexity of the throughput in heralded swapping scales quadratically in the maximum width of the path. More importantly, under any swapping policy, finding a path with maximum expected throughput in a network does not have the subpath optimality property, causing methods like Dijkstra and Bellman-Ford to fail in finding optimal paths [32], [35].

Thus, solving the routing problem formulated above remains a challenging task. Most existing works often fix a swapping policy or order (e.g., [30], [35], [36]) and consider the routing problem using Dynamic Programming and/or replacing exact path throughput with heuristics that are more efficient and/or Dijsktra friendly. For example, path computation can be done on-line or off-line using a heuristic path metric such as hop count, edge width, or link fidelity [29], [30], [37], [38].

It is important to note that the problem described above assumes a centralized, offline and synchronous setting. The following extensions can be more challenging but also more practical to consider:

- Distributed: Routing paths and swapping orders are computed in a (semi-)distributed fashion based on nodes' local view of the network.
- Online: Routing can be adaptive to request arrival patterns and available elementary link entanglements, which are possibly unknown in advance. In stead of static swapping policy/order, swapping can also be dynamic to better utilize elementary entanglements.
- Asynchronous: Different levels of asynchrony can also be considered. For example, nodes can have different time slot duration or operate based on discrete events. For each node, the internal and external phases for each memory can be asynchronous as well to better improve throughput and fidelity.

h: Physical parameters

One can illustrate the relevant time constraints involved in entanglement routing through the lens of several key physical parameters. The coherence time is an important factor in establishing and maintaining high-fidelity entanglements. Currently, quantum coherence times can be up to milliseconds for photons in optical fibers and up to seconds or more

[10] This is also related to the n-th Catalan number [32] in combinatorial mathematics.

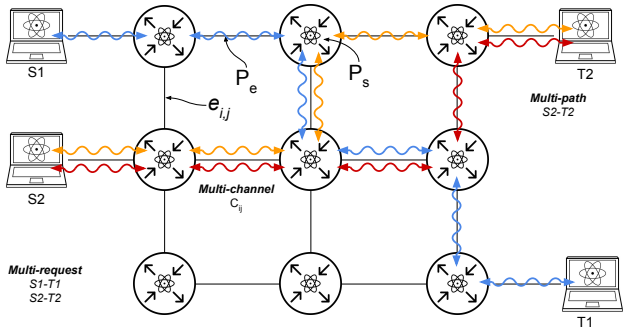

FIGURE 2. Entanglement routing on a grid network topology (adapted from [39]). Edge $(i, j) \in \mathcal{E}$ supports multiple entangled pairs, whose capacity is denoted as C_{ij}. Entanglement generation between S1 (S2) and T1 (T2) is requested, and one or more paths are provisioned for each request. The probabilities of successfully building an entangled pair between adjacent nodes and performing swapping are denoted as p_{ab} for $(a, b) \in \mathcal{E}$ and q_b for $b \in \mathcal{V}$, respectively.

for other systems [40], [41]. In particular, typical coherence times of spin qubits can be milliseconds up to seconds in silicon [42] and diamond [40], [43].[11] The success probability of a single entanglement attempt is dependent on the length of the optical fiber (besides other factors such as efficiencies of detectors and memories, as well as the success rate of BSM) and is typically measured on the order of 10^{-4} in a lab environment over a few-meter links [46] and 10^{-6} over $6-33$ km telecom fiber [47] at a frequency of several tens of kHz. A swapping success probability $q = 0.5$ is also feasible. Although BSMs based on linear optics have a maximum success probability of 0.5, a more complex measurement pattern using ancillary photons has been experimentally demonstrated recently to achieve a success probability of approximately 0.579 [33]. Depending on the attenuation coefficient of the optical fiber, repeater stations may be spaced at intervals of $L \approx 20\,\mathrm{km}$, and the one-way classical communication time is $L/(2 \times 10^5\ \mathrm{km/s}) = 0.1\ \mathrm{ms}$.

III. ENTANGLEMENT ROUTING APPROACHES

Drawing inspiration from classical networking, we decompose the E2E entanglement distribution process into two distinct phases, namely routing and forwarding [48]. The decision to examine the two phases separately stems from the realization that these phases can be implemented independently in quantum networks. For instance, in repeaterless QLANs, route computation approaches may be relevant without necessitating entanglement swapping operations for servicing entanglement requests. Conversely, long-distance connections between QLANs over linear paths with a chain of repeaters might utilize forwarding techniques such as swapping and purification without engaging in path selection.

It is important to note that the term *forwarding* in the context of quantum communication is somewhat a misnomer. Unlike classical networks, where forwarding refers to hop-by-hop packet transmission, in quantum networks, it involves generating and swapping entanglements across nodes. However, we use this classical term to provide a more familiar and abstract description of quantum networking processes. Furthermore, as we will discuss, the forwarding phase includes entanglement purification and swapping — two operations that are often viewed as distinct functions. Therefore, we retain the term *forwarding* to keep this phase conceptually separate from the two distinct processes it contains.

The routing phase is concerned with determining (optimal) paths for E2E entanglement requests and ensuring the necessary routing instructions are in place across the network to support these paths. This includes all the background processes needed to select paths such as collecting topology. The routing phase involves two key aspects: path computation and route installation.

Path computation consists of choosing the sequence of intermediate links (and nodes) that will generate the E2E entanglements. This process should not happen for each individual E2E entanglement. Instead, it should be launched once for a single or a group of E2E entanglement requests. Path computation is strongly tied to the topology, which can be physical and/or logical and/or virtual, and the resources (memory qubits) already in use by other paths. Path computation uses a routing algorithm to calculate paths for the requested E2E entanglements. These algorithms may model various network parameters and optimizations as discussed in Section III-C.

Route installation follows the path computation and involves the deployment of specific instructions at each node along the path to establish E2E entanglements. These instructions can be static and installed manually on the nodes at configuration time, or dynamically transferred from path computation results, akin to updating forwarding tables or setting up cross-connects in classical networks.

Path computation and route installation can be implemented in a distributed fashion across the nodes, in a centralized controller, or combine the two approaches (see Section IV).

The forwarding phase usually takes place after route installation. It includes the external phase for the production of elementary entanglements (which may not yet exist) and the internal phase where swapping is executed to establish E2E entanglements. The forwarding phase may include purification and mechanisms to ensure path reliability. The combination of entanglement swapping and purification has been a fundamental approach in early quantum repeater protocols, where nested entanglement purification was used alongside swapping to enhance fidelity [49], [50]. Although these techniques are crucial for enabling long-distance entanglement, they primarily concern improving link quality within a predetermined chain of repeaters. In contrast, the entanglement routing problem, as discussed in this work, includes the selection of an optimal path across a network

[11]Trapped ions can have a lifetime from minutes to hours [44], but the efficiency of frequency conversion to telecommunication wavelength is still

of repeaters to maximize entanglement generation efficiency under resource constraints. This distinction is important because routing focuses on network-wide path selection, whereas early repeater protocols optimized fidelity along pre-established routes.

Routing algorithms can be designed to operate on a partial or local view of the network topology (distributed), or, alternatively, they may require a global view of the topology (centralized). According to the routing problem formulation given in Section II, the routing algorithm may include forwarding operations, such as optimal swapping strategies, in addition to routing decisions (i.e., path computation). Hence, we discuss the routing algorithms separately without limiting them only to the routing or forwarding phase.

The specific concepts and approaches that underpin the routing phase, the forwarding phase, and the routing algorithms are organized in the taxonomy outlined in Figure 3. We chose to exclude the notion of a "routing protocol" in the taxonomy, considering it as a result of combining techniques from the categories identified (see Section IV). To use the taxonomy diagram for routing protocol design, one may start by selecting a path computation scheme for the network's technology and objectives, such as a proactive path computation. Then choose a route installation that matches the network's capabilities and implementation preferences, such as a centralized mode. Next, decide on the swapping strategy for the forwarding and the needed QoS features. Finally, pick or develop a model and algorithm to address the routing problem formulation. Note that not all combinations of these elements will be consistent or practical, so careful consideration is needed to ensure compatibility and effectiveness.

A. ROUTING PHASE

We distinguish three main routing schemes based on the network topology on which paths are computed and the time at which path computation takes place with respect to entanglement creation: proactive, reactive, and hybrid.

1) Proactive

In proactive routing (sometimes designated as *on demand* entanglement generation [16], [51]) the path computation and route installation take place before elementary entanglements are created[12].

A centralized proactive routing architecture is illustrated in Figure 4. Initially, E2E entanglement requests are received by the controller, which computes the paths and provides the nodes (routers and end-nodes) with instructions for entanglement creation and swapping for each path. Subsequently, the nodes coordinate to create E2E entanglements, incorporating purification processes if supported. End-nodes are notified of available E2E entanglements as part of the swapping process.

Distributed proactive routing is also possible and uses a connection phase to install paths and negotiate resources between nodes, as illustrated in Figure 5. Path computation can be done at the requester or the receiver, or hop-by-hop along the path.

The path computation may take into account parameters for entanglement creation, multiplexing, and swapping, and provides necessary instructions to the nodes. These instructions apply to a large stream of E2E entanglements until different instructions are installed. Once the paths are computed and the relevant instructions are installed, E2E entanglements are generated for each path until a stop condition is met (e.g., the request terminates) or new paths and instructions are installed.

Since path computation and route installation are not constrained by entanglement decoherence time, proactive routing allows more flexibility in its design. It can be envisioned in a centralized system where routing instructions are computed and installed from a central entity. Alternatively, proactive routing can be realized as a distributed system, where each node computes and installs its routing instructions. In the latter case, a routing algorithm with a partial knowledge of the physical topology can be used. Note that a majority of the proposed proactive routing approaches assume a global knowledge of the physical topology by the path computation algorithm. The routing algorithm typically considers the physical topology and can be as simple as pre-computing the paths for all the possible pairs of nodes at initialization time [30].

Since the path computation is executed before knowing which elementary entanglements have succeeded, a cost for physical links (paths) must be modeled in the algorithm. In its simplest form, link cost reflects the expected entanglement throughput (i.e., generation rate). The throughput is based on the channel loss and error rates, which is mainly dependent on the fiber length. For a more realistic throughput estimation, other characteristics may be included such as photon source power, detector efficiency, and quantum memory coherence, frequency, and efficiency [35]. Based on entanglement generation rates and swapping probabilities, the path throughput can be modeled as in Section II above, or approximated by using different heuristic metrics such as the sum of node distances or link entanglement generation rates [30]. Overall, proactive path selection tends to prefer shorter paths in physical distance, although this does not guarantee selection of paths with the highest throughput in the presence of links with varying capacities.

Without a-priori knowledge of the existing entanglement links, if the path computation algorithm does not plan for redundant links or paths, the E2E entanglement throughput may quickly degrade due to failures in creating the planned entanglements [37]. Moreover, proactive routing may induce a higher latency to satisfy the E2E entanglement requests as

[12]In this paper, we use "proactive" to describe path computation occurring before entanglement generation. However, these terms can be used inversely in other studies, such as in [7], [52], where what we call "reactive" might be referred to as "proactive" and vice versa. This inversion stems from the perspective of considering the entanglement generation process relative to path computation. Since we are discussing routing schemes, the terminology chosen here considers the path computation process relative to entanglement generation.

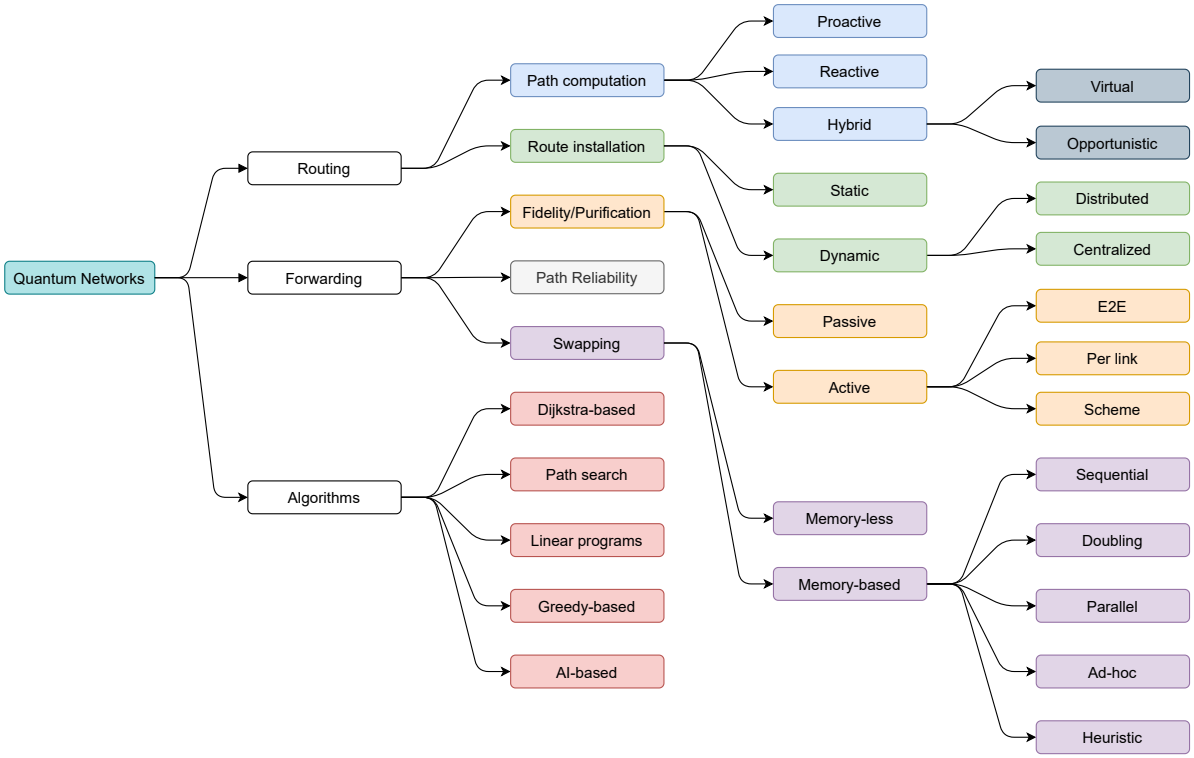

FIGURE 3. Taxonomy of quantum routing concepts discussed in this survey.

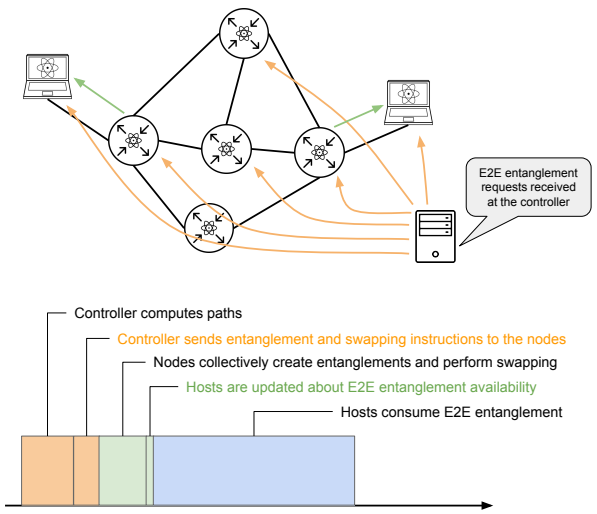

FIGURE 4. Architecture of a centralized proactive routing. The purpose of the slotted representation is only to illustrate the sequential operations involved in proactive routing.

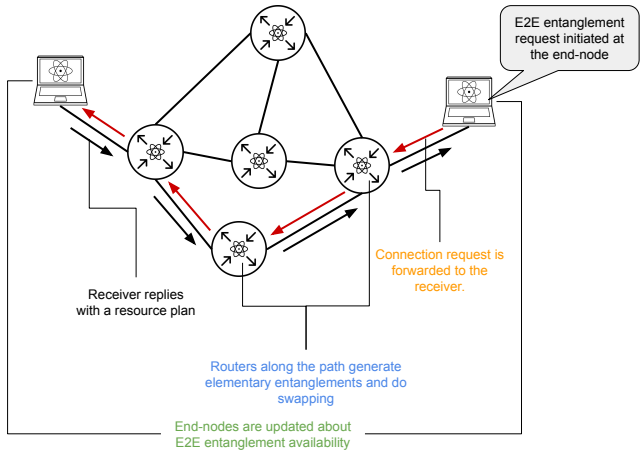

FIGURE 5. E2E entanglement distribution in distributed proactive routing. Paths can be computed by the requester, the receiver, or hop-by-hop along the path.

multiple attempts may be required before creating entanglement links. This latency may be further increased when path computation is distributed due to additional signaling. These limitations lead to considering the reactive routing scheme discussed next.

2) Reactive

In reactive routing (sometimes designated as *advance* [16] or *continuous* [51] entanglement generation), entanglements are constantly produced over each quantum channel across the network, and path computation is done on the instant logical topology formed by the created entanglements.

This scheme frequently assumes a synchronous network where the quantum and routing operations evolve sequentially

within discrete time slots of a fixed duration. Such a time-slotted system is used particularly with centralized path computation [53], although it can also be used in distributed routing [29].

A centralized slotted reactive routing is illustrated in Figure 6. At the beginning of a time slot, each pair of adjacent routers (i.e., sharing a quantum channel) attempts to generate entanglements. Entangled link-states are communicated to the routing element (e.g., controller) which computes paths for the current E2E entanglement requests based on the logical topology formed by the created entanglements. The routing element sends swapping instructions to the routers to establish the E2E entanglements.

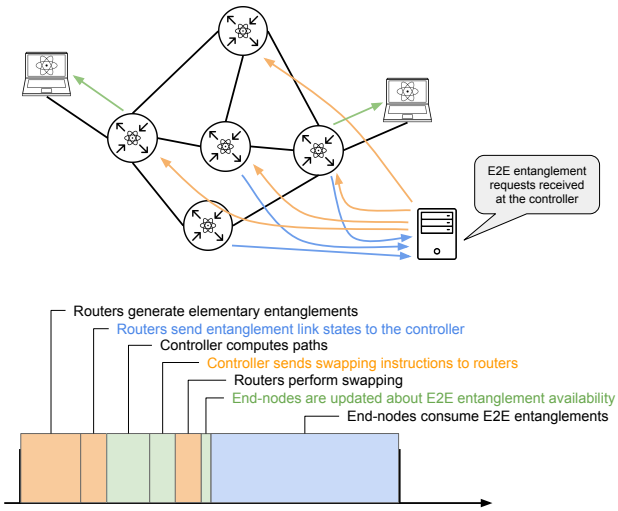

FIGURE 6. Reactive routing with a central controller and operations executed during a time slot (adapted from [53]).

As discussed in Section II (see footnote 1), a time slot typically includes the entanglement attempt/creation, the entanglement outcomes transmission, the path computation, the route installation, and the swapping execution time. The remaining time is used to consume the E2E entanglement by the layer above (e.g., application). Hence, the duration of a time slot depends on the hardware characteristics and entanglement protocol. It is upper bounded by the entanglement creation delay, plus the entanglement lifetime (i.e., quantum memories coherence time). Note that these operations are executed sequentially within the time slot without further synchronization. Note that the entanglement coherence time must be long enough for the path computation and route installation to take place, and allow enough time for the E2E entanglement to be consumed. However, nodes have a limited time to create elementary entanglements and must wait until the next time slot to re-attempt failed entanglements. This may lead to non-optimal utilization of network resources.

Reactive routing can also be designed in a distributed system, where routers rely only on the local knowledge of the logical topology to compute paths. The routing proposed in [29] uses a path-finding mechanism in which all elementary entanglements are consumed in each attempt to establish an E2E entanglement. A more elaborated distributed path selection is designed in [54] where nodes collectively maintain the logical topology as a distributed graph. As new entanglements arrive and others expire, each node interacts with its neighbors to join the graph by selecting a root node based on link cost (e.g., fidelity). Once the node has joined a graph, it has a route toward the graph root which is used to select entanglements to swap for each E2E entanglement request.

Figure 7 shows a simplified view of a distributed reactive routing scheme where two end nodes want to establish an E2E entanglement. Since each router is only aware of the status of entanglements with its neighbors, the routers need to perform swapping when possible to increase the probability of establishing an E2E entanglement. By consuming all entanglements in the logical topology, the E2E entanglement is established. With a more advanced algorithm, a path can then be identified within the distributed graph where each router selects the entanglements to swap.

Note that although reactive routing operates on a partial or global logical topology, it may assume that the global physical topology is also known to the routing algorithm [38], [39], [53], [55], [56].

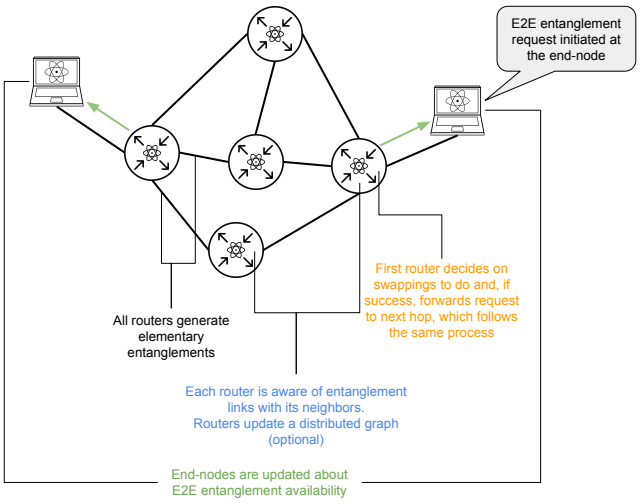

FIGURE 7. Simplified view of a reactive distributed routing.

Path computation metrics commonly used in reactive routing literature include hop count [38] and link fidelity [55] due to their simplicity and low computational complexity. However, more elaborated path computation algorithms may take into account a non-linear combination of relevant metrics to maximize throughput [39], [53], [57], [58], such as decoherence times, physical distance between routers, classical communication latency, requested fidelity, purification costs, and swapping, gates, and measurement error rates.

Selecting paths based on already created entanglements allows the routing algorithm to support purification schemes

[18], swapping strategies, and resource management. However, the logical topology graph formed by the entangled links must be connected enough to satisfy the E2E entanglement requests. Otherwise, most requests must wait until the next period (in the best case) to be satisfied. Therefore, entanglement technology should provide an efficient entanglement success rate to guarantee sufficiently connected logical graphs, or a high enough frequency to make multiple attempts within a short time period.

The impact of entanglement generation rate and success rate is still to be investigated in order to determine the technology characteristics required for the reactive routing to be viable. Authors in [59] study the entanglement generation rate in a quantum network and demonstrate that the performance of E2E entanglement distribution may be affected by the scheduling of entanglement generation and swapping.

Since entanglements are already created at path computation time, the delay to satisfy E2E entanglement requests may be reduced in reactive routing. However, this approach faces significant concerns regarding its feasibility and scalability. Communicating with a controller or neighbor routers for every E2E entanglement creation introduces latency and overhead, challenging the network's efficiency and practicality.

3) Hybrid

Elements of proactive and reactive schemes can be combined to leverage their respective strengths while mitigating their weaknesses. These hybrid approaches may use pre-shared entanglements or path selection concurrent to entanglement generation. For instance, virtual routing pre-establishes virtual topologies using higher-level entanglements to simplify path computation and reduce latency, while opportunistic routing dynamically selects paths at each node based on local information and ongoing entanglement attempts.

a: Virtual:

Entanglements can be created between any pair of non-adjacent routers using swapping at intermediate nodes. This allows the network to form arbitrary virtual topologies with entanglements that extend beyond one physical link. Such entanglements are referred to as virtual links or l-level entanglements [7], [60]. For a l-level entanglement, the hop distance between its two end nodes x and y is 2^{l-1} [61]. Elementary entanglements are considered 1-level entanglements.

Virtual links can be combined, or with elementary entanglements, to create E2E entanglements as illustrated in Figure 8. Virtual links can be chosen deterministically or randomly [51], allowing the routing problem to be divided into two sub-problems. On the one hand, virtual links are (pre-)calculated and created to form the virtual topology [62]. On the other hand, the routing algorithm uses the virtual topology to calculate the paths for requested E2E entanglements.

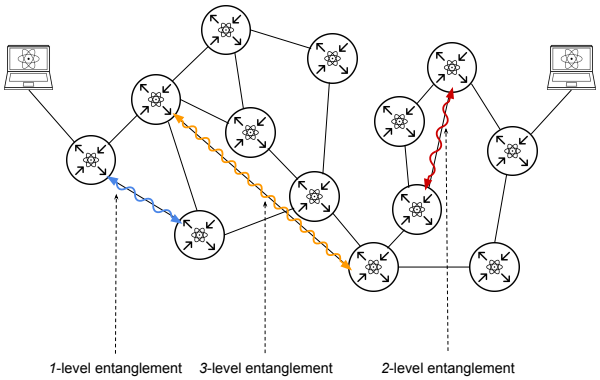

FIGURE 8. Virtual routing on a quantum network with various levels of entanglement links.

Using virtual links can increase the connectivity of the network. It also reduces the diameter of the network topology, which may reduce the complexity and execution time of the routing algorithm. The study [63] showed that an efficient construction of the virtual topology can reduce the latency of E2E entanglement creation. However, virtual entanglements expire and need to be regenerated just like elementary entanglements. This can be a time-consuming and challenging process that involves the generation and swapping of entanglements.

Virtual routing approaches are proposed in [60], [61], [64]. The work in [60] starts by establishing a physical topology and categorizing the nodes based on their capabilities to generate entanglements. Following this, a virtual topology is computed in a decentralized manner, employing the concept of a small world network as outlined by Kleinberg [65]. In the virtual topology graph, routing is performed using a distributed greedy algorithm with the entanglement success probability as a metric. The same approach is adopted in [64] using a metric derived from entanglement throughput statistics.

Researchers in [51] adopt a similar approach, but introduce a maximum limit on the distance of a pre-shared virtual link relative to the physical links. Additionally, they set a cap on the storage duration for the entanglement link. Considering a network that continuously generates entanglements, they designed hybrid routing algorithms that select a next hop even if no virtual link is established yet (similar to proactive routing). In this case, the entanglement generation is attempted over the selected virtual link.

This approach helps mitigate the downside of reactive routing resulting from its sub-optimal resource utilization. According to the reported results, when there is only one request in the network, relying on the continuously generated entanglements yields a lower latency than creating new ones. In the case of multiple concurrent requests, however, relying only on continuously generated entanglements can exhaust pre-shared entanglements before creating new ones to replace them; hence, requesting new entanglements can improve

performance.

Another virtual routing approach that addresses the lack of entanglements in reactive routing is proposed in [66], where the network accumulates l-level entanglements between nodes with storage capacity (either randomly or based on the degree of the node) and makes them available when needed.

In general, virtual routing can be seen as an enhancement of the reactive scheme, which can combine the advantages of proactive and reactive routing while limiting their shortcomings. On the one hand, the latency to establish E2E entanglements due to the entanglement creation in proactive routing can be reduced by the continuous generation of virtual links pre-configured on routers. On the other hand, virtual entanglements can artificially improve the availability of entanglements (e.g., [51], [66]) and accommodate path computation since the virtual topology to process can be simplified.

b: Opportunistic:

In opportunistic routing, the next entanglement (i.e., next hop) is selected at each router based on the results of the elementary entanglements attempted with a set of selected neighbors. One may think of quantum opportunistic routing as a scheme in which path computation and entanglement generation are executed in parallel and in a distributed manner at each hop along the path.

Note that one can distinguish opportunistic routing from reactive routing approaches through the fact that opportunistic routing can only be distributed, and path selection is done at each node without any coordination between nodes.

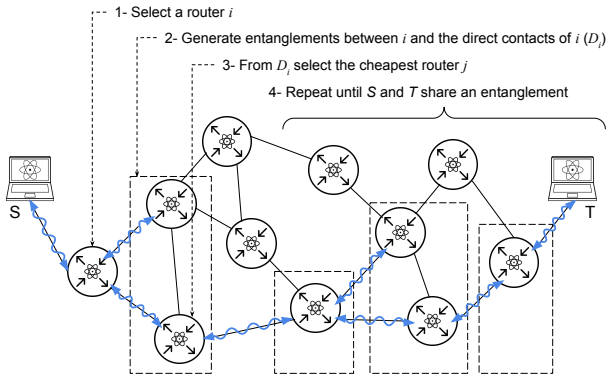

FIGURE 9. Opportunistic routing proposed in [67].

Figure 9 illustrates an opportunistic routing scheme as proposed in [67]. Starting with the first end node of the requested E2E entanglement, each node selects a distribution set of nodes among its neighbors. The distribution set of node i is made up of nodes with which it shares entanglements in the direction of the other end node of the request. The node i then selects one neighbor to forward the request based on criteria such as entanglement fidelity. The process is repeated at each hop until the entanglement is created with the end node. Note that opportunistic routing handles only the selection of entanglement links to create along the path and does not assume any specific swapping approach. Swapping can use a certain order or policy after all entanglement links are created, or the swapping can be executed opportunistically while forwarding the E2E entanglement request (see Section III-B). Quantum opportunistic routing typically considers a partial physical topology or the global physical topology if shared beforehand. However, the knowledge of logical topology is always local; within the next neighbor or k neighbors.

B. FORWARDING PHASE

Forwarding encompasses two main steps: the *external phase* for generating elementary entanglements and the *internal phase* for establishing E2E entanglements via swapping. The execution of these steps may differ depending on the routing scheme and supported services. In the proactive (and often virtual) routing, the external phase occurs after path computation. In contrast, in reactive routing, it happens before or concurrently to path selection.

The internal and external phases can be executed synchronously or asynchronously, as illustrated in Figure 10. In synchronous mode, nodes perform the external phase within a fixed duration, followed by the internal phase. If a swapping attempt fails, all nodes involved in the path re-execute the external phase again. Note that elementary entanglements do not need to succeed at the same time, but within the duration of the internal phase. In the asynchronous mode, each pair of nodes creates elementary entanglements independently as soon as quantum memories are available, and swapping is performed immediately when conditions are met. If a swapping attempt fails, only the involved entanglements are regenerated, while the other swapping attempts continue. Consequently, the asynchronous mode provides better resource utilization and achieves higher throughput compared to synchronized entanglement generation and swapping.

The forwarding phase includes two main operations: swapping and purification. These two operations may not be considered as a single functionality, as purification may be performed at the elementary entanglement generation phase, after swapping, or a combination of both. Additionally, purification can also be applied on E2E entanglements, although this may be handled by protocols above the network layer, such as transport or application. Lastly, different types of path reliability may be included depending on the supported QoS, as discussed below.

1) Swapping

The fidelity of E2E entanglements drops with each swapping operation [31] and the time elapsed before applying the swapping correction at the end-node [34]. Additionally, the E2E entanglement throughput may be affected by the order in which swappings are performed [32]. In the following, we will discuss different swapping approaches studied in the literature. In general, swapping can be either memoryless or based on quantum memory.

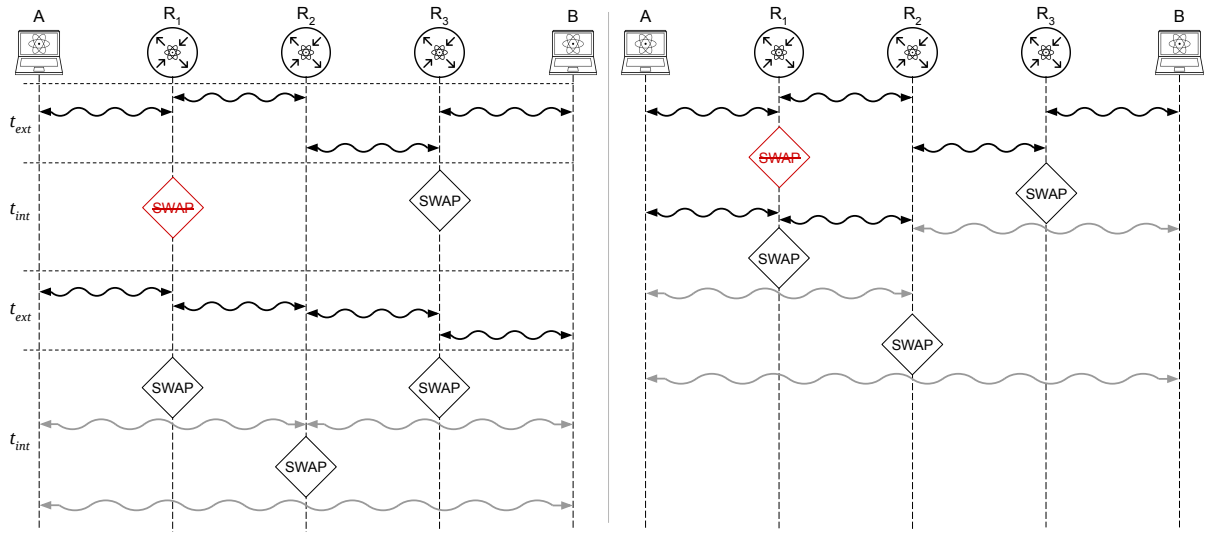

FIGURE 10. Entanglement generation and swapping. Synchronous mode (left) and asynchronous mode (right). The crossed SWAP indicates a failed swapping

a: Memory-less swapping:

This is often referred to as *synchronous* swapping or *waitless* protocol [31]. Here, strict synchronization among routers is required so that all entanglements along a path must be successfully created at the same time and all swapping operations can be carried out simultaneously. Thus, memoryless swapping is unheralded.[13] Clearly, an E2E entanglement can only be generated if all underlying processes are synchronized and succeed; otherwise, any failure will cause the whole process to restart from the generation of elementary entanglements. As a result, although this approach can provide E2E entanglements with high fidelity w hen s uccessful, i ts g eneration r ate i s v ery low, hindering its practicality.

b: Memory-based swapping:

This is also known as *asynchronous* swapping or *waiting* protocol [31], where a qubit of an entangled Bell pair may wait in memory for its swapping counterpart to become available and certain conditions to be met so that a swapping operation can be carried out. This is possible because coherence times of several seconds to minutes (and even hours, depending on the technology) have been demonstrated [69]. As a result, swapping can happen at different times at intermediate nodes along a path following certain rules and orders (including random ones).[14] Thus, different swapping policies can lead to different wait times, which in turn affect fidelity a nd t hroughput o f E 2E entanglements.

The asynchronous mode is expected to perform better than the synchronous mode because it does not require all the entanglements to succeed at the same time and allows the

[13]Of course, the swapping results will still need to be communicated to end-nodes via classical signaling in either centralized or distributed fashion.

[14]Multiple paths going through the same sequence of routers might t swapping policies as well.

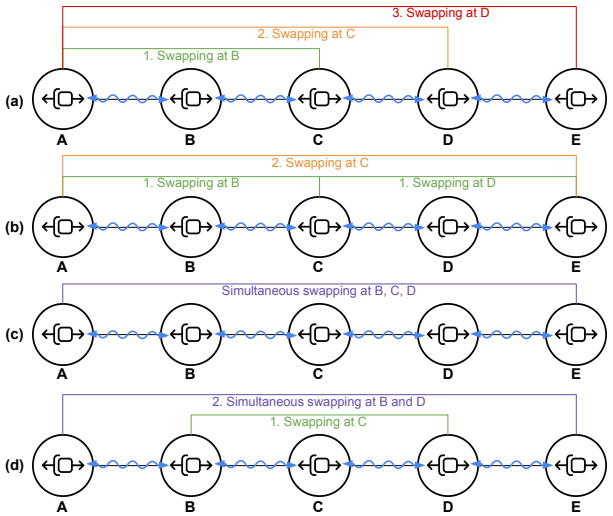

FIGURE 11. Examples of swapping orders in a single path without multi-hop purification (adapted from [68]).

regeneration of only the failed entanglements instead of all entanglements. Additionally, it may be useful for routers to know if a swapping operation has failed as soon as possible to avoid unnecessary waiting (at the cost of additional classical signaling). In the following, we will focus on asynchronous swapping and discuss several common swapping schemes using an example of a path with $n = 4$ hops shown in Figure 11 above.

- **Sequential:** Nodes perform swapping sequentially from left to right (or right to left) [70]; see Figure 11(a). This scheme requires $n - 1$ steps with one node performing swapping operations in each step after its immediate neighbor. As a result, this scheme yields the highest wait time due to swapping steps. However, it is often the

most useful scheme in a path-finding algorithm based on simple metrics (such as hop count, node distance, link entanglement creation rate, or link bottleneck capacity as described in Section III-C2 above) because the contribution of a new hop to the path length when extending the path is often much easier to find compared to other schemes below.

- **Doubling:** This is one of the most studied schemes, where the order of swapping corresponds to a balanced binary tree of height $\lceil \log_2 n \rceil$ and nodes on the same level can perform swapping in parallel; see e.g., [12], [35], [36], [71].[15] Thus, this scheme takes only $\lceil \log_2 n \rceil$ steps. Figure 11(b) shows the case for $n = 4$ where Node B and Node D perform swapping first, and Node C connects the end nodes afterward. Note that when n is a power of 2, the binary tree is called a perfect binary tree and, assuming sufficiently long decoherence time and repeaters with sufficiently large memory, the doubling scheme gives the optimal generation rate for homogeneous chains [36], [72]. For a general path, however, this scheme is not necessarily rate-optimal.
- **Parallel:** When all elementary entanglements have been successfully established on a path, all repeater nodes can perform entanglement swapping simultaneously and independently in the sense that one node's swapping does not depend on the swapping results of others, hence unheralded swapping; see Figure 11(c). This scheme is more suitable for reactive routing as it takes only one step to finish, but its generation rate may be low because, taking Figure 11(c) for example, if one swapping fails, the whole process needs to restart from scratch (from generation of all elementary entanglements just like in synchronous swapping).
- **Ad-hoc:** Entanglement swapping is performed based mainly on the availability of local resources (locally adjacent elementary entanglements) instead of following any predefined order. A popular policy in this scheme is often referred to as *opportunistic* [48] or *swap-as-soon-as-possible* [73]. Here, the order and the number of swapping steps can vary from execution to execution; as a result, any order in Figure 11 can happen. This policy is the default in opportunistic routing but can be adopted in the other routing approaches. It allows part of the waiting time for the entanglement generation to be used to perform some of the swapping operations. Hence, resources are consumed and freed more quickly, resulting in less waiting time for requests that are waiting for the completion of prior requests. The scheme can be applied after the whole path is computed by the routing phase (in proactive routing) or it can be combined with the opportunistic routing approach; see also [48].
- **Heuristic:** For a general non-homogeneous path, entanglement swapping can be performed in a particular order based on certain rules in each time slot; see Figure 11-(d) for example. Oftentimes, this is done to optimize certain performance objectives such as throughput or generation latency [32], [36], [70]. However, since optimizing swapping orders is combinatoric, this scheme is more computationally expensive than the other schemes described above, even when approximation/relaxation is in place. As a result, this scheme is more suitable for proactive routing where swapping orders are usually fixed and determined along with path computations. Recently, [74] has used Reinforcement Learning (RL) to design swapping policies for optimizing either wait time or fidelity of the end-to-end entanglement in a path with single channels (i.e., unit capacity) and cutoff (coherence) time of quantum memories. In this case, the swapping order varies between time slots as each swapping action depends not only on the policy but also on the state of the whole path.

Except for ad-hoc and certain heuristic (e.g., RL-based) swapping, all the other schemes need the routers to agree on a particular order to perform swapping along each path. Since these orders are predefined, they can be considered *static-swapping* policies. In this case, swapping instructions (excluding entanglement readiness messages between routers) for parallel schemes contain only the qubits to swap along the path, while those for other schemes must include some schedule or rules for the steps in addition to the qubits to swap.

Ad-hoc schemes and certain heuristic policies, such as RL-based policies can be seen as *dynamic swapping*; or more precisely, the swapping sequence is a stochastic process driven by the randomness of link entanglement generation and node entanglement swapping operations, as well as the swapping policy. Note however that swapping instructions for ad-hoc schemes can be as simple as satisfying certain local conditions [75], while RL-based swapping policies as in [74] must rely on non-local instructions because swapping actions in each step depend on the current state of the whole path.

Finally, it is important to note that optimal swapping policies cannot be determined without specifying practical assumptions and objectives. Different policies result in varying wait times due to swapping conditions and classical communication delays, which affect success rates, error rates, and reliability. On the one hand, longer wait times can increase the likelihood of decoherence, degrading the fidelity of entangled states and increasing the error rates. Of course, if the wait time exceeds the coherence time of the memory, the stored qubits will become corrupted, rendering them useless. However, if the coherence time is sufficiently long, policies with longer wait times may actually achieve higher success rates by making better swapping (and possibly purification) decisions, thus improving E2E entanglement throughput. Therefore, we believe that while

[15]This scheme is also referred to as *parallel order* in some references such as [32]; we refer to it as *doubling* instead so as to be consistent with the literature.

future implementation of quantum error correction techniques may mitigate decoherence effects and enhance reliability, in the short term, developing entanglement swapping and purification policies that not only optimize throughputs, but also consider wait times and efficiently manage quantum memory can improve both error rates and reliability.

2) Fidelity and Purification

Many routing algorithms operate under the assumption that entanglement links are successfully established and maintained in their ideal state. However, more realistic models take into account the quality of these links, acknowledging the possibility of creating lower-quality entanglements. To offer a more accurate assessment of path quality, some approaches incorporate entanglement fidelity in the routing algorithm and forwarding process.

a: Fidelity Estimation

Accurately estimating the fidelity of entanglements generated on a physical quantum channel is a challenging task, as it requires the generation and measurement of several entanglements to calculate the fidelity. Quantum State Tomography (QST) [76] offers complete information by reconstructing the density matrix of the prepared states through a comprehensive set of measurements on multiple copies [77], combined with statistical methods such as maximum likelihood estimation. Although QST is powerful for fully characterizing unknown states, it is computationally and experimentally intensive, with resource requirements scaling exponentially as the system grows. Direct Fidelity Estimation (DFE) [78] further reduces measurement needs by randomly sampling specific observables, directly computing fidelity between a prepared state and a target state without reconstructing the state. DFE is practical for states where particular components dominate, though it relies on a significant number of samples to maintain accuracy. Machine learning techniques [79], [80] have shown promise in improving DFE to support more general states with fewer measurements. Quantum State Verification (QSV) [81]–[83], in contrast, is a more resource-efficient approach to assess whether the fidelity of a prepared state exceeds a specified threshold [84]. Using adaptive measurement strategies, QSV scales better for large systems, though it provides a yes/no answer and is often state-specific. To estimate the entanglement fidelity without using any experimental resources, one may run simulations of network dynamics with detailed physical parameters to provide expectations (see the Appendix B for examples of simulation platforms).

Although the fidelity estimation and state verification methods may share similarities with purification, the former does not improve the fidelity of the remaining pairs, while the latter may identify errors and discard noisy pairs.

b: Passive Fidelity Support

Generated entanglements suffer from inherent imperfections due to hardware limitations. These imperfections are magnified during the entanglement swapping process, leading to a reduction in the quality of the resulting E2E entanglement [85]. Therefore, it is essential to provide high-fidelity entanglements to the swapping process to ensure the E2E entanglement meets the desired fidelity. Moreover, the noise introduced by quantum operations can be reduced with better hardware, although such improvements fall outside the scope of network protocols. Delivering high-fidelity entanglements requires a strategy that includes both hardware advancements and several optimizations at the protocol level.

During the routing phase, the length of the computed paths can be limited to meet the requested fidelity [31]. Furthermore, the routing algorithm can decide on the required fidelity of elementary entanglements to ensure a sufficiently high fidelity of E2E entanglements [85]. Decoherence significantly impacts the fidelity of qubits in memory. This can be mitigated during forwarding by optimizing the classical signaling to minimize the duration that the qubits remain unused in memory [85]. These approaches improve the chances of achieving higher fidelity without actively handling purification on the created entanglements. Moreover, they do not guarantee the satisfaction of the E2E fidelity requirements. Hence, they can be considered as a *passive* fidelity support.

c: Active Fidelity Support

Advanced approaches to satisfy requested fidelity use purification on elementary entanglements to individually increase fidelity (purify-then-swap). Alternatively, purification can be used on entanglements generated after swapping (swap-then-purify). Since these approaches use purification, they are considered as *active* fidelity support techniques. However, improving the fidelity of entangled links may require several rounds of purification, which consumes many entanglements and requires classical signaling. Hence, systematically applying a predefined number of purification rounds on each link requires a large amount of entanglements while introducing more latency to the E2E entanglement generation.

To improve single-hop purification decisions, Shi et al. [86] studied the performance of various purification protocols [87]–[89] under various errors (e.g., qubit decoherence, measurement error) and proposed a module that dynamically selects the appropriate purification protocol and number of rounds, considering the target fidelity and available qubits.

In [58] a routing algorithm is proposed that performs purification based on the maximum hops before purification is needed to guarantee the fidelity of the E2E entanglement for a given path. Dawar et al. [90] analyze the resource requirements to establish a path with target fidelity under gate and measurement error probabilities, employing multiple rounds of purification after each entanglement swapping. They propose a non-recursive method for estimating the number of entanglement pairs needed along a path, enabling fast calculations for path selection.

Li et al. [39] propose a path selection in which purification is applied only on entanglement links with fidelity below the required E2E fidelity. Then, the links with sufficient fidelity

are included in the path computation. However, purifying elementary entanglements up to a certain threshold does not guarantee the satisfaction of the requested fidelity in the E2E entanglement, which would still result in resource wastage. In a more effective approach, the authors in [57] use a purification cost table to determine the minimum purification rounds per entanglement link to satisfy the required E2E fidelity.

More elaborate routing models propose to compute for each selected path a purification scheme indicating which entanglement links need to be purified and, optionally, the number of purification rounds. Considering noisy quantum operations and finite memory storage time, Victora et al. [91] studied a routing model in which paths are selected based on the entanglement generation rate and the optimal combination of links to purify and the purification rounds to maximize the E2E entanglement throughput. In [55], a purification scheme is calculated for each path to satisfy the requested fidelity, indicating the minimum number of entanglements to consume to purify each link. Path selection is based on an estimate of the fidelity of each link in the network.

Considering deterministic purification with quantum error correction, Patil et al. [92] propose a routing model that searches for each path all possible swapping and purification sequences to find the best fidelity of the E2E entanglement where purification is used only if it improves the fidelity.

3) Path Reliability

In first-generation quantum networks, reliability is the ability to deliver uninterrupted and stable E2E entanglements for the duration of a request. In other words, the E2E entanglements remain constant, so reliability may translate to reducing the probability of connection drops. It is evident that mitigating entanglement generation and swapping failures is critical to ensuring the reliability of E2E entanglements generated along a selected path. Therefore, various path recovery procedures are envisioned in the reviewed literature.

Due to time constraints imposed by the entanglement decoherence, a path recovery procedure may be more realistic in a distributed way where each router cooperates with its neighboring nodes to find alternative links. However, the effectiveness of distributed/local path recovery is limited because nodes may not be able to optimize the usage of existing entanglements with only neighboring knowledge. This may lead to more frequently fragmented paths and unused entanglements [30], [37].

Path recovery is more critical in proactive routing, where entangled links are not known at path computation time, which increases uncertainty in forwarding. During the forwarding phase, routers can use various information (e.g., the physical topology, entanglement links outcomes, neighbors, etc.) to create alternative/redundant entanglements or re-affect unused entanglements to complete the creation of E2E entanglements (see Figure 12).

Path recovery can be enhanced through its integration with the path computation [30], [37]. In this case, the routing algorithm can be designed to provide routers with alternative links or entire paths to use for path recovery.

The research [30] proposes path recovery for proactive routing based on local entanglement link-state outcome following entanglement generation. In the routing phase, nodes use the global topology knowledge to choose a consistent set of paths and address link failures with entanglement generation information during the forwarding phase. For each E2E entanglement request, they identify multiple major paths that can be fully reserved according to available resources, and partial paths which share resources with major paths and thus lack complete reservation. The entanglements are created for both path types, with the partial paths serving as a fallback during the forwarding phase to compensate for any failed entanglement in a major path. As illustrated in Figure 12 (center), a routing algorithm computes two concurrent paths for request A-B. Path ACDEB is the main one while path AC'D'DEB does not have enough resources but some of its entanglements can be created and used as a recovery path. In the forwarding phase, node D finds that the main path is disconnected. It chooses to route through AC'D'D, and swaps link DE with link DD', instead of CD-DE.

Authors in [37] assume that when sufficient network resources are available, the path computation should involve provisioning extra, potentially redundant resources for establishing entanglement links. This strategy ensures that in the event of failures in creating some entanglements, alternative ones can be utilized to establish E2E entanglements subsequently. To avoid creating extra entanglements on every quantum channel, an algorithm determines an optimal set of backup entanglement links to be created, considering resource and path constraints. This approach is depicted in Figure 12 (right) where an algorithm computes path ACDEB and provisions extra links entanglements (AC' and C'D) for redundancy. As entanglement CD failed, the backup links are used to connect A to B.

Some path recovery aspects can be supported by the forwarding phase without requiring special features from the path computation algorithm. Typically, entanglements that cannot be used to establish paths can be re-affected to complete other paths. These entangled but unused qubit pairs, called fragments [93], negatively impact resource utilization efficiency and reduce overall network throughput. Defragmentation in quantum networks presents significant challenges due to the unpredictable nature of entanglements and their limited lifetime, preventing link states from being propagated throughout the network. Consequently, nodes must make local decisions to connect entanglement links based on k-hop entanglement states as proposed in [93]. As depicted in Figure 12 (left), nodes E and F possess the information of selected paths and the states of entanglements within a few hops. This enables them to assign entanglement between E and F to path AB. This approach requires that path computation does not assign specific qubits to specific paths.

A similar strategy is adopted in [48] in which available entanglements that are not used to establish a path are affected to create other paths. For that, the outcome of all entanglement links is shared by each node to all other nodes within a few hops.

Note that path recovery is also useful in reactive and virtual routing to overcome unexpected failures that may occur due to memory issues or errors incurred during swapping and purification routines. To overcome link failures in the virtual routing scheme, the strategy proposed in [61] selects alternative replacement paths for primary paths. These replacement paths provide provisional routes to be used when the main path fails until all disrupted entanglement links are fully restored.

C. ROUTING ALGORITHMS

A typical objective of routing is to maximize the overall Entanglement Generation Rate (EGR); i.e., the number of E2E entangled pairs generated per unit of time. To this end, the routing algorithm includes the definition of a metric that is easy to calculate, general enough to be used independently of the physical technology, and reliably leads to choosing reasonable, if not optimal, paths.

The first-generation quantum repeater networks impose specific constraints that a routing algorithm must consider to effectively solve the entanglement routing problem. As discussed above, these constraints include the limited time available to utilize entanglement before it begins to decohere, the probabilistic nature of entanglement creation and swapping, the variable fidelity of entanglement which may necessitate one or multiple rounds of purification to meet path-fidelity requirements, and the exclusive reservation of qubits for a single entanglement path, preventing their concurrent consideration for multiple paths. In the following, we provide a classification of the routing algorithms and models proposed in the reviewed studies.

1) Dijkstra Algorithms

The applicability of the Dijkstra algorithm in quantum routing is certainly the oldest approach [12], [13]. The Dijkstra algorithm operates on the principle that the total cost of a path is derived from the sum of the costs associated with each edge along that path. This does not always match the formulation of the entanglement routing problem regarding quantum network characteristics. However, by defining an appropriate aggregated cost metric for paths, Dijkstra's algorithm offers a straightforward method for selecting paths. Using Dijkstra, one simple link cost could be the inverse of the link throughput, measured in seconds per Bell pair of a particular fidelity (e.g., in proactive routing) [12]. Although lower-level metrics such as quantum measurements and hardware operations serve as useful indicators to gauge the actual work expended to establish a path, they fall short as criteria for link prioritization [34]. This is because they mainly mirror the physical attributes of a link rather than the overall E2E entanglement distribution. Utilizing a variant of Dijkstra's algorithm, where the link cost is represented by the inverse of each hop's throughput, creates a relatively accurate correlation between the simplified path cost and the actual throughput, as well as between the cost and the quantum physical operations involved. This is because the use of the inverse of each hop's throughput as a metric should already take into account such lower-level physical parameters, as well as the fidelity of the entanglement generation. Furthermore, this approach can be implemented with a manageable level of computational complexity.

A variant of Dijkstra's algorithm is examined in [13] aiming at maximizing the E2E fidelity of entanglements. The study shows that the final E2E fidelity cannot be systematically inferred from the fidelity of each hop. Ideally, entanglement routing should be a function of both EGR and fidelity. Hence, more recent studies [30], [57] argue that reducing path selection to a simple shortest path problem may not always achieve optimal routing decisions. Recent studies frequently use Dijkstra in combination with other methods for path computation.

2) Path Search on Graph

Path search and enumeration in graphs are common approaches used to compute paths in both proactive and reactive routing with multiple concurrent requests. Path search algorithms such as Dijkstra, Yen's algorithm, and Bellman-Ford are extended with mechanisms to satisfy multiple requests without resource contention, provide fairness among requests, and optimize entanglement resource utilization. Yen's algorithm is used to find the k shortest paths between nodes, providing multiple alternative routes for route diversity and fault tolerance. Bellman-Ford handles graphs with negative weight edges, ensuring that the shortest paths are found while also detecting any negative weight cycles. However, these algorithms require knowledge of the global topology, whether the routing is implemented in a central controller or distributed across routers.

In the reactive routing scheme, the research [29], [38] uses the Node-Disjoint-Path (NDP) problem in the logical topology graph to find a set of paths linking a specified pair of nodes so that no two paths share a node. The hop count is used as a metric. In [38], several algorithms are proposed for the NDP problem: Sequential Multi-Path Scheduling Algorithm (SMPSA), Min-Cut-based Multi-Path Scheduling Algorithm (MCSA), and Random and Distance (physical) Scheduling Algorithm. In [57], path search with k-shortest path and an extended Dijkstra algorithm are used in the logical topology graph to find paths for a single request. The algorithm is extended with a resource allocation mechanism to support multiple E2E entanglement requests.

In proactive routing, the most commonly used metric in path search algorithms is the expected EGR [31], [37] estimated according to the network model similar to the one presented in Section II.

The research in [35] relies on path enumeration within a physical topology graph with the expected E2E entanglement

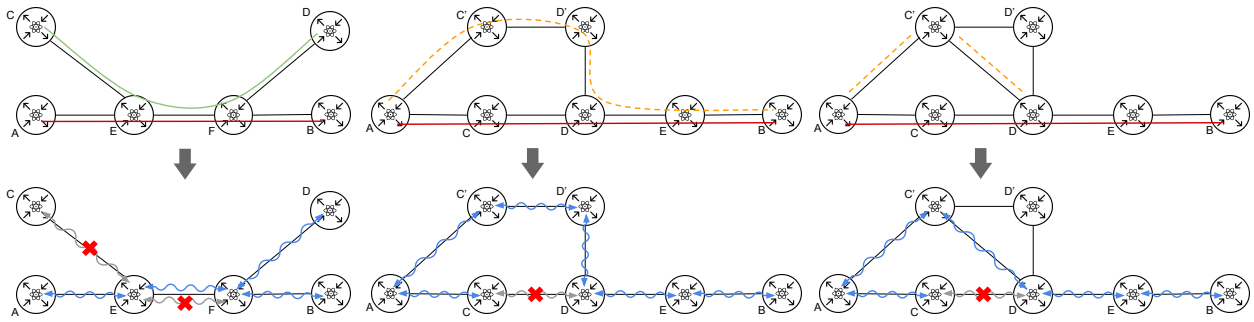

FIGURE 12. Path recovery strategies. Defragmentation (left), Secondary path (center), Redundant links (right).

throughput as a metric. The model incorporates low-level components involved in entanglement distribution, including optical fiber attenuation length, efficiency of Bell state measurements, and duration of atom pulses. Despite its advantages, the proposed routing scheme faces challenges in scaling across arbitrary topologies due to its computational time complexity.

Study [30] applies Yen's algorithm for identifying paths within the physical topology graph, alongside an augmented version of Dijkstra's algorithm to mitigate resource contention. It investigates multiple metrics for path evaluation. The sum of node distances, which sums up the lengths of channels at each hop, reflects path difficulty due to the exponential decrease in channel success rate with physical distance. The creation rate is calculated as the inverse of the product of success rates for channels along the path, offering insight into path width, while the bottleneck capacity is defined to favor paths with greater width. The creation rate is used to resolve ties among paths of equal width. However, the algorithms do not address the scheduling of demands for multi-path routing to optimize the use of resources.

3) Linear Programs

Linear programs are also used to formulate the routing problem whether or not fidelity requirements are included. The problem is usually formulated as mixed-integer linear programs (MILP) that capture the traffic pattern for all demands in the network. Solving exactly these problems for large-scale networks is impractical as they are usually NP-hard and thus cannot be solved in polynomial time. As a result, a relaxation step is usually employed in conjunction with some rounding techniques to obtain approximated solutions possibly with certain provable approximation ratios for the objective functions. Note, however, that as the time complexity of most LP solvers is polynomial in the problem size, this approach might not be scalable, especially under fidelity and decoherence time conditions.

Sometimes the routing problem is formulated as a multi-commodity flow optimization, in both proactive and reactive routing. Since the formulated problems are complex to resolve in a reasonable routing algorithm, a mix of linear programming optimization and path search in a graph is used in the routing algorithm. For example, [31] considers a multicommodity flow-based approach for maximizing the flow rate of entanglement distribution for all demands subject to certain fidelity requirements. The paper proposed to replace the fidelity requirements with hop limits and then formulate the problem as an edge-based LP with $O(|\mathcal{V}||\mathcal{E}||\mathcal{R}|)$ variables and $O(|\mathcal{V}|^2|\mathcal{E}||\mathcal{R}|)$ constraints, where $\mathcal{R}$ is the set of E2E entanglement requests. The solution to this LP can then be converted to path selection and rate allocation using an algorithm with a time complexity of $O(|\mathcal{V}|^4|\mathcal{E}||\mathcal{R}|)$. The overall complexity of this approach appears to be high; in particular, for a network with 70 nodes, solving the edge-based LP alone already takes roughly 45 seconds.

In reactive routing, [55] combines the k-shortest paths found with Yen's algorithm and linear program resolution with link fidelity as a metric while including a purification scheme to satisfy the requested fidelity. In particular, this paper considers maximizing the network throughput defined as the number of entanglement connections among multiple E2E pairs subject to fidelity constraints. This is done by first preparing multiple candidate entanglement paths, determining optimal purification schemes, and then selecting a final set of entanglement paths that can maximize network throughput. The entanglement path selection problem therein is solved through iterative LP relaxation and incremental rounding, which results in high time complexity.

In proactive routing, [56] uses Yen's algorithm and linear programming resolution with the expected rate as a metric. Specifically, this paper aims to simultaneously maximize the number of user pairs and their combined expected throughput. This problem is formulated as two sequential MILP steps, the first of which maximizes the number of user pairs that can be served with a main routing path selected from a subset of paths, and the second step is to maximize the expected throughput of user pairs. Here, both steps involve solving LP relaxations in $O(|\mathcal{R}|^3|\mathcal{V}|^3 + |\mathcal{R}|^{4.7})$ and using a Branch-and-Bound technique for rounding. For a network with $|\mathcal{V}| = 200$ nodes, the runtime of the proposed approach is in the order of hundreds of seconds.

The study [37] also formulates the routing problem as a linear program and searches the shortest path using the expected EGR as a metric, where the routing protocol includes request scheduling and selects redundant paths to overcome link failures (see Section III-B3). Specifically, this study proposed to provide additional entanglement links for redundancy and then select paths and links to maximize the expected throughput of multiple E2E entanglement pairs. The link provisioning and entanglement path selection problems are formulated as MILP with moderate sizes and then solved using LP relaxation and randomized rounding techniques, while the entanglement link selection is heuristic simply based on the number of entanglement paths selected and probabilities of success. In simulations with networks of up to 500 nodes, their algorithm runs in tens of milliseconds mainly because of relatively small problem sizes (in terms of the number of variables and constraints).

4) Greedy Algorithms

Greedy routing algorithms operate by choosing the neighboring node that is closest to the destination according to a specific metric, ensuring that no node is selected more than once. Greedy algorithm variants are used to find near-optimal paths using local link-state knowledge in the reactive routing in [29], [51], [60], [61]. Greedy algorithms do not guarantee finding optimal paths, although they represent a solution to implement decentralized reactive routing.

As in a path search on a graph, greedy algorithms commonly use the hop count or the physical distance of the link as a metric. In [51], a greedy algorithm is applied assuming nodes with global physical topology knowledge and local entanglement knowledge, using the hop count as a metric. In [60], a virtual topology of the entanglement network is represented with a base graph where the Manhattan distance between the nodes corresponds to the probability of the existence of the entanglement. Using the Manhattan distance function on the base graph, a greedy routing finds the shortest path. Similarly to [60], the routing algorithm in [61] finds the shortest NDP for a temporary replacement path in case of memory failure. In [29], to find multiple paths for a request, a greedy routing considers the subgraph induced by the successfully generated entanglements and the repeater nodes and finds in it the shortest path connecting the end nodes using hop counts. Then, all the links of the path are pruned from the subgraph, and another shortest path is computed in the pruned subgraph.

5) AI-based Routing

In proactive routing, [94] formulates the routing problem as a reinforcement learning problem. Path selection is implemented with a deep neural network to select a request to be fulfilled and the shortest path algorithm to find the best path.

Study [64] introduces a decentralized approach based on swarm intelligence for path selection in quantum networks. measures of entanglement success. Initially, the relevance of each entangled link is determined by an entanglement utility coefficient, reflecting the link's throughput statistics and its contribution to reaching the current node. Additionally, the attractiveness of a quantum node is defined by the link entanglement gradient coefficient, calculated from the deviation in entanglement throughput from statistics and the utility coefficient. This entanglement gradient is also applied to entire paths, creating a path entanglement gradient coefficient for routes composed of entangled links. Pathfinding is then achieved through the deployment of multiple threads that navigate local segments of the network topology, leveraging these defined metrics.

D. DISCUSSION

Table 1 summarizes the discussed routing approaches with their models and features, and Table 2 summarizes the forwarding approaches.

Overall, the routing phase of entanglement distribution seems to attract more interest among the reviewed studies, with a more exploration of the forwarding phase in recent works. The definition of the network topology appears as an important design aspect. The common assumption across the routing schemes of a knowledge of the global physical topology seems viable for intermediate-scale quantum networks. However, as networks scale, the feasibility and scalability of maintaining global physical topology knowledge (at each node or at a controller) become increasingly complex, hinting at a pivotal challenge for the evolution towards a quantum Internet [95].

Proactive routing's reliance on global physical topology knowledge at the time of routing stands out. It contrasts with reactive routing's focus on the logical topology, which is dynamically formed by successfully established entanglements. This logical post-entanglement routing is less dependent on the underlying, potentially very heterogeneous, hardware technology. Although much less explored, the virtual and opportunistic routing's potential for leveraging both global and local topology knowledge introduces a flexible framework, allowing for trade-offs between scalability, adaptability, and optimality in routing decisions.

Design strategies also vary. Both proactive and reactive routing can be realized through centralized or distributed systems, each with its advantages and challenges. Proactive routing benefits from not being constrained by the ephemeral nature of entanglement, allowing for flexibility in implementation and optimization. Conversely, reactive routing, while more abstracted, must contend with the challenges posed by entanglement's short existence. This does not prevent reactive schemes from being envisioned with both centralized and distributed systems. Path computation tends to favor the global knowledge of topology, either logical or physical, for feasibility and efficiency. However, the rapid pace at which the logical topology can evolve and the ephemeral nature of entanglements render approaches that rely solely on the logical topology less practical if not unrealistic. Hence,

Authors	Year	Routing	Metric	Phy. topo.	Log. topo.	Request	Path	Fidelity	Model
Shi et al.	2020	Proactive	Link rate	Global	N/A	Multiple	Single	No	PS
Caleffi et al.	2017	Proactive	E2E rate	Global	N/A	Single	Single	No	PS
Zeng et al.	2022	Proactive	Link rate	Global	N/A	Multiple	Single	No	LP
Chakraborty et al.	2020	Proactive	Link rate	Global	N/A	Multiple	Multiple	Passive	MCF
Le et al.	2022	Proactive	Link capacity	Global	N/A	Multiple	Single	No	ML
Zhao et al.	2021	Proactive	Link rate	Global	N/A	Multiple	Multiple	No	LP
Nguyen et al.	2022	Reactive	Hops	Global	Global	Multiple	Multiple	No	PS
Yang et al.	2024	Reactive	Hops	N/A	Local	Single	Multiple	No	PS
Pant et al.	2019	Reactive	Hops	Global	Local	Multiple	Multiple	No	PS
Cicconetti et al.	2021	Reactive	Hops/Distance	Global	Global	Multiple	Single	No	MCF
Li et al.	2021	Reactive	Hops	Global	Global	Multiple	Multiple	Link	PS/MCF
Zhao et al.	2022	Reactive	Fidelity	Global	Global	Single	Multiple	Scheme	PS/LP
Li et al.	2022	Reactive	Entang. cost	Global	Global	Multiple	Single	Passive	PS
Chakraborty et al.	2019	Virtual	Hops	Global	Local	Single	Single	No	Greedy
Gyongyosi et al.	2018	Virtual	Entang. prob.	Global	Local	Single	Single	No	Greedy
Gyongyosi et al.	2019	Virtual	Entang. prob.	Global	Local	Single	Single	No	Greedy/PS
Gyongyosi et al.	2017	Virtual	Entang. prob.	Local	Local	Multiple	Single	No	ML
Pouryousef et al.	2022	Virtual	Link rate	Global	Global	Multiple	Multiple	No	LP
Gyongyosi et al.	2019	Opportunistic	Fidelity/Coherence	Global	Local	Single	Single	No	Greedy

TABLE 1. Notable path computation approaches and routing algorithms. PS: Path Search, LP: Linear Program, MCF: Multi-commodity Flow Optimization, ML: Machine Learning

Authors	Year	Log. topo.	Swapping	Path recovery	Fidelity
Shi et al.	2020	Local	Heuristic	Yes	No
Li et al.	2022	Local	Sequential	No	No
Kozlowski et al.	2020	Local	Sequential	No	Passive
Li et al.	2020	Global	Heuristic	No	No
Farahbakhsh et al.	2022	Local	Ad-hoc	No	No
Farahbakhsh et al.	2021	Local	Heuristic	Yes	No
Zhao et al.	2021	Global	Heuristic	Yes	No
Wang et al.	2022	Local	Heuristic	No	No

TABLE 2. Notable forwarding approaches.

a hybrid approach that combines global management with localized distributed computation may be the most realistic.

The study of routing algorithms and models unveils a rich set of strategies covering various quantum networking aspects. Reactive routing strategies, particularly with local topology knowledge, emphasize metrics like hop count for a straightforward approach for path selection, with a service limited to a single path for one or a few requests. In contrast, when global topology knowledge is considered, more complex metrics such as channel capacity, E2E fidelity, and purification costs come into play, allowing deeper optimization of routing decisions for multiple paths and multiple concurrent requests. This underscores the importance of detailed topology knowledge (more than in classical) on quantum routing efficiency and features. Proactive routing, with an emphasis on the expected entanglement rate as a key metric, allows for more optimization of resources. Advanced algorithms, such as k-shortest paths and multi-commodity flow, are suitable to support multi-request scheduling and multi-path routing, further highlighting the adaptability of the proactive routing approach.

Regarding route installation and forwarding, the process of establishing E2E entanglements goes beyond the simplicity of traditional networking tasks, such as populating forwarding tables with route prefixes and matching routes based on the longest prefix. Instead, it involves the orchestration of point-to-point elementary entanglements to create E2E connections, a task that is inherently more complex and stateful. The state information and signaling required to handle this process is, therefore, significantly more challenging than basic forwarding protocols. The swapping policies and strategies, relatively less explored, present their own sets of challenges and strategic considerations. The degradation of entanglement fidelity with each swapping operation, alongside the varied impact of swapping strategies on E2E entanglement throughput and fidelity, emphasizes the need for a deeper understanding of these mechanisms and their implications for quantum routing tasks. Our intuition is that the degree of network heterogeneity, such as node capabilities and link quality, would reflect in the complexity of the swapping scheme, which can range from arbitrarily pre-defined orders (e.g., left-to-right), to topology-adapted orders, to predefined local (e.g., swap asap) or quasi-local (e.g., wider neighbor first) policies, to dynamic per-path strategies.

IV. PROTOCOLS FOR ENTANGLEMENT ROUTING

In this section, we adopt an engineering perspective by exploring how classical networking protocols and architectures could be adapted to support the entanglement routing approaches discussed earlier. Although classical routing protocols cannot be directly applied to entanglement routing, this discussion aims to present a realistic view of the practical strategies that could be employed to create a suite of protocols for routing entanglements in quantum networks.

A. DISTRIBUTED PACKET SWITCHING

Distributed wired packet-switching networks rely on a set of protocols to learn routes to destinations. Typical routing protocols include Border Gateway Protocol (BGP) where routes are selected based on deterministic criteria such as distance or origin from the received route advertisements. BGP provides path selection without global knowledge of the topology and can scale to large networks with many routes. Open Shortest Path First (OSPF), on the other hand, is used for routing in smaller networks, where routers share information about their links with all other routers in the network. This ensures that all routers have a consistent view of the global topology and each router uses the Dijkstra algorithm to calculate the paths. These protocols are sufficient to realize a functioning packet-switching classical network providing best-effort forwarding.

Quantum routing algorithms that rely on simple cost metrics, such as link entanglement throughput with global physical topology knowledge, can be supported by adapting the OSPF protocol to proactive scenarios with limited efficiency, such as a small number of concurrent requests with a single path per request without fidelity guarantees [31]. However, many routing algorithms assume that the network state shared among routers needs to extend beyond simple link cost (e.g., link fidelity [96], E2E capacity [97]) to select high-quality paths, and may require more frequent updates than in classical networks [36], [72]. This is the case in QBGP [80], [98] where the BGP protocol is adapted to entanglement routing, but the path selection rules had to be replaced by a more dynamic decision process. At the request time, QBGP adapts its path preference according to link fidelity using an online learning algorithm instead of deterministic rules.

Routing solutions can also be borrowed from the wireless world. Similarly to quantum networks, routing in wireless multi-hop networks such as wireless sensor networks often works with partial knowledge of the network topology and handles intermittent connectivity. The Routing Protocol for Low-Power and Lossy Networks (RPL) is designed for Low-Power and Lossy Networks (LLNs). RPL maintains a distributed Destination-Oriented Directed Acyclic Graph (DODAG) to organize nodes in a tree-like structure where each node has a single path to the root. The Asynchronous Entanglement Routing [54] adopts a similar approach to RPL for distributed entanglement routing in the reactive scheme. When a node receives an E2E entanglement request, it determines the path to the destination and attempts a swapping between its previous and next entanglement links. If the swapping succeeds, the node forwards the request to the next hop, which repeats the same process until the destination is reached. Otherwise, the node waits until necessary entanglements are available to continue. This approach shows an effective distributed protocol for reactive entanglement routing but still requires high entanglement coherence times and generation rates to be effective.

With distributed protocols, the E2E entanglement request is forwarded to the next hop parallel to the creation of the entanglement link (similar to opportunistic routing [67]). The swapping measurements are also routed along the path to the end nodes [64], [80] (see Figure 13). Hence, this scheme requires memory-based swapping as entanglement creation along the path is asynchronous. Given the various swapping schemes available in memory-based scenarios, the relatively limited metrics supported in distributed routing may be overcome with a more advanced swapping policy.

In another routing approach discussed below, quantum versions of packet-switching distributed routing protocols can serve to pre-share network state information among the routers to compute paths for more robust and resilient protocols, similar to the classical virtual circuit switching architectures.

B. VIRTUAL CIRCUIT SWITCHING

In Virtual Circuit Switching (VCS), paths are established before data transfer begins. This connection-oriented architecture is widely used in classical networks through data-plane protocols such as Multiprotocol Label Switching (MPLS) which forwards packets based on their labels and pre-installed forwarding instructions. Generalized Multiprotocol Label Switching (GMPLS) extends the notion of labels to represent packet flows, optical fibers or wavelengths, timeslots, etc. The label-switched path scheme represents a natural choice to manage E2E entanglement paths [99].

VCS protocols use signaling mechanisms to negotiate paths and install labels. Resource Reservation Protocol Traffic Engineering (RSVP-TE) is a signaling protocol that uses paths computed via distributed routing such as OSPF with traffic engineering extension (OSPF-TE). OSPF-TE distributes information on link attributes such as bandwidth, delay, administrative constraints, and available resources, and can use constrained shortest path first (CSPF) algorithms to take these factors into account to compute traffic-engineered paths for RSVP-TE.

With a similar architecture, the RuleSet protocol [95] uses a two-pass communication initiated from a requester node to provision E2E entanglements. The outbound request is forwarded to the responder based on a standard next-hop table while collecting information about the nodes and links it traverses. The responder node replies with a set of rules for each intermediate node along the path. A rule has conditions (e.g., entanglement created, timeout) and corresponding actions (e.g., measure qubit). The RuleSet approach uses an outbound request to avoid the drawbacks of relying on globally pre-shared topology state via quantum adaptations of distributed routing protocols [100].

Building on a similar architecture to RuleSet, the REDiP protocol [101] provides a way to specify swapping orders (as node ranks) and purification instructions (i.e., purification rounds to be performed at each rank) during the connection phase. The authors also provide elaborated guidelines and metrics on defining swapping and purification strategies to meet E2E fidelity requirements and maximize throughput.

Segment Routing (SR) is an alternative to RSVP-TE in which routing instructions are carried within packet headers, eliminating the need for a per-flow state in intermediate routers. This design choice reduces the need for additional signaling protocols and simplifies the architecture. SR leverages a stateless model, with the head-end router managing paths from either centralized or distributed routing. Quantum-SR [52] adapts the SR mechanism to provision E2E entanglement paths using centralized routing to select paths based on topology state shared via distributed routing protocols such as OSPF.

Entanglement distribution protocols inspired by VCS are considered more adapted for the current and near-term technologies (e.g., memory-less repeaters, short coherence time, limited qubits) [99]. The path reservation process allows for both memory-less and memory-based swapping. Moreover, by reserving the entire path beforehand, this approach can support more elaborated path computation algorithms with E2E entanglement throughput [36], path length [31], and multipath [97] while providing routers with the capability to enforce specific paths through the network.

C. SOFTWARE DEFINED NETWORKING

Given the complexity of quantum path selection and the diversity of parameters it considers, many routing approaches are developed assuming a Software Defined Networking (SDN) architecture for various scenarios [38], [102]–[105]. This brings to quantum networks a familiar and easy-to-design framework which positively impacted classical networking. The controller can dynamically select paths in response to changing network conditions, such as link quality, demands, congestion, or failures.

SDN is most commonly considered in the reactive entanglement routing scenario, where elementary entanglement outcomes are collected by the controller, which uses them to compute paths for E2E entanglement requests. In the proactive scenario, the SDN controller collects the network topology and provides path computation in the same way that segment routing can rely on centralized routing to install forwarding instructions [52]. The controller can also be used to schedule the E2E entanglements to create, similar to the Quantum-adapted Time Sensitive Networks architecture envisioned in [106].

As mentioned, factors other than throughput per fidelity can be relevant in path selection, such as channel capacities and their distribution (especially for swapping strategies), multiplexing, and routing algorithm fairness. Moreover, distributed traffic engineering is usually sub-optimal in classical networks [107], [108] and one could expect this to be exacerbated in quantum networks. Distributed routing protocols and virtual circuit switching could certainly benefit from the global view of the SDN architecture for centralized traffic engineering [109], [110]. Rather than trying to dynamically control the network's logical topology and E2E entanglement requests, the SDN controller could be leveraged to understand communication demands, manage network resources, predict failures, select adapted swapping policies [74], and efficiently distribute entanglements in the virtual routing scheme [61].

D. SWAPPING

Entanglement swapping can be viewed as a distributed computation task performed on a reserved path. Figure 13 illustrates the implementation of a *swap-as-soon-as-possible* policy along a path of three routers between the end nodes A and B as commonly envisioned [85], [101]. Router R1 is the first to get two neighboring entanglements available, swaps immediately, and sends a `SWAP_UPDATE` to R2 containing the swapping measurement. Upon receiving the message, R2 waits until an upstream entanglement is available and swaps it with the entanglement specified in the message received from R1. Then, R2 adds the swapping measurement to R1's `SWAP_UPDATE` and forwards it to R3. When R2 receives the SWAP_UPDATE from R3 (which swapped its local entanglements in the meantime), it can immediately include its swapping measurement and forward it to R1. This process continues until both `SWAP_UPDATE`s are delivered to end nodes A and B.

Other swapping policies or orders can be implemented in this scheme using specific swapping conditions instead of "*as soon as upstream and downstream entanglements are available*". For example, a swapping condition in R2 specifying the upstream entanglement only with B will prevent R2 from swapping until it receives a SWAP_UPDATE from R3, thereby achieving a *doubling* swapping policy.

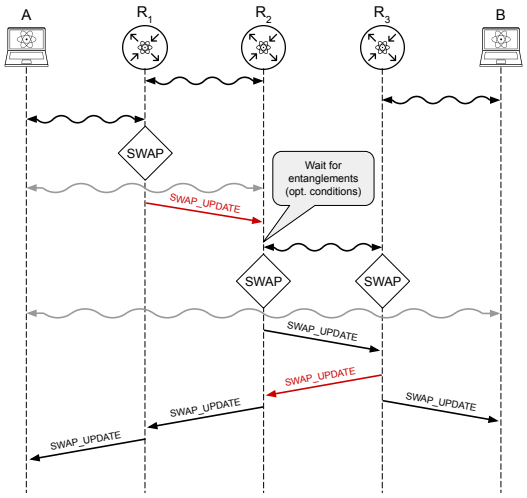

FIGURE 13. Swapping process along a path (adapted from [101]). Dark wavy lines indicate elementary entanglements, and light wavy lines indicate post-swapping entanglements.

V. OPEN QUESTIONS

A. NETWORK DESIGN AND TOPOLOGY

Identifying suitable topologies for quantum networks remains unresolved, as their effectiveness varies based on specific application needs and technological limitations [111], [112].

Different topologies, such as rings, stars, meshes, and fully connected networks, offer varying trade-offs in terms of routing complexity, scalability, resource requirements, and entanglement distribution efficiency. An open question is determining which topologies perform optimally in different scenarios, such as long-distance communication versus distributed quantum computing.

Moreover, unlike classical networks, the logical topologies in quantum networks can change dynamically due to the probabilistic nature of entanglement generation, decoherence, and swapping. This dynamism allows for logical topologies that can change on the fly or be programmed to optimize network performance for specific tasks. For instance, the ability to pre-share entanglement links can enable the network to be tailored to current communication demands and opens unique approaches to circumvent faulty nodes and links [113]. Designing routing protocols that can efficiently handle these dynamic and programmable logical topologies is an open challenge. Such protocols must account for the transient availability of entanglement links and adapt to the continuously changing network state while optimizing resource utilization and maintaining high fidelity of quantum information.

B. NETWORK AND HARDWARE HETEROGENEITY

Ensuring interoperability and optimality across highly heterogeneous components of intermediate-scale quantum networks is considerably challenging [114]. This heterogeneity impacts the abstraction level of routing models regarding the hardware, operating wavelengths, physical link architectures, and multiplexing schemes to ensure efficient interconnection of quantum devices and networks, facilitate maintenance and upgrades, and minimize communication disruptions and inefficiencies.

Ideally, routing protocols and architectures should be designed to ensure the continuity and interoperability of first-generation quantum networks with upcoming (different) generations. This may concern the appearance of second- and third-generation repeaters (that is, changing from the circuit-switched to packet-switched quantum paradigm) [115] or the integration of multipartite entanglements and all-photonic repeaters [116], [117]. Quantum network design may benefit from the Recursive Inter Network Architecture (RINA), an approach that treats the network as a single, recursive programmable layer, to manage heterogeneity and scalability and envisioned in [95].

C. ROUTING AND FORWARDING INTERDEPENDENCE

The complexity of routing and forwarding, each composed of multiple functions and considering various network parameters, raises questions about their interdependence. While some routing approaches incorporate forwarding considerations to achieve higher performance, others treat them as separate functions linked by standard interfaces [103], [104].

Among the aspects that blur the lines between these functions is multiplexing. In quantum networks, multiplexing involves managing multiple quantum channels or qubits over the same physical resources, which is closely tied to memory management in quantum repeaters. Efficient multiplexing strategies are crucial for optimizing the utilization of limited quantum memory resources and minimizing waiting times for entanglement generation. This raises open questions about how routing protocols should account for memory management and multiplexing. Should routing decisions be influenced by the current state of quantum memories and multiplexing schemes, or should these concerns be handled separately? Additionally, the interaction between routing schemes and qubit allocation strategies remains relatively unexplored, particularly with respect to how they impact path computation, purification, and swapping operations.

D. RELIABILITY AND AVAILABILITY

Quantum networks face fundamental challenges in reliability and availability that differ significantly from classical networks. A key issue is that measuring a quantum state collapses it, destroying the very entanglement needed for quantum communication. This makes traditional network monitoring techniques, which rely on probing and measuring network states, inapplicable. As a result, ensuring resource availability in quantum networks is an open problem. Network operators cannot directly verify the presence or quality of entangled states without compromising them. This limitation complicates tasks such as fault detection, performance monitoring, and dynamic resource allocation. Addressing this challenge requires developing new methods to assess network conditions indirectly. Possible approaches include using classical signaling to infer quantum states, leveraging quantum non-destructive measurements, or designing protocols that incorporate redundancy and error correction to mitigate the effects of undetected failures.

Furthermore, the inability to measure quantum states affects reliability. Network failures, such as photon loss or decoherence, may go unnoticed until they impact communication outcomes. In addition to the path reliability strategies discussed in Section III-B3, developing approaches to enhance the resilience of quantum networks, such as adaptive routing that can respond to inferred network conditions, is a crucial area for future research.

E. EVALUATION AND PROTOTYPE IMPLEMENTATION

The absence of a reference architecture for quantum repeaters introduces difficulties in standardizing control and configuration interfaces, crucial to ensuring interoperability and scalability across quantum networks. Moreover, the lack of specialized hardware platforms further impedes the rapid prototyping and testing of repeater technologies. This makes it difficult to define clear assumptions on underlying mechanisms for repeaters (e.g., memory management, multiplexing, heterogeneity), which in turn leads to uncertainties

in designing routing models, signaling mechanisms, and interoperable protocols.

Evaluating proposed routing schemes through simulations or analytical methods [118], while useful, does not fully capture the complexities of a real-world implementation. Moreover, current tools are not designed for the quick evaluation of dynamic routing protocols and swapping strategies over various abstractions of entanglement distribution technologies. Consequently, many entanglement routing approaches are considered in isolation and not evaluated under realistic network scenarios such as random topologies with heterogeneous networks.

A more comprehensive approach, possibly involving an emulation environment for quantum repeater networks, could facilitate the development and evaluation of effective routing protocols, device APIs, network controllers, and management systems [119], [120].

Traditional network simulation and emulation methods may be inefficient in accurately modeling and running a virtual quantum network. New evaluation paradigms such as digital twins [121], [122] and generative artificial intelligence [123] may be considered to address these challenges.

F. METROLOGY AND HARDWARE CHARACTERIZATION

Given the unprecedented precision required in synchronization and the susceptibility of quantum states to noise, identifying and prioritizing the relevant metrology metrics is crucial for enhancing the design and efficacy of quantum networks. For example, relevant metrics related to fibers in quantum networks are already emerging, such as polarization stability, photon time-of-travel, and noise. These are equivalent to, for example, the detection efficiency, timing jitter, or dark counts of a single-photon detector.

The advancement of measurement tools and protocols will not only impact the standardization of quantum network measurements but also drive innovative solutions to address practical challenges in real-world quantum deployments. This raises the critical question of how best to leverage these metrics and tools to refine entanglement routing schemes, ultimately optimizing quantum network performance and resilience.

G. NETWORK MANAGEMENT AND OPERATION

Quantum networks present unique challenges in operation and management due to their dependence on both quantum and classical communication channels. The classical control plane is essential for coordinating quantum processes, which demand high-speed and precise timing and latency for control signals [124]–[126]. Unlike classical networks, quantum networks face stricter latency and synchronization constraints, driven by the short coherence times of quantum memories and the requirement for indistinguishable photon pairs. This requires specialized hardware, such as FPGAs or ASICs, to meet real-time processing needs and low-latency data formatting. Additionally, quantum networks operate with probabilistic processes where metrics such as throughput and error rate must consider entanglement fidelity and qubit error, rather than simple data packets. These metrics are further complicated by the inability to buffer or retransmit quantum signals arbitrarily.

Network failures introduce additional complexity, with operational failures such as photon loss or desynchronization having high sensitivity due to the fragility of quantum states. Infrastructural problems, such as signal attenuation and hardware malfunctions, require advanced mitigation strategies, including error correction and path redundancy. Calibration and synchronization are especially critical due to quantum protocols' sensitivity to timing, meaning even minor misalignments can disrupt entire operations.

VI. CONCLUSIONS AND PERSPECTIVES

In concluding this survey, it is apparent that entanglement routing is rich with theoretical innovations and practical challenges. Our exploration ties quantum networking concepts to classical terminology to provide a grounding framework that makes these complex concepts more accessible and aligned with familiar networking paradigms. The separation of routing and forwarding, and the modular approach to network design reflected in our proposed taxonomy, could help in understanding and organizing the domain of quantum networks.

The formal definition of the quantum routing problem presented is abstract but captures the major aspects necessary for developing entanglement routing models. This is complemented by our discussion on practical deployment and implementation strategies, which bring the theoretical concepts closer to real-world application. The integration of topology knowledge with routing and forwarding mechanisms, alongside the consideration of routing algorithms and metrics, provides a holistic view of current and possible directions in quantum network design.

From our analysis, while it is clear that there is a rich body of work on routing, the area of forwarding within quantum networks remains less explored. The importance of disaggregating control and data planes, a principle that significantly advanced classical networks, remains attractive in quantum networking, as it simplifies network operations and enhances flexibility. However, the boundaries of such disaggregation will certainly be different from the ones known in classical networks, and a clear distinction is still to emerge.

Our observations suggest that a software-defined and centralized approach to routing, similar to SDN, could streamline operations and optimize performance by including multiple aspects of the network, such as classical communication delays, physical topology, and unused resources, rather than focusing on a single aspect such as logical topologies. However, the real-time requirements of entanglement distribution and decoherence pose significant challenges, suggesting that a hybrid model incorporating both centralized control and localized decision-making at each node could be more effective. Such an approach would mitigate the

rapid decoherence of quantum states and the unpredictability of entanglement generation and swapping, while providing efficient communication and resource management.

The management and monitoring of entanglement-based quantum networks will necessitate more complex strategies that are distinct from classical approaches [127], such as the use of simulation-aided management systems, for example through the digital twins technology [106], given the impossibility of duplicating qubits, the inadequacy of traditional traffic monitoring techniques, and the huge amount of log data generated by quantum processes.

Furthermore, the necessity for the synchronization operations across the network and the application of multiplexing strategies are essential for deploying viable, reliable, and scalable quantum networks [9], [10], [128]. Additionally, robust security protocols cannot be overstated, as quantum networks introduce new vulnerabilities and attack vectors [129], [130]. The development of robust protocols and secure authentication schemes is crucial for protecting these emerging networks against sophisticated attacks.

As we look to the longer-term future, the promise of third-generation quantum repeaters and the exploration of multipartite entanglements are expected to profoundly impact quantum network design and routing protocols. The ongoing research and development in this field promise to lead to significant breakthroughs and innovative solutions that will (re-)shape the evolution of quantum routing.

Finally, it becomes apparent that the traditional concepts and terminologies rooted in classical networking—such as the distinctions between control and data planes, as well as the concepts of forwarding, route metrics and costs may not fully encapsulate the quantum networking characteristics. Quantum networks, particularly the forwarding phase, introduce important shifts in communication paradigms, necessitating a reevaluation of these foundational definitions to better align with all aspects of quantum communications. While in this survey we adhere to classical networking terms for familiarity and reference points for classical networking engineers, this also highlights the transformative impact of quantum technologies on communication, urging a reimagined understanding of network operations that bridges the gap between classical precedents and quantum innovations.

DISCLAIMER

Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NIST, nor does it imply that the products mentioned are necessarily the best available for the purpose.

REFERENCES

[1] Ibm raises the bar with a 50-qubit quantum computer. https://www.technologyreview.com/2017/11/10/147728/ibm-raises-the-bar-with-a-50-qubit-quantum-computer/. Accessed: 2024-01-11.

[2] Google's 72-qubit chip is the largest yet. https://www.newscientist.com/article/2162894-googles-72-qubit-chip-is-the-largest-yet/. Accessed: 2024-01-11.

[3] Davide Castelvecchi. China's quantum satellite clears major hurdle on way to ultrasecure communications. Nature, 15, 2017.

[4] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters. Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. Phys. Rev. Lett., 70:1895–1899, Mar 1993.

[5] M. Żukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert. "event-ready-detectors" bell experiment via entanglement swapping. Phys. Rev. Lett., 71:4287–4290, Dec 1993.

[6] H. de Riedmatten, I. Marcikic, J. A. W. van Houwelingen, W. Tittel, H. Zbinden, and N. Gisin. Long-distance entanglement swapping with photons from separated sources. Phys. Rev. A, 71:050302, May 2005.

[7] Jessica Illiano, Marcello Caleffi, Antonio Manzalini, and Angela Sara Cacciapuoti. Quantum internet protocol stack: A comprehensive survey. Computer Networks, 213:109092, 2022.

[8] Alexander N. Craddock, Anne Lazenby, Gabriel Bello Portmann, Rourke Sekelsky, Mael Flament, and Mehdi Namazi. Automated distribution of high-rate, high-fidelity polarization entangled photons using deployed metropolitan fibers, 2024.

[9] Marzieh Bathaee and Jawad A. Salehi. Entangled-based quantum wavelength-division-multiplexing and multiple-access networks. Entropy, 25(12), 2023.

[10] Emily A Van Milligen, Eliana Jacobson, Ashlesha Patil, Gayane Vardoyan, Don Towsley, and Saikat Guha. Entanglement routing over networks with time multiplexed repeaters, 2024.

[11] Si-Chen Li, Bang-Ying Tang, Han Zhou, Hui-Cun Yu, Bo Liu, Wan-Rong Yu, and Bo Liu. First request first service entanglement routing scheme for quantum networks. Entropy, 24(10), 2022.

[12] Rodney Van Meter, Takahiko Satoh, Thaddeus D Ladd, William J Munro, and Kae Nemoto. Path selection for quantum repeater networks. Networking Science, 3:82–95, 2013.

[13] Carlo Di Franco and D Ballester. Optimal path for a quantum teleportation protocol in entangled networks. Physical Review A, 85(1):010303, 2012.

[14] Rodney Van Meter, Thaddeus D Ladd, William J Munro, and Kae Nemoto. System design for a long-line quantum repeater. IEEE/ACM Transactions On Networking, 17(3):1002–1013, 2008.

[15] Wojciech Kozlowski and Stephanie Wehner. Towards large-scale quantum networks. In Proceedings of the sixth annual ACM international conference on nanoscale computing and communication, pages 1–7, 2019.

[16] Fabrice Dupuy, Claire Goursaud, and Fabrice Guillemin. A survey of quantum entanglement routing protocols—challenges for wide-area networks. Advanced Quantum Technologies, page 2200180, 2023.

[17] Binayak Kar and Pankaj Kumar. Routing protocols for quantum networks: Overview and challenges. arXiv preprint arXiv:2305.00708, 2023.

[18] Pei-Shun Yan, Lan Zhou, Wei Zhong, and Yu-Bo Sheng. Advances in quantum entanglement purification. Science China Physics, Mechanics & Astronomy, 66(5):250301, 2023.

[19] Barbara M. Terhal. Quantum error correction for quantum memories. Rev. Mod. Phys., 87:307–346, Apr 2015.

[20] S.-Y. Lan, A. G. Radnaev, O. A. Collins, D. N. Matsukevich, T. A. B. Kennedy, and A. Kuzmich. A multiplexed quantum memory. Opt. Express, 17(16):13639–13645, Aug 2009.

[21] Tian-Shu Yang, Zong-Quan Zhou, Yi-Lin Hua, Xiao Liu, Zong-Feng Li, Pei-Yun Li, Yu Ma, Chao Liu, Peng-Jun Liang, Xue Li, Yi-Xin Xiao, Jun Hu, Chuan-Feng Li, and Guang-Can Guo. Multiplexed storage and real-time manipulation based on a multiple degree-of-freedom quantum memory. Nature Communications, 9(1):3407, August 2018.

[22] Neil Sinclair, Erhan Saglamyurek, Hassan Mallahzadeh, Joshua A. Slater, Mathew George, Raimund Ricken, Morgan P. Hedges, Daniel Oblak, Christoph Simon, Wolfgang Sohler, and Wolfgang Tittel. Spectral multiplexing for scalable quantum photonics using an atomic frequency comb quantum memory and feed-forward control. Phys. Rev. Lett., 113:053603, Jul 2014.

[23] 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. Nature Communications, 8(1):15359, May 2017.

[24] A. G. Radnaev, Y. O. Dudin, R. Zhao, H. H. Jen, S. D. Jenkins, A. Kuzmich, and T. A. B. Kennedy. A quantum memory with telecom-wavelength conversion. Nature Physics, 6(11):894–899, November 2010.

[25] Xavier Fernandez-Gonzalvo, Giacomo Corrielli, Boris Albrecht, Marcel.li Grimau, Matteo Cristiani, and Hugues de Riedmatten. Quantum frequency

conversion of quantum memory compatible photons to telecommunication wavelengths. Opt. Express, 21(17):19473–19487, Aug 2013.
[26] M S Shahriar, P Kumar, and P R Hemmer. Connecting processing-capable quantum memories over telecommunication links via quantum frequency conversion. Journal of Physics B: Atomic, Molecular and Optical Physics, 45(12):124018, jun 2012.
[27] Amoldeep Singh, Kapal Dev, Harun Siljak, Hem Dutt Joshi, and Maurizio Magarini. Quantum internet—applications, functionalities, enabling technologies, challenges, and research directions. IEEE Communications Surveys & Tutorials, 23(4):2218–2247, 2021.
[28] Sreraman Muralidharan, Linshu Li, Jungsang Kim, Norbert Lütkenhaus, Mikhail D Lukin, and Liang Jiang. Optimal architectures for long distance quantum communication. Scientific reports, 6(1):20463, 2016.
[29] Mihir Pant, Hari Krovi, Don Towsley, Leandros Tassiulas, Liang Jiang, Prithwish Basu, Dirk Englund, and Saikat Guha. Routing entanglement in the quantum internet. npj Quantum Information, 5(1):25, 2019.
[30] Shouqian Shi and Chen Qian. Concurrent entanglement routing for quantum networks: Model and designs. In Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, pages 62–75, 2020.
[31] Kaushik Chakraborty, David Elkouss, Bruno Rijsman, and Stephanie Wehner. Entanglement distribution in a quantum network: A multicommodity flow-based approach. IEEE Transactions on Quantum Engineering, 1:1–21, 2020.
[32] Alena Chang and Guoliang Xue. Order matters: On the impact of swapping order on an entanglement path in a quantum network. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pages 1–6. IEEE, 2022.
[33] Matthias J Bayerbach, Simone E D'Aurelio, Peter van Loock, and Stefanie Barz. Bell-state measurement exceeding 50% success probability with linear optics. Science Advances, 9(32):eadf4080, 2023.
[34] Rodney Van Meter. Quantum networking. John Wiley & Sons, 2014.
[35] Marcello Caleffi. Optimal routing for quantum networks. Ieee Access, 5:22299–22312, 2017.
[36] Mohammad Ghaderibaneh, Caitao Zhan, Himanshu Gupta, and CR Ramakrishnan. Efficient quantum network communication using optimized entanglement swapping trees. IEEE Transactions on Quantum Engineering, 3:1–20, 2022.
[37] Yangming Zhao and Chunming Qiao. Redundant entanglement provisioning and selection for throughput maximization in quantum networks. In IEEE INFOCOM 2021-IEEE Conference on Computer Communications, pages 1–10. IEEE, 2021.
[38] Tu N Nguyen, Kashyab J Ambarani, Linh Le, Ivan Djordjevic, and Zhi-Li Zhang. A multiple-entanglement routing framework for quantum networks. arXiv preprint arXiv:2207.11817, 2022.
[39] Changhao Li, Tianyi Li, Yi-Xiang Liu, and Paola Cappellaro. Effective routing design for remote entanglement generation on quantum networks. npj Quantum Information, 7(1):10, 2021.
[40] Mohamed H Abobeih, Julia Cramer, Michiel A Bakker, Norbert Kalb, Matthew Markham, Daniel J Twitchen, and Tim H Taminiau. One-second coherence for a single electron spin coupled to a multi-qubit nuclear-spin environment. Nature communications, 9(1):2552, 2018.
[41] Christopher P Anderson, Elena O Glen, Cyrus Zeledon, Alexandre Bourassa, Yu Jin, Yizhi Zhu, Christian Vorwerk, Alexander L Crook, Hiroshi Abe, Jawad Ul-Hassan, et al. Five-second coherence of a single spin with single-shot readout in silicon carbide. Science advances, 8(5):eabm5912, 2022.
[42] Jarryd J Pla, Kuan Y Tan, Juan P Dehollain, Wee H Lim, John JL Morton, Floris A Zwanenburg, David N Jamieson, Andrew S Dzurak, and Andrea Morello. High-fidelity readout and control of a nuclear spin qubit in silicon. Nature, 496(7445):334–338, 2013.
[43] David D Awschalom, Ronald Hanson, Jörg Wrachtrup, and Brian B Zhou. Quantum technologies with optically interfaced solid-state spins. Nature Photonics, 12(9):516–527, 2018.
[44] Pengfei Wang, Chun-Yang Luan, Mu Qiao, Mark Um, Junhua Zhang, Ye Wang, Xiao Yuan, Mile Gu, Jingning Zhang, and Kihwan Kim. Single ion qubit with estimated coherence time exceeding one hour. Nature communications, 12(1):233, 2021.
[45] James Schneeloch, Samuel H Knarr, Daniela F Bogorin, Mackenzie L Levangie, Christopher C Tison, Rebecca Frank, Gregory A Howland, Michael L Fanto, and Paul M Alsing. Introduction to the absolute brightness and number statistics in spontaneous parametric down-conversion. Journal of Optics, 21(4):043501, 2019.
[46] Axel Dahlberg, Matthew Skrzypczyk, Tim Coopmans, Leon Wubben, Filip Rozpundefineddek, Matteo Pompili, Arian Stolk, Przemysław Pawełczak, Robert Knegjens, Julio de Oliveira Filho, Ronald Hanson, and Stephanie Wehner. A link layer protocol for quantum networks. In Proceedings of the ACM Special Interest Group on Data Communication, SIGCOMM '19, page 159–173, New York, NY, USA, 2019. Association for Computing Machinery.
[47] Tim van Leent, Matthias Bock, Florian Fertig, Robert Garthoff, Sebastian Eppelt, Yiru Zhou, Pooja Malik, Matthias Seubert, Tobias Bauer, Wenjamin Rosenfeld, et al. Entangling single atoms over 33 km telecom fibre. Nature, 607(7917):69–73, 2022.
[48] Ali Farahbakhsh and Chen Feng. Opportunistic routing in quantum networks. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications, pages 490–499. IEEE, 2022.
[49] 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., 81:5932–5935, Dec 1998.
[50] W. Dür, H.-J. Briegel, J. I. Cirac, and P. Zoller. Quantum repeaters based on entanglement purification. Phys. Rev. A, 59:169–181, Jan 1999.
[51] Kaushik Chakraborty, Filip Rozpedek, Axel Dahlberg, and Stephanie Wehner. Distributed routing in a quantum internet. arXiv preprint arXiv:1907.11630, 2019.
[52] Ling Zhang and Qin Liu. Concurrent multipath quantum entanglement routing based on segment routing in quantum hybrid networks. Quantum Information Processing, 22, 03 2023.
[53] Claudio Cicconetti, Marco Conti, and Andrea Passarella. Request scheduling in quantum networks. IEEE Transactions on Quantum Engineering, 2:2–17, 2021.
[54] Zebo Yang, Ali Ghubaish, Raj Jain, Hassan Shapourian, and Alireza Shabani. Asynchronous entanglement routing for the quantum internet. AVS Quantum Science, 6(1), January 2024.
[55] Yangming Zhao, Gongming Zhao, and Chunming Qiao. E2E fidelity aware routing and purification for throughput maximization in quantum networks. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications, pages 480–489. IEEE, 2022.
[56] Yiming Zeng, Jiarui Zhang, Ji Liu, Zhenhua Liu, and Yuanyuan Yang. Multi-entanglement routing design over quantum networks. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications, pages 510–519. IEEE, 2022.
[57] Jian Li, Mingjun Wang, Kaiping Xue, Ruidong Li, Nenghai Yu, Qibin Sun, and Jun Lu. Fidelity-guaranteed entanglement routing in quantum networks. IEEE Transactions on Communications, 70(10):6748–6763, 2022.
[58] HaoRan Hu, HuaZhi Lun, ZhiFeng Deng, Jie Tang, JiaHao Li, YueXiang Cao, Ya Wang, Ying Liu, Dan Wu, HuiCun Yu, XingYu Wang, JiaHua Wei, and Lei Shi. High-fidelity entanglement routing in quantum networks. Results in Physics, 60:107682, 2024.
[59] Vyacheslav Semenenko, Xuedong Hu, Eden Figueroa, and Vasili Perebeinos. Entanglement generation in a quantum network with finite quantum memory lifetime. AVS Quantum Science, 4(1), 2022.
[60] Laszlo Gyongyosi and Sandor Imre. Decentralized base-graph routing for the quantum internet. Physical Review A, 98(2):022310, 2018.
[61] Laszlo Gyongyosi and Sandor Imre. Adaptive routing for quantum memory failures in the quantum internet. Quantum Information Processing, 18:1–21, 2019.
[62] Mohammad Ghaderibaneh, Himanshu Gupta, C.R. Ramakrishnan, and Ertai Luo. Pre-distribution of entanglements in quantum networks. In 2022 IEEE International Conference on Quantum Computing and Engineering (QCE), pages 426–436, 2022.
[63] Eddie Schoute, Laura Mancinska, Tanvirul Islam, Iordanis Kerenidis, and Stephanie Wehner. Shortcuts to quantum network routing. arXiv preprint arXiv:1610.05238, 2016.
[64] Laszlo Gyongyosi and Sandor Imre. Entanglement-gradient routing for quantum networks. Scientific reports, 7(1):14255, 2017.
[65] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In Proceedings of the thirty-second annual ACM symposium on Theory of computing, pages 163–170, 2000.
[66] Shahrooz Pouryousef, Nitish K. Panigrahy, and Don Towsley. A quantum overlay network for efficient entanglement distribution, 2022.
[67] Laszlo Gyongyosi and Sandor Imre. Opportunistic entanglement distribution for the quantum internet. Scientific Reports, 9(1):2219, 2019.
[68] Takaaki Matsuo, Clément Durand, and Rodney Van Meter. Quantum link bootstrapping using a ruleset-based communication protocol. Physical Review A, 100(5):052320, 2019.

[69] H. P. Bartling, M. H. Abobeih, B. Pingault, M. J. Degen, S. J. H. Loenen, C. E. Bradley, J. Randall, M. Markham, D. J. Twitchen, and T. H. Taminiau. Entanglement of spin-pair qubits with intrinsic dephasing times exceeding a minute. Phys. Rev. X, 12:011048, Mar 2022.

[70] Jian Li, Qidong Jia, Kaiping Xue, David SL Wei, and Nenghai Yu. A connection-oriented entanglement distribution design in quantum networks. IEEE Transactions on Quantum Engineering, 3:1–13, 2022.

[71] H-J Briegel, Wolfgang Dür, Juan I Cirac, and Peter Zoller. Quantum repeaters: the role of imperfect local operations in quantum communication. Physical Review Letters, 81(26):5932, 1998.

[72] Wenhan Dai, Tianyi Peng, and Moe Z Win. Optimal remote entanglement distribution. IEEE Journal on Selected Areas in Communications, 38(3):540–556, 2020.

[73] Lars Kamin, Evgeny Shchukin, Frank Schmidt, and Peter van Loock. Exact rate analysis for quantum repeaters with imperfect memories and entanglement swapping as soon as possible. Physical Review Research, 5(2):023086, 2023.

[74] Stav Haldar, Pratik J Barge, Sumeet Khatri, and Hwang Lee. Fast and reliable entanglement distribution with quantum repeaters: principles for improving protocols using reinforcement learning. Physical Review Applied, 21(2):024041, 2024.

[75] Stav Haldar, Pratik J. Barge, Xiang Cheng, Kai-Chi Chang, Brian T. Kirby, Sumeet Khatri, Chee Wei Wong, and Hwang Lee. Reducing classical communication costs in multiplexed quantum repeaters using hardware-aware quasi-local policies, 2024.

[76] Marcus Cramer, Martin B. Plenio, Steven T. Flammia, Rolando Somma, David Gross, Stephen D. Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu. Efficient quantum state tomography. Nature Communications, 1(1):149, 2010.

[77] Jonas Helsen and Stephanie Wehner. A benchmarking procedure for quantum networks. npj Quantum Information, 9(1):17, 2023.

[78] Steven T. Flammia and Yi-Kai Liu. Direct fidelity estimation from few pauli measurements. Phys. Rev. Lett., 106:230501, Jun 2011.

[79] Xiaoqian Zhang, Maolin Luo, Zhaodi Wen, Qin Feng, Shengshi Pang, Weiqi Luo, and Xiaoqi Zhou. Direct fidelity estimation of quantum states using machine learning. Phys. Rev. Lett., 127:130503, Sep 2021.

[80] Maoli Liu, Zhuohua Li, Kechao Cai, Jonathan Allcock, Shengyu Zhang, and John C.S. Lui. Quantum bgp with online path selection via network benchmarking. In Proceedings of the IEEE Conference on Computer Communications (INFOCOM), Vancouver, Canada, May 2024. (AR: 256/1307=19.6%).

[81] Sam Pallister, Noah Linden, and Ashley Montanaro. Optimal verification of entangled states with local measurements. Phys. Rev. Lett., 120:170502, Apr 2018.

[82] Xiao-Dong Yu, Jiangwei Shang, and Otfried Gühne. Optimal verification of general bipartite pure states. npj Quantum Information, 5(1), December 2019.

[83] Jorge Miguel-Ramiro, Ferran Riera-Sàbat, and Wolfgang Dür. Collective operations can exponentially enhance quantum state verification. Phys. Rev. Lett., 129:190504, Oct 2022.

[84] Ferran Riera-Sàbat, Jorge Miguel-Ramiro, and Wolfgang Dür. Nondestructive verification of entangled states via fidelity witnessing. Phys. Rev. A, 107:022414, Feb 2023.

[85] Wojciech Kozlowski, Axel Dahlberg, and Stephanie Wehner. Designing a quantum network protocol. In Proceedings of the 16th international conference on emerging networking experiments and technologies, pages 1–16, 2020.

[86] Yue Shi, Chenxu Liu, Samuel Stein, Meng Wang, Muqing Zheng, and Ang Li. Design of an entanglement purification protocol selection module, 2024.

[87] Charles H. Bennett, Gilles Brassard, Sandu Popescu, Benjamin Schumacher, John A. Smolin, and William K. Wootters. Purification of noisy entanglement and faithful teleportation via noisy channels. Phys. Rev. Lett., 76:722–725, Jan 1996.

[88] David Deutsch, Artur Ekert, Richard Jozsa, Chiara Macchiavello, Sandu Popescu, and Anna Sanpera. Quantum privacy amplification and the security of quantum cryptography over noisy channels. Phys. Rev. Lett., 77:2818–2821, Sep 1996.

[89] Naomi H. Nickerson, Ying Li, and Simon C. Benjamin. Topological quantum computing with a very noisy network and local error rates approaching one percent. Nature Communications, 4(1):1756, 2013.

[90] Manik Dawar, Ralf Riedinger, Nilesh Vyas, and Paulo Mendes. Quantum internet: Resource estimation for entanglement routing, 2024.

[91] Michelle Victora, Stefan Krastanov, Alexander Sanchez de la Cerda, Steven Willis, and Prineha Narang. Purification and entanglement routing on quantum networks. arXiv preprint arXiv:2011.11644, 2020.

[92] Ashlesha Patil, Michele Pacenti, Bane Vasić, Saikat Guha, and Narayanan Rengaswamy. Entanglement routing using quantum error correction for distillation, 2024.

[93] Shengyu Zhang, Shouqian Shi, Chen Qian, and Kwan L Yeung. Fragmentation-aware entanglement routing for quantum networks. Journal of Lightwave Technology, 39(14):4584–4591, 2021.

[94] Linh Le and Tu N Nguyen. Dqra: Deep quantum routing agent for entanglement routing in quantum networks. IEEE Transactions on Quantum Engineering, 3:1–12, 2022.

[95] Rodney Van Meter, Ryosuke Satoh, Naphan Benchasattabuse, Kentaro Teramoto, Takaaki Matsuo, Michal Hajdusek, Takahiko Satoh, Shota Nagayama, and Shigeya Suzuki. A quantum internet architecture. In 2022 IEEE International Conference on Quantum Computing and Engineering (QCE). IEEE, September 2022.

[96] Maoli Liu, Zhuohua Li, Xuchuang Wang, and John C.S. Lui. LinkSelFiE: Link Selection and Fidelity Estimation in Quantum Networks. In Proceedings of the IEEE Conference on Computer Communications (INFOCOM), Vancouver, Canada, May 2024. (AR: 256/1307=19.6%).

[97] G. Vardoyan, E. van Milligen, S. Guha, S. Wehner, and D. Towsley. On the bipartite entanglement capacity of quantum networks. IEEE Transactions on Quantum Engineering, 5(01):1–14, jan 2024.

[98] Zhuohua Li. lizhuohua/quantum-bgp-online-path-selection: Release of QBGP paper codebase, December 2023.

[99] Wojciech Kozlowski, Stephanie Wehner, Rodney Van Meter, Bruno Rijsman, Angela Sara Cacciapuoti, Marcello Caleffi, and Shota Nagayama. Architectural Principles for a Quantum Internet. RFC 9340, March 2023.

[100] Rodney Van Meter and Takaaki Matsuo. Connection Setup in a Quantum Network. Internet-Draft draft-van-meter-qirg-quantum-connection-setup-01, Internet Engineering Task Force, September 2019. Work in Progress.

[101] Leonardo Bacciottini, Luciano Lenzini, Enzo Mingozzi, and Giuseppe Anastasi. Redip: Ranked entanglement distribution protocol for the quantum internet. IEEE Open Journal of the Communications Society, 5:397–411, 2024.

[102] Alejandro Aguado, Victor López, Juan Pedro Brito, Antonio Pastor, Diego R. López, and Vicente Martin. Enabling quantum key distribution networks via software-defined networking. In 2020 International Conference on Optical Network Design and Modeling (ONDM), pages 1–5, 2020.

[103] Alejandro Aguado, Victor Lopez, Diego Lopez, Momtchil Peev, Andreas Poppe, Antonio Pastor, Jesus Folgueira, and Vicente Martin. The engineering of software-defined quantum key distribution networks. IEEE Communications Magazine, 57(7):20–26, 2019.

[104] Francesco Chiti, Romano Fantacci, Roberto Picchi, and Laura Pierucci. Towards the quantum internet: Satellite control plane architectures and protocol design. Future Internet, 13(8):196, 2021.

[105] Roberto Picchi, Francesco Chiti, Romano Fantacci, and Laura Pierucci. Towards quantum satellite internetworking: A software-defined networking perspective. IEEE Access, 8:210370–210381, 2020.

[106] Stephen F Bush, William A Challener, and Guillaume Mantelet. A perspective on industrial quantum networks. AVS Quantum Science, 3(3), 2021.

[107] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, Jon Zolla, Urs Hölzle, Stephen Stuart, and Amin Vahdat. B4: experience with a globally-deployed software defined wan. In Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM, SIGCOMM '13, page 3–14, New York, NY, USA, 2013. Association for Computing Machinery.

[108] B. Fortz, J. Rexford, and M. Thorup. Traffic engineering with traditional ip routing protocols. Comm. Mag., 40(10):118–124, oct 2002.

[109] Chi-Yao Hong, Subhasree Mandal, Mohammad Al-Fares, Min Zhu, Richard Alimi, Kondapa Naidu B., Chandan Bhagat, Sourabh Jain, Jay Kaimal, Shiyu Liang, Kirill Mendelev, Steve Padgett, Faro Rabe, Saikat Ray, Malveeka Tewari, Matt Tierney, Monika Zahn, Jonathan Zolla, Joon Ong, and Amin Vahdat. B4 and after: managing hierarchy, partitioning, and asymmetry for availability and scale in google's software-defined wan. In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, SIGCOMM '18, page 74–87, New York, NY, USA, 2018. Association for Computing Machinery.

[110] Sajad Khorsandroo, Adrián Gallego Sánchez, Ali Saman Tosun, JM Arco, and Roberto Doriguzzi-Corin. Hybrid sdn evolution: A comprehensive survey of the state-of-the-art. Computer Networks, 192:107981, 2021.

[111] Laszlo Gyongyosi and Sandor Imre. Topology adaption for the quantum internet. Quantum Information Processing, 17:1–12, 2018.
[112] Hyeongrak Choi, Marc G. Davis, Álvaro G. Iñesta, and Dirk R. Englund. Scalable quantum networks: Congestion-free hierarchical entanglement routing with error correction, 2023.
[113] Mohammad Ghaderibaneh, Himanshu Gupta, C.R. Ramakrishnan, and Ertai Luo. Pre-distribution of entanglements in quantum networks. In 2022 IEEE International Conference on Quantum Computing and Engineering (QCE), pages 426–436, 2022.
[114] Vinay Kumar, Claudio Cicconetti, Marco Conti, and Andrea Passarella. Routing in quantum repeater networks with mixed noise figures, 2023.
[115] W. J. Munro, Nicolo' Lo Piparo, Josephine Dias, Michael Hanks, and Kae Nemoto. Designing tomorrow's quantum internet. AVS Quantum Science, 4(2):020503, 06 2022.
[116] Koji Azuma, Kiyoshi Tamaki, and Hoi-Kwong Lo. All-photonic quantum repeaters. Nature Communications, 6(1):6787, 2015.
[117] Naphan Benchasattabuse, Michal Hajdušek, and Rodney Van Meter. Architecture and protocols for all-photonic quantum repeaters, 2024.
[118] Koji Azuma, Stefan Bäuml, Tim Coopmans, David Elkouss, and Boxi Li. Tools for quantum network design. AVS Quantum Science, 3(1):014101, 02 2021.
[119] Mininet. http://mininet.org/. Accessed: 2024-01-11.
[120] Welcome to mininet-optical. https://mininet-optical.org/README.html. Accessed: 2024-01-30.
[121] Qunbi Zhuge, Xiaomin Liu, Yihao Zhang, Meng Cai, Yichen Liu, Qizhi Qiu, Xueying Zhong, Jiaping Wu, Ruoxuan Gao, Lilin Yi, and Weisheng Hu. Building a digital twin for intelligent optical networks – Invited Tutorial. J. Opt. Commun. Netw., 15(8):C242–C262, Aug 2023.
[122] Danshi Wang, Yuchen Song, Yao Zhang, Xiaotian Jiang, Jiawei Dong, Faisal Nadeem Khan, Takeo Sasai, Shanguo Huang, Alan Pak Tao Lau, Massimo Tornatore, and Min Zhang. Digital twin of optical networks: A review of recent advances and future trends. Journal of Lightwave Technology, 42(12):4233–4259, 2024.
[123] Nguyen Van Huynh, Jiacheng Wang, Hongyang Du, Dinh Thai Hoang, Dusit Niyato, Diep N. Nguyen, Dong In Kim, and Khaled B. Letaief. Generative ai for physical layer communications: A survey. IEEE Transactions on Cognitive Communications and Networking, 10(3):706–728, 2024.
[124] Wojciech Kozlowski, Fernando Kuipers, and Stephanie Wehner. A p4 data plane for the quantum internet. In Proceedings of the 3rd P4 Workshop in Europe, CoNEXT '20. ACM, December 2020.
[125] Venkat R Dasari and Travis S Humble. Openflow arbitrated programmable network channels for managing quantum metadata. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 16(1):67–77, October 2016.
[126] Jessica Illiano, Angela Sara Cacciapuoti, Antonio Manzalini, and Marcello Caleffi. The impact of the quantum data plane overhead on the throughput. In Proceedings of the Eight Annual ACM International Conference on Nanoscale Computing and Communication, pages 1–6, 2021.
[127] Iván García-Cobo. Quantum network intelligent management system. Optics, 3(4):430–437, 2022.
[128] Christopher Spiess, Sebastian Töpfer, Sakshi Sharma, Andrej Kržič, Meritxell Cabrejo-Ponce, Uday Chandrashekara, Nico Lennart Döll, Daniel Rieländer, and Fabian Steinlechner. Clock synchronization with correlated photons. Phys. Rev. Appl., 19:054082, May 2023.
[129] Takahiko Satoh, Shota Nagayama, Shigeya Suzuki, Takaaki Matsuo, Michal Hajdusek, and Rodney Van Meter. Attacking the quantum internet. IEEE Transactions on Quantum Engineering, 2:1–17, 2021.
[130] Hongyi Zhou, Kefan Lv, Longbo Huang, and Xiongfeng Ma. Quantum network: Security assessment and key management. IEEE/ACM Transactions on Networking, 30(3):1328–1339, June 2022.
[131] M.A. Nielsen and I.L. Chuang. Quantum Computation and Quantum Information: 10th Anniversary Edition. Cambridge University Press, 2010.
[132] Howard Barnum, Jonathan Barrett, Matthew Leifer, and Alexander Wilce. Generalized no-broadcasting theorem. Phys. Rev. Lett., 99:240501, Dec 2007.
[133] Suzanne B van Dam, Peter C Humphreys, Filip Rozpedek, Stephanie Wehner, and Ronald Hanson. Multiplexed entanglement generation over quantum networks using multi-qubit nodes. Quantum Science and Technology, 2(3):034002, 2017.
[134] Stefanie Barz, Gunther Cronenberg, Anton Zeilinger, and Philip Walther. Heralded generation of entangled photon pairs. Nature Photonics, 4(8):553–556, June 2010.
[135] Khabat Heshami, Duncan G England, Peter C Humphreys, Philip J Bustard, Victor M Acosta, Joshua Nunn, and Benjamin J Sussman. Quantum memories: emerging applications and recent advances. Journal of modern optics, 63(20):2005–2028, 2016.
[136] Charles H. Bennett, David P. DiVincenzo, John A. Smolin, and William K. Wootters. Mixed-state entanglement and quantum error correction. Phys. Rev. A, 54:3824–3851, Nov 1996.
[137] M. Zwerger, H. J. Briegel, and W. Dür. Robustness of hashing protocols for entanglement purification. Phys. Rev. A, 90:012314, Jul 2014.
[138] M. Zwerger, W. Dür, and H. J. Briegel. Measurement-based quantum repeaters. Phys. Rev. A, 85:062326, Jun 2012.
[139] F. Riera-Sàbat, P. Sekatski, A. Pirker, and W. Dür. Entanglement-assisted entanglement purification. Phys. Rev. Lett., 127:040502, Jul 2021.
[140] F. Riera-Sàbat, P. Sekatski, A. Pirker, and W. Dür. Entanglement purification by counting and locating errors with entangling measurements. Phys. Rev. A, 104:012419, Jul 2021.
[141] Sreraman Muralidharan, Jungsang Kim, Norbert Lütkenhaus, Mikhail D Lukin, and Liang Jiang. Ultrafast and fault-tolerant quantum communication across long distances. Physical review letters, 112(25):250501, 2014.
[142] Stefan Bäuml and Koji Azuma. Fundamental limitation on quantum broadcast networks. Quantum Science and Technology, 2(2):024004, 2017.
[143] Siddhartha Santra, Liang Jiang, and Vladimir S Malinovsky. Quantum repeater architecture with hierarchically optimized memory buffer times. Quantum Science and Technology, 4(2):025010, 2019.
[144] Angela Sara Cacciapuoti, Marcello Caleffi, Francesco Tafuri, Francesco Saverio Cataliotti, Stefano Gherardini, and Giuseppe Bianchi. Quantum internet: Networking challenges in distributed quantum computing. IEEE Network, 34(1):137–143, 2019.
[145] Luciano Aparicio and Rodney Van Meter. Multiplexing schemes for quantum repeater networks. In Ronald E. Meyers, Yanhua Shih, and Keith S. Deacon, editors, Quantum Communications and Quantum Imaging IX, volume 8163, page 816308. International Society for Optics and Photonics, SPIE, 2011.
[146] Charles H. Bennett, Gilles Brassard, and N. David Mermin. Quantum cryptography without bell's theorem. Phys. Rev. Lett., 68:557–559, Feb 1992.
[147] Tim Coopmans, Robert Knegjens, Axel Dahlberg, David Maier, Loek Nijsten, Julio de Oliveira Filho, Martijn Papendrecht, Julian Rabbie, Filip Rozpedek, Matthew Skrzypczyk, et al. NetSquid, a network simulator for quantum information using discrete events. Communications Physics, 4(1):164, 2021.
[148] Xiaoliang Wu, Alexander Kolar, Joaquin Chung, Dong Jin, Tian Zhong, Rajkumar Kettimuthu, and Martin Suchara. SeQUeNCe: a customizable discrete-event simulator of quantum networks. Quantum Science and Technology, 6(4):045027, 2021.
[149] Ryosuke Satoh, Michal Hajdušek, Naphan Benchasattabuse, Shota Nagayama, Kentaro Teramoto, Takaaki Matsuo, Sara Ayman Metwalli, Poramet Pathumsoot, Takahiko Satoh, Shigeya Suzuki, et al. QuISP: a quantum internet simulation package. In 2022 IEEE International Conference on Quantum Computing and Engineering (QCE), pages 353–364. IEEE, 2022.
[150] Axel Dahlberg and Stephanie Wehner. SimulaQron—a simulator for developing quantum internet software. Quantum Science and Technology, 4(1):015001, 2018.
[151] Stephen DiAdamo, Janis Nötzel, Benjamin Zanger, and Mehmet Mert Beşe. QuNetSim: A software framework for quantum networks. IEEE Transactions on Quantum Engineering, 2:1–12, 2021.
[152] Julius Wallnöfer, Frederik Hahn, Fabian Wiesner, Nathan Walk, and Jens Eisert. Faithfully simulating near-term quantum repeaters. PRX Quantum, 5(1), March 2024.
[153] Lutong Chen, Kaiping Xue, Jian Li, Nenghai Yu, Ruidong Li, Qibin Sun, and Jun Lu. Simqn: A network-layer simulator for the quantum network investigation. IEEE Network, 37(5):182–189, 2023.

APPENDIX. QUANTUM COMMUNICATION BACKGROUND

This section introduces the fundamental quantum physics operations and properties essential in quantum communication. Instead of detailing each principle in isolation, as is common in the literature, we will illustrate these principles through the step-by-step design of a hypothetical quantum

network, intended to transfer quantum states to two end-nodes over a long distance. This allows us to present these operations as they functionally coexist in a quantum network. Some technological specifics a nd h ardware components are intentionally understated to maintain an abstracted perspective on quantum networks and entanglement routing.

A. FROM QUANTUM BITS TO QUANTUM NETWORKS

a: Qubits and Quantum States

Quantum computing operates on quantum bits (qubits) encoded in the quantum state of particles such as photons, electrons, and atoms. A quantum state has the particularity of being in a superposition; i.e., 0 and 1 at the same time [27]. Various technologies can be used to represent a qubit such as encoding a quantum state in photon polarization or electron spin. The principles of quantum mechanics hold regardless of the underlying qubit technology.

Scaling the computational power of quantum computers and expanding the potential of quantum computing to more applications requires the transfer of qubits between distant nodes without destroying their superposition states. For example, Alice begins by encoding her message into the quantum states of photons, which will be sent to Bob. This initial step introduces the challenge of maintaining the integrity of the qubits during transmission. The no-cloning principle [131] prohibits the exact replication of quantum states, ensuring that quantum information remains fundamentally secure but also challenging to manipulate. Additionally, any measurement of a quantum system inevitably alters its state. Consequently, qubits cannot be measured and then replicated to repeat a quantum signal or retry a transmission. These principles do not hold in with classical signals, which is why classical networks can use amplifiers, f eed-forward e rror correction, and retransmission to reliably send units of data across the network.

b: Elementary Entanglement

Bipartite entanglement is a special connection between two quantum particles, like photons, where their properties are linked together regardless of the distance between them. This means if one of the entangled qubits is observed or measured, it instantly determines the state of the other, breaking the entangled state, no matter how far apart they are. It is as if the two qubits are acting as one unit, even when they are separated by an arbitrary distance. Unlike a mix of separate states, this entangled state cannot be expressed by simpler, individual states for each particle [7]. Since a qubit cannot be replicated, a third qubit cannot be entangled with either of the two entangled qubits. This is referred to as the no-broadcasting theorem [132].

Alice and Bob are considered *entangled nodes* if they share one or more pairs of entangled qubits. These pairs of entangled qubits are known as *Bell pairs* or Einstein-Podolsky-Rosen (EPR) pairs (terms that we will use interchangeably throughout this paper) symbolizing the four orthogonal states possible for two maximally entangled qubits located with Alice and Bob, respectively. As a result, when Alice independently measures her qubit from a given Bell pair, she receives a random result, with the probabilities of zero and one outcomes being equally likely. Bob experiences the same phenomenon with his measurement. If Alice and Bob measure their respective qubits in the same basis, the results will be perfectly correlated. This is regardless of the distance between Alice and Bob and without any further interaction between the two parties.

Some prominent schemes to generate Bell pairs include Spontaneous Parametric Down Conversion (SPDC), laser beam excitation of a single atom, and simultaneous excitation of two atoms by a laser beam [27]. Most entangled photon pair sources (EPPS) based on these schemes are non-deterministic as to when Bell pairs are generated. Based on these schemes, the so-called elementary entanglement generation (EEG) protocols are designed to attempt to generate Bell pairs and distribute each entangled qubit to one end node of a direct link [133]. Elementary entanglement is defined as entanglement between two neighboring nodes (i.e., directly connected through an optical channel).

c: Entanglement Generation and Heralding

One way to overcome the limitations of nondeterministic sources is by heralding. Heralding refers to the process of mutually acknowledging the confirmed presence or absence of an entangled pair between the nodes attempting an entanglement creation. For example, one can implement a heralded entangled photon source by generating bipartite entangled states conditioned on the detection of additional auxiliary photons [134]. If these auxiliary photons are found to meet the requisite conditional measurement patterns, they herald the successful creation of the entangled pair. Another example scheme would be to integrate heralded quantum memories at each end node, which can confirm the successful storage of a photon upon its arrival. As such, a Heralded Entanglement Generation (HEG) protocol may require a two-way classical signaling message to each qubit recipient node [28] to confirm the presence of an entangled state between the two end-nodes. It is important to note that this additional classical communication overhead of HEG protocols has critical implications for higher-layer network protocols such as routing.

There are several architectures one can consider for HEG between two neighboring nodes, as illustrated in Figure 14. We summarize these architectures as node-source, midpoint-source, and meet-in-the-middle.

Let the physical distance between Alice and Bob be L. In a node-source architecture, the sender (Alice) has a local EPPS and a quantum memory that receives one entangled qubit from the EPPS, and the receiver (Bob) has a local quantum memory that receives the other entangled qubit from Alice's EPPS after it crosses distance L.

To make an HEG protocol from a node-source architecture, only one-way classical communication is required from Alice to Bob over distance L after each attempt. In a midpoint-

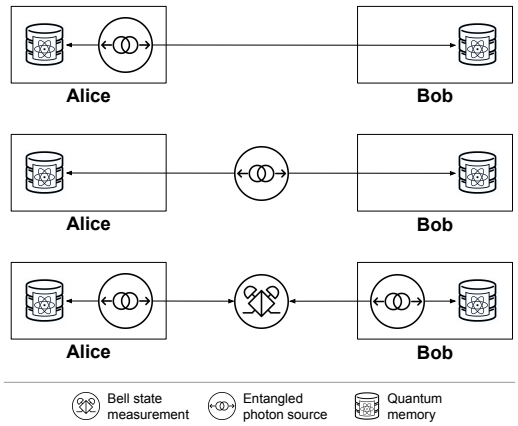

FIGURE 14. Varying physical architectures for entanglement generation between quantum memories on Alice and Bob: node-source (top), midpoint-source (middle), and meet-in-the-middle (bottom).

source architecture, Alice and Bob each receive an entangled qubit into their local quantum memory from an EPPS that lies somewhere on the channel between them. To make an HEG protocol from a midpoint-source architecture, the EPPS must send a heralding signal to both Alice and Bob after each attempt.

In a meet-in-the-middle architecture, Alice and Bob both have a local EPPS and quantum memory, from which they each emit one entangled qubit from their EPPS to their local memory and send the other entangled qubit to a Bell-state measurement (BSM) station that lies somewhere on the channel between them. If the qubits arriving at the BSM are indistinguishable (i.e., have identical or perfectly correlated properties such as their polarization states, phase, wavelength, and timing), a successful BSM will result in entanglement between the quantum memories at Alice and Bob. Depending on the capabilities of the BSM station, the emitted qubits from each pair generation attempt at Alice and Bob may need to arrive at the BSM station simultaneously, requiring a time-synchronized system. To make an HEG protocol from a meet-in-the-middle architecture, the BSM station must send a heralding signal to both Alice and Bob after each attempt. In addition to the heralding signal itself, these classical messages may also include data indicating which Bell state was prepared in the process, which allows the end nodes to execute any relevant state transformations, if necessary.

While the HEG protocols described herein assume that the classical communication of heralding signals is reliable, the HEG protocol implementation must be robust in the face of lost or garbled messages. It is straightforward to see that each of these architectures and HEG protocols has different tradeoffs with classical communication overhead, hardware and infrastructure requirements, and time synchronization.

d: Quantum Memory

Entanglement generation schemes frequently use photons (flying qubits) to encode quantum information and deliver the generated entangled states to end nodes over optical fiber, and atoms, or matter qubits, to store the entangled state at each node in a quantum memory. Photons are a promising medium for encoding and transmitting quantum information due to their degrees of freedom and their fast traversal speed over fiber and free-space optical links.

Capturing and storing a photon in a quantum memory can be achieved with various technologies [135], either with solid-state platforms or free-space platforms. Some platforms support on-demand photon emission from the memory, and others may only emit at set time intervals. Memories can be characterized by key parameters such as their storage time representing the time interval beyond which the stored quantum state is irreversibly degraded and can no longer be used. This results from entanglement decoherence, where the entangled pair of particles degrades over time due to interactions with their surrounding environment. Other memory parameters can be storage efficiency, retrieval efficiency, and an acceptable wavelength range.

e: Entanglement Purification

Decoherence, fiber loss, and noisy quantum operations prevent entangled states from achieving or maintaining optimal-quality entanglement with respect to a desired state, thus compromising the utility of entanglement links in the network. The probability that generated entanglements are in a certain desired state is quantified by a value known as the fidelity.

Entanglement purification (or distillation) provides an efficient means of generating high-fidelity entangled states from multiple copies of noisy entangled pairs. Unlike error correction, these protocols assume that the target state is known. Purification is particularly critical in long-distance communication with quantum repeaters, where it poses a bottleneck due to substantial overhead that affects transmission rates and error resilience.

Entanglement purification protocols use LOCC to reduce noise in an ensemble of entangled pairs, enhancing the entanglement of a subset while discarding the rest. These protocols work by using some of the entangled pairs to collect information about others without consuming them. This information enables the protocol to identify, filter, and sometimes correct errors within the remaining pairs, effectively "purifying" the ensemble.

There are different types of entanglement purification protocols, varying in the number of pairs involved, LOCC, and whether they are deterministic or probabilistic. Among these, *recurrence protocols* (two-way classical communication [87]–[89]) are widely used in practice. These protocols operate on two pairs at a time, where one pair is measured to gain information about the other. With each successful iteration, the remaining pairs become less noisy, converging asymptotically toward the desired entangled state. However, since one pair is discarded in each step, the protocol yield (the fraction of perfectly entangled pairs generated from the

noisy ensemble) approaches zero when approaching unit fidelity.

The recurrence protocol or Heralded Entanglement Purification (HEP) is illustrated in Figure 15. The HEP process can sometimes fail and requires bidirectional classical communication to notify both end nodes about the outcome of the purification effort [27].

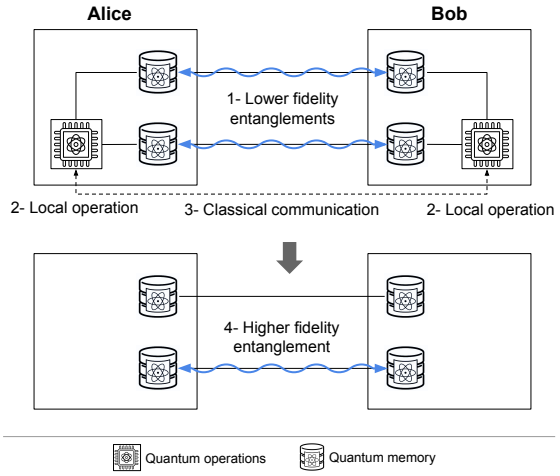

FIGURE 15. Entanglement purification. With local physical operations and communication through classical channels, an entangled pair with high fidelity can be distilled from two pairs with lower fidelity.

Alternatively, hashing and breeding protocols [136] are deterministic and require one-way classical communication. These protocols treat the entire ensemble at once. They operate on a large number of identical noisy pairs and measure a subset to reveal information about the rest and progressively exclude incorrect states. If the initial fidelity is sufficiently high [136], these protocols achieve a maximally entangled state with non-zero yield. Although hashing and breeding protocols are theoretically efficient, they are infeasible with noisy gate-based implementations because the number of required operations grows infinitely in the asymptotic limit, causing noise to accumulate and corrupt the ensemble information [137]. This issue can be addressed with measurement-based implementations [138], where entangled resource states are prepared locally and coupled by Bell measurements to the particles in the ensemble.

This approach inspired a newer class of purification protocols using high-dimensional auxiliary states (qudits) to purify qubit entanglement [139], [140]. These protocols exploit the extra levels of qudits, enabling more efficient purification processes. Unlike hashing protocols, they utilize counter gates to transfer specific error information to the auxiliary qudit system, allowing not only parity checks but also precise identification of error locations and types. These protocols can operate in deterministic mode, always identifying the error configuration, or in probabilistic mode, where the protocol may terminate if the error count exceeds a certain threshold.

f: Quantum Error Correction

In addition to purification, Quantum Error Correction (QEC) is essential to manage errors in quantum states during transmission [27], [28]. QEC techniques encode a logical qubit into a block of several physical qubits, which helps to preserve the integrity of the quantum state despite potential errors that may arise from decoherence or other quantum operations. By encoding and then decoding the information, QEC allows for the correction of errors without needing to directly observe the quantum state, thereby not violating the no-cloning principle. However, the effectiveness of QEC is restricted by the no-cloning theorem to less than a 50% threshold [141]. This error threshold represents the maximum rate of errors that can be corrected using a QEC protocol. Beyond this threshold, the accumulation of errors outpaces the capability of the error correction process, failing to recover the original quantum information.

g: Entanglement Swapping

Given the inherent lossy nature of optical channels, the probability of successfully distributing an entangled pair decreases exponentially as the physical distance of the channel increases. There are several approaches for mitigating photon loss to extend the distance over which high-quality entanglements are distributed. For example, using high-rate sources or multiplexing techniques (increase the number of attempts per unit time), high-efficiency photon detectors (more accurate measurements), low-loss optical fiber, or free-space channels can all help improve the HEG success probability. However, these approaches can only partially mitigate the exponential nature of photon loss and rapidly become insufficient [142]. To distribute entanglement across even longer distances, quantum repeaters are used to chain a sequence of elementary entanglements to create a longer one using swapping [143].

The swapping process is illustrated in Figure 16. One repeater is placed between Alice and Bob to split the end-to-end distance into two smaller distances. Two elementary entanglements are generated; one between Alice and the repeater and one between the repeater and Bob. The repeater then measures its two entangled local qubits and sends the measurement outcome to Bob. Based on the received measurement, Bob applies a correction to its local qubit, creating an E2E entanglement between Alice and Bob. It is important to note that the measurement process at the repeater destroys the two initial entanglements, thus requiring the generation of new elementary entanglements whenever swapping is used. The success of the swapping operation is also probabilistic.

To support longer distances between Alice and Bob, more repeaters can be placed between the two end nodes. Elementary entanglements are created from Alice to the first repeater, between repeaters, and from the last repeater to Bob. These elementary entanglements are consumed as the swapping operations are executed at each repeater, creating the E2E entanglement.

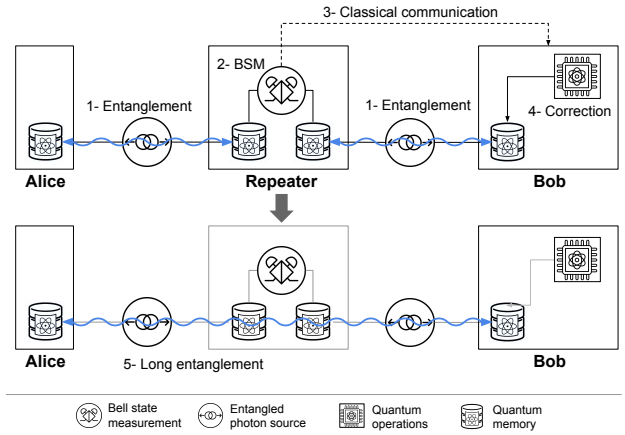

FIGURE 16. Entanglement swapping where two distant nodes share an entanglement with the assistance of an intermediate repeater.

If all swapping operations are successful, the E2E entanglement can be obtained by applying the swapping corrections in any order, since they are performed within the coherence time of the entangled pairs [7], [144]. This allows for using different swapping orders to create the E2E entanglement. However, it is worth noting that the fidelity of the resulting entanglement depends on the time elapsed before applying corrections. That is, the resulting entanglement fidelity is also affected by the time taken by the classical measurements to reach the end node.

h: Quantum Repeaters

The main function of a repeater is to implement the swapping process described above. Entanglement distribution over quantum repeaters is subject to two main types of errors [28]: loss errors arising from the attenuation of fibers and operational errors resulting from inaccuracies in manipulating and measuring quantum states.

By repeating the heralded process until the neighboring repeaters confirm the successful creation of an entanglement, the HEG protocol serves as a mechanism for natively mitigating loss errors. During this period, the entangled qubits are maintained by the repeaters until a clear indication of success or failure is communicated.

On the other hand, QEC offers a more efficient approach, though it demands more resources. At each repeater, QEC can be implemented to restore the original logical qubit. In this scenario, the quantum and classical signals flow unidirectionally. That is, the end nodes do not store any quantum states once the logical qubit is restored and transmitted.

To handle operational errors, the choice falls between QEC [131] and Heralded Entanglement Purification (HEP) [18]. HEP utilizes a set of low-fidelity Bell pairs to probabilistically generate a reduced number of high-fidelity pairs, necessitating bidirectional classical communication for the verification of success. In contrast, QEC addresses these errors through a unidirectional classical communication system but requires quantum gates with higher fidelity.

The evolution of quantum repeaters can be delineated into three distinct generations based on their error mitigation strategies [28].

The first generation integrates HEG for loss error management and HEP for the correction of operation errors. The communication is based on the production of high-fidelity entangled pairs between neighboring repeaters, and employing HEP at each stage of entanglement swapping to counteract fidelity degradation due to quantum operations. It is important to note that executing an HEP protocol on multi-hop entanglement requires classical signaling between non-adjacent nodes, adding a potentially costly latency overhead.

The second generation employs HEG for the identification of loss errors and QEC for the mitigation of operation errors. In this architecture, entangled qubits distributed through HEG enable the formation of a Bell pair for each qubit within an encoded physical block. QEC is applied to restore the logical qubit from the block of entangled physical qubits, thus obviating the necessity for bidirectional classical communication by substituting HEP with QEC.

The third generation exclusively relies on QEC to address both loss and operational errors, directly encoding the logical qubit within a series of physical qubits dispatched through channels as photons. Provided the errors are minimal, the incoming physical qubits can be employed to reconstruct and transmit the encoding block to the subsequent station. As a result, the signaling is unidirectional, and the communication rate can potentially be maximized.

The first-generation repeaters are currently in the development phase. Although not commercially available yet, a high-level view of the architecture of a first-generation quantum repeater is illustrated in Figure 17. Such a repeater consists of quantum measurement and operation components, such as BSM and quantum gates, and quantum memory modules. Depending on the specific features and technology, the architecture may include an entangled photon source and a quantum error correction circuit. In addition to optical ports, repeaters also include classical communication ports to coordinate actions and share measurement outcomes. These components and interactions would be controlled with a pre-programmed hardware component such as an FPGA to meet the real-time requirements of low-level quantum processes. At a higher level, software-based control can be used for relatively slower operations such as network instruction processing and table lookups.

Note that quantum repeaters are limited to extending only entanglement distribution distances over a linear path. Path computation and routing are implemented in quantum routers which, as their classical counterpart, are quantum repeaters with more than two interfaces; i.e., they can service direct quantum connections between more than two neighbors.

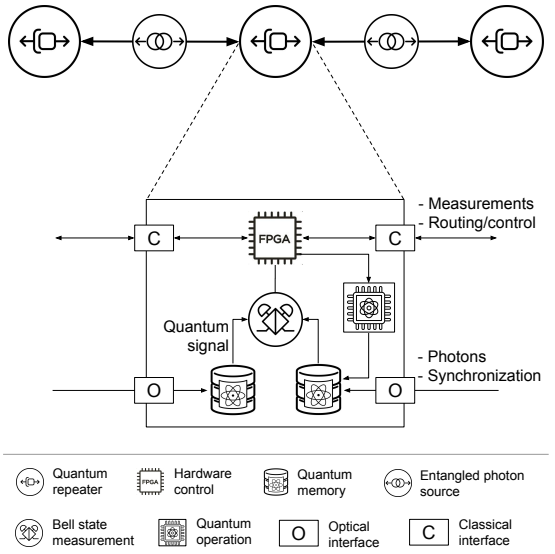

FIGURE 17. Simplified architecture of a first-generation quantum repeater assuming midpoint-source entanglement architecture.

i: Multiplexing

First-generation quantum networks utilize circuit switching, which establishes a dedicated path between the two communicating end nodes before transmitting quantum information. To serve E2E entanglements across the network, a policy is needed to allocate the qubits available at repeaters to different paths. Various approaches have been proposed for this purpose. A straightforward one is to reserve the entire quantum channel along the path for the duration of the communication session. Although easy to implement, this pure circuit switching method has limitations in terms of network utilization and scalability. Alternatively, buffer space multiplexing divides the limited memory or buffer space at each repeater among the different paths that pass through it [145]. This method allows multiple communications to be supported concurrently, thereby improving utilization over pure circuit switching. Statistical multiplexing dynamically shares the available buffer space between different paths based on demand, rather than statically dividing it [145]. Time division multiplexing allocates the available time slots at each repeater node among different paths.

Multiplexing techniques are also used to manage entanglement creation over a single link, for example through time-division [10] or wavelength-division [9]. Overall, the choice of multiplexing approach involves balancing performance, complexity, and implementation feasibility trade-offs for the specific quantum network design.

j: Teleportation

Distributing entanglements between Alice and Bob does not constitute an exchange of quantum data by itself. Instead, it serves as a connection to transmit data qubits through the teleportation process [34]. Since qubits cannot be copied, teleportation is used for the transmission of a quantum state without physically transferring the quantum particle that encodes the qubit. The teleportation process is depicted in Figure 18. The procedure for teleporting an arbitrary qubit X from Alice to Bob involves generating an E2E entanglement between Alice and Bob, where qubit A is located at Alice and qubit B at Bob. Alice then performs a BSM of X and A, and the two-bit result of this measurement is communicated to Bob through a classical channel. Upon receiving this information, Bob applies specific operations to qubit B based on the measurement outcome, resulting in the retrieval of qubit X [7].

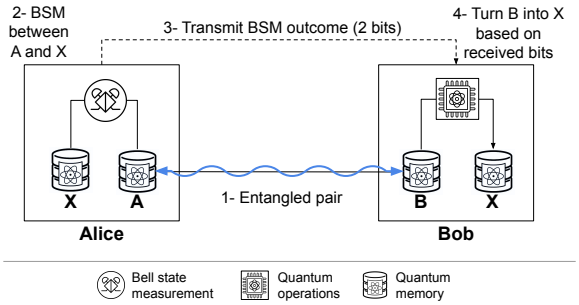

FIGURE 18. Quantum teleportation: an unknown qubit is transferred from Alice to Bob by consuming an entangled pair A-B shared between Alice and Bob.

While the teleportation operation is described to illustrate an end-to-end exchange of quantum information, it is worth noting that there are applications, such as entanglement-based quantum key distribution [146], that consume E2E entanglements directly without teleportation.

B. QUANTUM NETWORK SIMULATORS

Several quantum network simulators are currently available as open-source packages or commercial products, each supporting different levels of abstraction and catering to various aspects of quantum networking. Some notable examples include:

- NetSquid [147]: a modular and scalable quantum network simulator that models time-dependent processes such as qubit decoherence. It supports detailed hardware modeling and multiple quantum state representations, for efficient evaluation of quantum systems and network designs.
- SeQUeNCe [148]: a full-stack discrete event simulator designed to study quantum links, systems, and networks. It uses Bra-Ket notation and density matrices, with a noise model that accounts for qubit decoherence, photon loss, and gate imperfections.
- QuISP [149]: a simulation package for the quantum internet that implements the RuleSet protocol [95] based on the OMNET++ simulator. The qubit model uses error tracking instead of full state representations.
- SimulaQron [150]: a simulator for the application layer of quantum networks. Network layer software such

as routing protocols can also be implemented and simulated.

- QuNetSim [151]: a high-level framework that simulates the network layer and above. It does not implement lower-level quantum device operations and thus is suitable for quick testing of quantum applications and routing protocols.
- ReQuSim [152]: a lightweight Python-based quantum network simulator that simulates the entanglement distribution across quantum repeaters including loss and a wide range of imperfections such as memories with time-dependent noise.
- SimQN [153]: A modular, discrete-event-driven simulator specifically designed for evaluations of the network layer. SimQN aims to balance ease of configuration, scalability, and accuracy for simulating diverse quantum networking scenarios, including entanglement distribution, purification, and routing.

Similar to classical network simulators, quantum network simulators provide significant benefits in modeling and testing, protocol development and optimization, and scalability and security assessments. They play a crucial role in advancing quantum communication research, enabling innovation, and preparing for the future quantum Internet.