FENDI: Toward High-Fidelity Entanglement Distribution in the Quantum Internet

Huayue Gu, *Graduate Student Member, IEEE*, Zhouyu Li, *Graduate Student Member, IEEE*, Ruozhou Yu, *Senior Member, IEEE*, Xiaojian Wang, *Graduate Student Member, IEEE*, Fangtong Zhou, *Graduate Student Member, IEEE*, Jianqing Liu, *Member, IEEE*, and Guoliang Xue, *Fellow, IEEE*

***Abstract*— A quantum network distributes quantum entanglements between remote nodes, and is key to many applications in secure communication, quantum sensing and distributed quantum computing. This paper explores the fundamental trade-off between the throughput and the quality of entanglement distribution in a multi-hop quantum repeater network. Compared to existing work which aims to heuristically maximize the entanglement distribution rate (EDR) and/or entanglement fidelity, our goal is to characterize the maximum achievable worst-case fidelity, while satisfying a bound on the maximum achievable expected EDR between an arbitrary pair of quantum nodes. This characterization will provide fundamental bounds on the achievable performance region of a quantum network, which can assist with the design of quantum network topology, protocols and applications. However, the task is highly non-trivial and is NP-hard as we shall prove. Our main contribution is a *fully polynomial-time approximation scheme* to approximate the achievable worst-case fidelity subject to a strict expected EDR bound, combining an optimal fidelity-agnostic EDR-maximizing formulation and a worst-case isotropic noise model. The EDR and fidelity guarantees can be implemented by a post-selection-and-storage protocol with quantum memories. By developing a discrete-time quantum network simulator, we conduct simulations to show the characterized performance region (the approximate Pareto frontier) of a network, and demonstrate that the designed protocol can achieve the performance region while existing protocols exhibit a substantial gap.**

***Index Terms*— Quantum network, entanglement routing, entanglement fidelity, network optimization, approximation algorithm.**

Manuscript received 24 October 2023; revised 16 May 2024; accepted 15 August 2024; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor D. Elkouss. The work of Huayue Gu, Zhouyu Li, Ruozhou Yu, Xiaojian Wang, and Fangtong Zhou was supported in part by NSF under Grant 2045539 and Grant 2350152. The work of Jianqing Liu was supported in part by NSF under Grant 2304118 and Grant 2326746. The work of Guoliang Xue was supported in part by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) Program, under Contract ERKJ432; in part by the Performance Integrated Quantum Scalable Internet (PiQSci) Quantum Networking Project; and in part by NSF under Grant 2007083. *(Huayue Gu and Zhouyu Li contributed equally to this work.) (Corresponding author: Ruozhou Yu.)*

Huayue Gu, Zhouyu Li, Ruozhou Yu, Xiaojian Wang, Fangtong Zhou, and Jianqing Liu are with the Department of Computer Science, North Carolina State University, Raleigh, NC 27606 USA (e-mail: hgu5@ncsu.edu; zli85@ncsu.edu; ryu5@ncsu.edu; xwang244@ncsu.edu; fzhou@ncsu.edu; jliu96@ncsu.edu).

Guoliang Xue is with the School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281 USA (e-mail: xue@asu.edu).

Digital Object Identifier 10.1109/TNET.2024.3450271

I. INTRODUCTION

A QUANTUM network enables efficient quantum communication [29]. The ability to transmit information between remote nodes is key to many astonishing quantum applications, such as quantum secure communication [5], distributed quantum computing [15], and quantum sensor network [61].

Though current systems are built in ideal conditions and on a small-scale [17], [23], [42], [46], [58], research has explored how such small-scaled networks could potentially be extended to a fully-fledged, global-scale quantum internet to distribute entangled quantum states between remote nodes across long distances. Future applications would require a steady stream of high-quality entanglements between arbitrary remote ends.

This paper considers a first-generation quantum network built with quantum repeaters [39], which performs entanglement distribution via *entanglement generation* and *entanglement swapping*. If a quantum link connects a pair of repeaters, one repeater can generate an entangled photon pair and send one of the pair to the other repeater directly. Entanglements generated over multiple links can further be swapped at joint intermediate nodes to entangle qubits at indirectly connected nodes. This way, each end-to-end entanglement is generated along an *entanglement path* in a quantum network.

As entanglements are a critical resource, attention has been drawn to the efficient distribution protocol design to balance the quantity (*aka* entanglement distribution rate or EDR) and quality (*aka* fidelity) of entanglement distribution. A quantum network has unique characteristics imposed by the underlying physics or technology deficits. First, entanglement distribution efficiency is fundamentally limited by transmission loss of entangled photons and failures in swapping. To mitigate these, existing works have studied *efficient entanglement routing* to find paths with maximum success probability [18], [48], [62]. Second, uncontrollable noise and operation errors can degrade the quality of distributed entanglements. Low fidelity results in low communication efficiency due to excessive error correction needed, even when the EDR is high. Thus, it is essential to consider both EDR and fidelity to support various applications.

This paper explores the tradeoff between the achievable EDR and the fidelity of a general quantum network. We start with characterizing the end-to-end fidelity of entanglements distributed over an entanglement path. We then propose a novel decomposition theorem based on a new *primitive entanglement flow* abstraction that generalizes an entanglement path by considering the order of swapping along a path, which we show to be an exact abstraction for characterizing

both the maximum achievable EDR and the end-to-end fidelity at the same time. Next, we formulate the problem of computing the maximum achievable worst-case fidelity while trying to satisfy a lower bound on the achievable expected EDR. This bi-criteria formulation can be used to optimize for many applications that desire a steady entanglement rate and can benefit from improved end-to-end fidelity. Our solution, named **FENDI**, is a *fully polynomial-time approximation scheme* to the formulated bi-criteria problem. We further show that the computed solution can be implemented with a post-selection-and-storage protocol[1] to achieve both the expected EDR and the end-to-end fidelity. With the help of discrete event simulation, we demonstrate that FENDI can be used to approximate the EDR-fidelity Pareto frontier of a network efficiently and show that existing algorithms exhibit a substantial gap from the approximate frontier that can be achieved by the post-selection-and-storage protocol. Our main contributions are summarized as follows:

1) We model a general quantum network with Werner states and derive an end-to-end fidelity parameter as a product of link and node attributes based on isotropic noise.
2) We present a novel decomposition theorem that bridges between the maximum achievable EDR and fidelity, based on a new *primitive entanglement flow (pflow)* abstraction.
3) Based on the above, we formulate a bi-criteria problem called *high-fidelity remote entanglement distribution (HF-RED)* between a pair of nodes and prove it is NP-hard.
4) We propose a *fully polynomial-time approximation scheme (FPTAS)* to maximize the worst-case end-to-end fidelity subject to a lower bound on the expected EDR.
5) We develop a discrete event quantum network simulator implementing the protocol, characterize the (approximate) EDR-fidelity frontier and compare existing protocols to the post-selection-and-storage protocol.

Organization: §II reviews background and related work. §III introduces the network model. §IV presents our decomposition theorem for characterizing the EDR-fidelity trade-off, formulating the HF-RED problem, and showing its NP-hardness. §V presents our approximation scheme, analysis and discussion. §VI presents simulation results. §VII concludes the paper.

II. Background and Related Work

The idea of a quantum network was first proposed by the DARPA quantum network project [23]. Early work in quantum networking focused on feasibility demonstration in ideal situations. Much of the literature has derived analytical and simulation models for quantum repeater chains [9], [25] and other specialized topologies including lattices [41], star [52] and ring-like topologies [11], [47]. In reality, a quantum internet is unlikely to have such ideal topologies due to physical and geographical limitations.

Recent studies have focused on *entanglement routing* in general quantum networks [12], [33]. A common approach was to find paths with the highest success probability using modified shortest path algorithms [51]. Shi and Qian [48] first showed that maximum-success paths do not lead to the highest throughput and proposed QCAST and QPASS with optimal single-path routing metrics. Zhaoand Qiao [62] proposed an algorithm to achieve higher throughput by provisioning redundant intermediate entanglements for swapping. Zeng et al. [60] proposed an integer programming-based solution using branch-and-price with very limited quantum memories. Chen et al. [13], [14] proposed a decentralized routing design for congestion avoidance through adaptive evaluation of neighbor nodes. Dai et al. [18], [19] proposed the first optimal remote entanglement distribution (ORED) protocol for end-to-end EDR maximization, giving an upper bound on the achievable expected EDR between a pair of nodes. The above works only considered the quantity (EDR) but ignored the quality (fidelity) of entanglements.

To enable high-quality quantum communication, some works have focused on ensuring or improving fidelity [32], [50]. Zhao et al. [64] derived an end-to-end fidelity model based on *bit flip errors* and proposed a purification-based fidelity-aware routing algorithm. Li et al. [34] further proposed end-to-end fidelity-guaranteed entanglement routing design to achieve high-performance and low-complexity routing. Pouryousef et al. [44] proposed a quantum overlay network architecture, utilizing entanglement purification to satisfy the fidelity requirements of end-user applications.

Despite many recent efforts, we find that most entanglement routing solutions are based on handcrafted heuristic algorithms that only provide a *lower bound* on the achievable EDR and/or a heuristic trade-off between EDR and fidelity. Many papers assume quantum memories are limited and ephemeral [48], [60], [64] and build their solutions directly upon this assumption. However, this limitation is mostly technological rather than fundamental, as demonstrated in recent rapid advances in quantum memories [7], [8], [16], [21], [45], and it remains unclear whether the above solutions will achieve close-to-optimal EDR or EDR-fidelity trade-off when memory technology matures. ORED [18] is the only protocol that provides an *upper bound* on the achievable EDR in a general quantum network, and the bound is shown to be tight with perfect memories. However, ORED does not consider fidelity. Hence, the resulting protocol may not be applicable to fidelity-sensitive application scenarios. Overall, there is no existing work that characterizes the fundamental EDR-fidelity trade-off in a complex, general quantum network scenario, which we aim to address in this paper.

III. System Model

In this section, we present preliminaries of a quantum network. Notations related to modeling are summarized in Table I.

A. Quantum Basics

Consider a common 2-state quantum system with orthonormal basis states $|0\rangle$ and $|1\rangle$. A quantum bit (*qubit*) is a superposition of $|0\rangle$ and $|1\rangle$, written as $|b\rangle = \alpha|0\rangle + \beta|1\rangle$, satisfying $|\alpha|^2+|\beta|^2 = 1$. A perfect measurement on $|b\rangle$ yields classical state 0 with probability $|\alpha|^2$ and 1 with probability $|\beta|^2$. A two-qubit system is a superposition of four basis states $|00\rangle$, $|01\rangle$, $|10\rangle$ and $|11\rangle$. Let $|b_1b_2\rangle = \alpha_{00}|00\rangle + \alpha_{01}|01\rangle + \alpha_{10}|10\rangle + \alpha_{11}|11\rangle$, such that $|\alpha_{00}|^2+|\alpha_{01}|^2+|\alpha_{10}|^2+|\alpha_{11}|^2 = 1$. A maximally entangled pair (Bell pair) is a two-qubit system in one of the four *Bell states*: $|\Phi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle)$, and

[1]In a post-selection-and-storage protocol, the qubits can be stored in the quantum memories of repeaters after being post-selected following successful generation or swapping, and waiting for the next quantum operation.

TABLE I
KEY NOTATIONS IN MODELING

Parameters	Description
$G = (N, L)$	quantum network with nodes N and links L
c_l, F_l	capacity and fidelity of link l
W_l, W_n	fidelity loss parameters of link l and node n
q_l, q_n	ebit generation & swapping success probabilities
$F^{\mathsf{E2E}}, P^{\mathsf{E2E}}_{st}$	the end to end fidelity and success probability
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	induced graph of an eflow or pflow
η_{st}	the expected EDR between SD st
mn	an enode, *i.e.*, an unordered pair of nodes $m, n \in N$
Ψ_{st}	the set of all possible pflows between s and t
Δ_{st}	expected EDR bound between s and t
Υ_{st}	end-to-end fidelity bound between s and t
$\mathcal{P}_{st}$	the set of st-pflows with fidelity no lower than Υ_{st}
Variables	**Description**
g_{mn}	elementary ebit generation rate along link $mn \in L$
f^{mk}_{mn}	rate of mk-ebits to be swapped to generate mn-ebits
$I(mn)$	total ebit rate generated between node pair mn
$\Omega(mn)$	total ebit rate contributed by mn to swapping
$\eta(\psi)$	the pflow value (expected EDR) assigned to $\psi \in \Psi_{st}$

$|\Psi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle)$. A Bell pair is *maximally entangled* since it only contains two of the four basic states with equal probability, where in both states the two qubits are perfectly correlated. Bell pairs (also called ***ebits***) form the basis of two-party quantum communications: if Alice and Bob each holds one of two entangled qubits, they can use this pair to send any single-qubit quantum state via *local operations and classical communication (LOCC)*. Bell pairs can also be used to construct arbitrary multipartite entangled states needed by applications such as distributed quantum sensing [55].

B. Quantum Operations

Entanglement generation: A quantum network mainly relies on the generation and transmission of photonic entangled states. A pair of entangled photons is first generated by a physical process such as *spontaneous parametric down-conversion (SPDC)* at an entanglement source. Then, both photons are transmitted to two nearby nodes via a quantum link[2]. The photons can be transmitted via links such as optical fiber, free space, or optical switch, but suffer from transmission loss that is commonly exponential to the distance traversed [43], [49]. We consider generating an entangled photon pair and transmitting one/both photons jointly as *entanglement generation*. Entanglements generated via this process are *elementary ebits*.

Notably, this is a *probabilistic* process because of both the generation process with non-linear optics and the probabilistic transmission loss. A heralding and post-selection process is commonly employed after this process to detect successfully entangled and transmitted pairs, and the process can be repeated many times until one entangled pair is generated.

Entanglement swapping: Considering photon loss during transmission, entanglement swapping via quantum repeaters is essential for long-distance entanglement distribution. A swap takes two remote entangled pairs as input—each with one photon on a shared repeater. The repeater first entangles the two local photons, performs Bell state measurement (BSM), and then sends the result to either of the two remote nodes via classical communication. The node receiving the result then performs a local unitary operation on its own qubit, and the two remote photons become entangled without physical interaction.

[2]Alternatively, the entanglement source can be placed at a repeater, then only one photon needs to traverse the link to the other repeater.

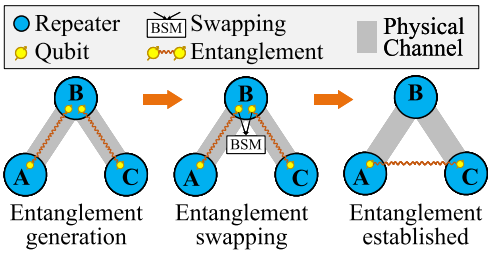


Fig. 1. Basic quantum network operations: entanglement generation and entanglement swapping.

Similar to generation, swapping is also *probabilistic* with near-term devices. Fundamentally, the success probability of linear optics-based BSM cannot exceed 50% without auxiliary photons, since two of the four Bell states are not distinguishable [10], [28]. To boost the BSM success probability, auxiliary single- or entangled-photon states may be used, and success rates of 62.5% and 78.125% have been demonstrated experimentally with single-photon ancillae using single photon detectors and photon number resolvers, respectively [3] and [24]. In principle, the success probability of BSM can be boosted to be arbitrarily close to 100% with an infinite number of ancillae and photon detectors [26], but the requirement of simultaneously generating many indistinguishable single photons at once (or storing them in quantum memories) significantly limits the applicability of such schemes. Besides theoretical limits, device deficits may further degrade the success probability.

Fig. 1 illustrates the process of entanglement distribution. Two *elementary ebits* are first generated along links A–B and B–C via entanglement generation. To swap, B entangles and measures its two local qubits and sends the result to either A or C via classical communications. According to the result, A or C applies a unitary operation on its qubit. If all operations succeed, the two qubits at A and C are then entangled without interacting with each other. This can be done recursively along a path until an end-to-end ebit between source and destination[3] is established for quantum information exchange.

C. Quantum Network Model

Formally, we model a quantum internet with an undirected graph $G = (N, L)$, where N is the set of quantum repeaters, and L is the set of physical channels (links) between repeaters. Each link $l \in L$ has a capacity $c_l \in \mathbb{Z}^+$, denoting the number of channels that can be attempted for ebit generation; $\mathbb{Z}^+$ denotes the positive integer set. To model the aforementioned probabilistic processes, we assume each link $l \in L$ has a success probability q_l for entanglement generation and each repeater $n \in N$ has a success probability q_n for swapping.

To ease illustration and facilitate comparison to existing work, we adopt a time-slotted system model following existing

[3]Although entanglements are undirected, we use traditional network terms "source" and "destination" to denote an undirected pair of end nodes involved in quantum communications for simplicity.

work [62], [64] where each time slot consists of four phases that are executed in a sequential order[4].

1) **Entanglement generation:** For a pair of nodes $mn \in L$ with a direct link, they will attempt to generate elementary mn-ebits at a pre-defined rate.[5]
2) **Entanglement swapping:** When ebits are available between both node pair mk and node pair kn sharing a common repeater k, repeater k can attempt to perform entanglement swapping between each pair of mk- and kn-ebits to create ebits between remote nodes m and n.

We assume a central controller collects network information, monitors network status, and allocates resources by defining the rates of generation and the order of swapping.

D. Quantum Noise and Fidelity

While the above models assume perfect quantum channels and operations—meaning the final distributed ebits are in the exact same state as the generated ones—the inevitable noise in quantum operation and transmission can introduce error and make the final state differ from the initial state. In classical communication, errors can be measured, detected, and corrected on-the-fly or end-to-end. In quantum, however, errors cannot be detected without destroying the quantum state due to the no-cloning theorem. Thus when a pure entangled state is affected by noise, it becomes a *mixed state* that cannot be distinguished from the pure state without measurement.

Let $|\Phi^+\rangle$ be our desired pure entangled state[6]. A mixed state M can result from $|\Phi^+\rangle$ through a noisy channel or noise in quantum operations. Fidelity is a key quantum metric quantifying how close a mixed state is to the desired state, defined as $F \triangleq \langle\Phi^+|M|\Phi^+\rangle$, and denoting the probability that M (represented by a density matrix) is in the desired state $|\Phi^+\rangle$. To provide a rigorous fidelity guarantee, we assume a worst-case isotropic error model [53], as compared to the bit flip error model in [64]. As in [6], an arbitrary mixed state M with fidelity F can be transformed to a Werner state with the same F via *random bilateral rotations* (RBR) as

$$\mathcal{W}_F = F|\Phi^+\rangle\langle\Phi^+| + \frac{1-F}{3}(|\Phi^-\rangle\langle\Phi^-| + |\Psi^+\rangle\langle\Psi^+| + |\Psi^-\rangle\langle\Psi^-|).$$

This Werner state can be viewed as a mixture of the pure state $|\Phi^+\rangle$ with isotropic noise [53]. We assume all elementary and intermediate mixed-state ebits are transformed to the Werner state before further operation.

For an elementary ebit generated along a physical channel l, we define the fidelity as $F_l \in [0,1]$, which is decided by the quantum circuit that generates the entanglements and the channel noise during transmission.

Given two ebits with fidelity F_1 and F_2 respectively and a *perfect* swapping performed between them, the presence of noise in the ebits means that even a perfect swapping might still fail due to the two ebits not being in the desired state $|\Phi^+\rangle$, leading to measurement error. Two cases may result in a successful swap: 1) both ebits were in $|\Phi^+\rangle$ with probability $F^* = F_1F_2$, in which case the swapped ebit is also in $|\Phi^+\rangle$; 2) both ebits were not in $|\Phi^+\rangle$ but had equal states with probability $F^{**} = 3\frac{(1-F_1)}{3}\frac{(1-F_2)}{3}$, in which case the swapped ebit is in another Bell state instead of $|\Phi^+\rangle$, but can be transformed to $|\Phi^+\rangle$ via LOCC [6]. In the other cases, the swap fails because of unknown and unequal states of the two ebits. By combining these cases, a *perfect* entanglement swap will result in a new ebit with fidelity F' [22], where

$$F' = F^* + F^{**} = \frac{1}{4}\cdot\left(1 + 3\frac{(4F_1-1)}{3}\frac{(4F_2-1)}{3}\right). \quad (1)$$

In practice, the swapping operation is also noisy or imperfect, and hence incurs additional fidelity loss. Such loss is due to the (un)reliability of BSM, 1-qubit operation, and 2-qubit operation involved. For instance, if a swap is performed with two elementary ebits with F_1 and F_2 at node n where the accuracy of BSM and probabilities of ideal 1-qubit, 2-qubit operations are α_n, $o_{1,n}$, and $o_{2,n}$, respectively, the fidelity of a successfully generated ebit after swapping [22] is

$$F' = \frac{1}{4}\cdot\left(1 + 3o_{1,n}o_{2,n}\frac{4\alpha_n^2-1}{3}\frac{4F_1-1}{3}\frac{4F_2-1}{3}\right). \quad (2)$$

Based on Eq. (2), we facilitate notation by defining fidelity parameters $W_l \triangleq \frac{4F_l-1}{3}$ and $W_n \triangleq o_{1,n}o_{2,n}\frac{4\alpha_n^2-1}{3}$ for each link l and repeater n respectively, and the fidelity of a successfully generated ebit after swapping is

$$F' = \frac{1}{4}\cdot(1 + 3W_1W_2W_n). \quad (3)$$

Assume an end-to-end ebit is established by swapping elementary ebits created along links $\{l_1, l_2, \ldots, l_{X+1}\} \subseteq L$ recursively at nodes $\{n_1, n_2, \ldots, n_X\} \subseteq N$. Recursively applying Eq. (3), the end-to-end fidelity of the ebit is

$$F^{\text{E2E}} = \frac{1}{4}\cdot\left(1 + 3\prod_{i=1}^{X+1}W_{l_i}\prod_{j=1}^{X}W_{n_j}\right). \quad (4)$$

From Eq. (4), the end-to-end fidelity decreases exponentially with increasing number of hops [31]. Eq. (4) will serve as the basic tool to quantify and optimize the end-to-end fidelity.

Note that fidelity cannot be measured for a single ebit—the measurement itself will destroy the ebit. As such, fidelity parameters can only be inferred from measuring and profiling ebits generated on elementary links (or after swap) for many times. We assume that each node or link will be independently profiling the W_n or W_l value continuously throughout the network operations and will regard these values (or some binary encoding of them) as input to further modeling and formulation.

E. Performance Metrics and Problem Statement

We consider two important performance metrics in the end-to-end entanglement distribution process, which has also been widely adopted in the literature:

1) **Entanglement distribution rate (EDR):** Similar to throughput in a classical network, EDR is the number of ebits distributed between an SD pair in unit time. Due

[4]The assumption on time slots is non-restrictive, as our design can be extended to the continuous-time model with the help of quantum memories. With even short-lived memories as temporary buffers, the four phases can be executed in parallel in continuous time, each phase independently based on (probabilistic) output from the other phases. With more advances in long-term optical quantum memories [21], [45], we believe quantum repeaters built upon continuous-time asynchronous operations will be more realistic in the near future and provide better performance (such as EDR) [56], [63].

[5]We use mn to abbreviate an unordered node pair $\{m,n\}$. Hence $mn = nm$.

[6]Since all Bell states are symmetric, we use $|\Phi^+\rangle$ as the desired state without loss of generality throughout this paper.

to the probabilistic operations, we use η_{st} to denote the *expected EDR* between the source s and destination t.

2) **End-to-end fidelity:** A higher end-to-end fidelity F^{E2E} leads to higher communication efficiency.

For realistic applications, it is commonly required that both metrics are high enough. For instance, in distributed quantum computing, the EDR decides how long a data qubit needs to wait for an ebit for teleportation, while the fidelity decides the quality of the data qubit after teleportation. A general quantum network needs to support applications with varying needs in terms of EDR and fidelity. It is thus important to (1) characterize the achievable EDR and fidelity between an arbitrary SD pair in a general quantum network and (2) devise arbitrary trade-offs between achievable EDR and fidelity.

While the first task has been somewhat tackled in existing work, only heuristic solutions exist for the second task. The difficulty lies in simultaneously optimizing for two metrics while having theoretical guarantee on both. In the classical network, such a problem is called *quality-of-service (QoS) routing* [59]. Compared to the classical network, the difficulty in quantum is that the path-based formulation—while enabling clear modeling of end-to-end fidelity as in Eq. (4)—does not lead to any optimality or approximation guarantee on the achievable EDR, while the EDR-optimal formulation (ORED) cannot encode end-to-end fidelity being linear program (LP)-based. To tackle these challenges, we aim to achieve two goals:

1) Define a unifying mathematical abstraction that can accurately characterize both the achievable EDR and end-to-end fidelity between an SD pair in a quantum network.
2) Design an algorithm to find an arbitrary (approximately) optimal trade-off between achievable EDR and fidelity.

We address the first in Sec. IV, and the second in Sec. V. Specifically, we provide a formal trade-off problem definition in Sec. IV-C and computational complexity in Sec. IV-D.

IV. Characterizing Achievable EDR and Fidelity

A. Characterizing Achievable Expected EDR

We start with the question of how to characterize the maximum achievable EDR between two nodes in a given network. Assuming no quantum memory is available, the generated ebits would decohere within one time slot, meaning that all generation and swapping processes along an end-to-end path must succeed in unit time. The probability of successful generation along one path is thus the product of all node and link probabilities: $P_{st}^{\mathsf{E2E}} = \prod_{i=1}^{X+1} q_{l_i} \prod_{j=1}^{X} q_{n_j}$ for a path $\rho = (n_0, n_1, \ldots, n_{X+1})$ where $s = n_0$, $t = n_{X+1}$ and $l_i = n_{i-1}n_i \in L$. The achievable expected EDR is then the bottleneck capacity $c_{st}^* \triangleq \min_i\{c_{l_i}\}$ times the end-to-end success probability: $\eta_{st} = c_{st}^* \cdot P_{st}^{\mathsf{E2E}}$.

It is expected that future quantum repeaters will be equipped with quantum memories acting as temporary buffers. In this case, rate characterization becomes more complicated. In [48], it has been shown that post-selection and storage can increase the maximum achievable EDR beyond the simple product of probabilities times capacity, since the quantum memories can temporarily buffer and rematch the post-selected ebits that are unmatched for swapping due to unsuccessful ebit generation on other links. Subsequently, many works have explored how to design entanglement routing and distribution protocols with limited or ephemeral quantum memories to improve EDR [56], [62], [64]. However, the extent to which post-selection and storage can increase the optimal expected EDR remains unclear.

A recent breakthrough is a tight *upper bound* on the maximum achievable EDR between a pair of nodes with post-selection and storage as in [18], [19]. Their result is based on an abstraction called the *entanglement flow*, or **eflow**, which formulates the maximum achievable expected EDR as a linear program. Below, we present the general definition of an eflow in [19], which will be used in our decomposition theorem.

Definition 1 (Eflow [19]): Given a network $G = (N, L)$ and an SD pair st, an eflow in G is defined by variables noitemsep,topsep=0pt

- *$g_{mn} \in [0, 1], \forall mn \in L$, denoting the rate of elementary ebit generation along the physical link mn, as a ratio of the capacity c_{mn} of the link, and*
- *$f_{mn}^{mk} \geq 0, \forall m, n, k \in N$, denoting the expected rate of ebits established between nodes m and k that will be used for swapping to generate ebits between nodes m and n.*

A feasible eflow must have f and g satisfying:

$$f_{mn}^{mk} = f_{mn}^{kn}, \quad \forall m, n, k \in N; \tag{5a}$$

$$I(mn) = \Omega(mn), \quad \forall m, n \in N, mn \neq st; \tag{5b}$$

$$\Omega(st) = 0; \tag{5c}$$

where for $\forall m, n \in N$,

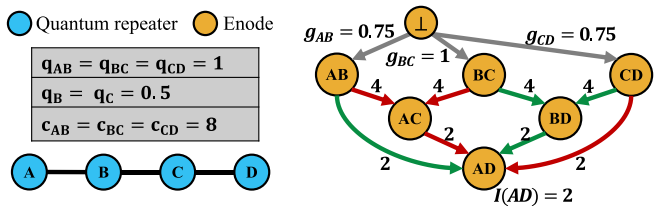
$$I(mn) \triangleq q_{mn} c_{mn} g_{mn} \cdot \mathbf{1}_{mn \in L} + \sum_{k \in N \setminus \{m,n\}} \frac{q_k}{2} \left(f_{mn}^{mk} + f_{mn}^{kn}\right), \tag{5d}$$

$$\Omega(mn) \triangleq \sum_{k \in N \setminus \{m,n\}} \left(f_{mk}^{mn} + f_{kn}^{mn}\right), \tag{5e}$$

*and $\mathbf{1}_{mn \in L}$ is an indicator function of whether $mn \in L$. The **eflow value** of SD pair st is defined as $\eta_{st} \triangleq I(st)$.* □

Explanation: For brevity, each node pair mn with $m, n \in N$ is called an **enode**, meaning that post-selected ebits may be established between the pair of nodes at some stage of remote distribution. Here $I(mn)$ denotes the ebits generated between mn (including elementary ebits and ebits generated by swapping), and $\Omega(mn)$ denotes the ebits contributed by mn to generate ebits between other node pairs via swapping. Note that the elementary ebits (first term in Eq. (5d)) are discounted by generation probability q_{mn}, and ebits received from swapping at node k are discounted by swapping probability q_k. Eq. (5a) enforces the two pairs mk and kn, whose ebits will be swapped to form ebits for mn, contribute equal number of ebits. Eq. (5b) enforces an intermediate pair mn does not keep generated ebits, but contributes all ebits for establishing end-to-end ebits. Eq. (5c) constraints that the SD pair st should not contribute any established ebits to further swapping. An eflow describes how ebits "flow through" different enodes and "merge" at repeaters until some are "landed in" (established between) the SD pair st, with flow conservation at repeaters as in (5b).

One way to visualize an eflow is to define its *induced graph*, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \subseteq (N \times N) \cup \{\perp\}$ is a set of enodes (with a special enode $\perp$ denoting the generation process), and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ are directed edges denoting generation and swapping processes. An enode $mn \in \mathcal{V}$ corresponds to one with $I(mn) > 0$ in the eflow. An edge $(mk, mn) \in \mathcal{E}$ then denotes one swapping variable $f_{mn}^{mk} > 0$. An edge $(\perp, mn) \in \mathcal{E}$ specially denotes a generation variable $g_{mn} > 0$.



(a) Quantum network with SD pair A-D (b) Induced graph of eflow with EDR=2

Fig. 2. Induced graph of an eflow. Value on each edge denotes a variable: g_{mn} or f^{mk}_{mn}. Two same-color edges pointing to one enode are matched for swapping.

From Eq. (5a), it is clear that swapping edges (mk, mn) and (kn, mn) must appear simultaneously in $\mathcal{G}$—either they both present or they both absent. An example is shown in Fig. 2.

We summarize the importance of the eflow formulation with the following theorem, which restates Theorems 1–2 in [19].

Theorem 1 (Characterizing Maximum EDR [19]): The optimal solution to the following problem (called ORED in [19]),

$$\eta^*_{st} \triangleq \max_{f,g}\{\eta_{st} \,|\, (5a)\text{–}(5e)\}, \tag{6}$$

is a tight upper bound on the maximum expected EDR between s and t in G. The induced graph $\mathcal{G}$ of at least one optimal solution is a directed acyclic graph (DAG). Furthermore, there exists an entanglement distribution protocol that can achieve expected EDR of η^*_{st} between st. □

The primary limitation with the eflow formulation is that it **cannot model ebit fidelity loss** in generation and swapping. Since ebits may arrive at an enode from any possible sequence of swaps at arbitrary repeaters, there may be an exponential number of possible paths in G from which an ebit might have been generated, and some may result in low fidelity that can render the distributed ebits unusable. In the next subsection, we propose a novel abstraction, called *primitive eflow*, to characterize the end-to-end fidelity of the distributed ebits.

B. Eflow Decomposition & Characterizing End-to-End Fidelity

To characterize the end-to-end fidelity, we adopt an abstraction, named **primitive eflow (pflow)**, that naturally encodes the fidelity of entanglement paths while still maintaining the same tight upper bound on achievable EDR. This enables an alternative formulation that is equivalent to Program (5), and is similar to the path-flow formulation in classical network flow as an alternative to the edge-flow formulation [2]. We establish this equivalence with a novel *eflow decomposition theorem*.

Definition 2 (Pflow): A primitive eflow (pflow) *is a feasible eflow as defined by Program (5), which additionally satisfies that: for every enode mn, either $g_{mn} > 0$, or there exists exactly one $k \in N$ such that $f^{mk}_{mn} = f^{kn}_{mn} > 0$, but not both.* □

A pflow is *primitive* in that ebits at each enode mn is generated in exactly one way: either they are elementary ebits generated directly along link $mn \in L$, or they are generated by swapping mk- and kn-ebits at a single intermediary k. The induced graph $\mathcal{G}$ of a pflow, excluding the special $\perp$ vertex, is always a binary tree rooted at enode st by the definition; Fig. 2 shows two such binary trees with different colors. A pflow naturally represents exactly one *path* in the quantum internet, and the final st-ebits generated along a pflow have **identical fidelity**, which can be directly computed via Eq. (4).

Algorithm 1 Computing Ebit Generation Ratios of a Pflow

Input: Induced graph $\mathcal{G}$ of an st-pflow
Output: Ebit generation ratios $\{\bar{g}_{mn}, \bar{f}^{mk}_{mn}, \bar{f}^{kn}_{mn}\}$

```
Initialize all ratios to 0, and Q ← {(st, 1)};
while Q ≠ ∅ do
    (mn, ψ) ← Q.pop();
    if ∄k such that (mk, mn) ∈ E then
        ḡ_mn ← g_mn + ψ/(q_mn · c_mn);
    else
        f̄^mk_mn ← f̄^mk_mn + ψ/q_k, f̄^kn_mn ← f̄^kn_mn + ψ/q_k;
        Q.push((mk, ψ/q_k)), Q.push((kn, ψ/q_k));
return {ḡ_mn, f̄^mk_mn, f̄^kn_mn}.
```

Another property of a pflow is that the ratio between each variable in $\{g_{mn}, f^{mk}_{mn} \,|\, m, k, n \in N\}$, and the end-to-end EDR η_{st}, is fixed. Let $\bar{g}_{mn}$ or $\bar{f}^{mk}_{mn}$ be the ratio between the corresponding variable and the EDR of the pflow. Given the induced graph $\mathcal{G}$ of the pflow, these ratios can be computed as in Algorithm 1, backtracking from enode st which has a ratio of 1 (one generated ebit between st translates into one end-to-end st-ebit). For each enode mn, its output ebit rate $\Omega(mn)$ is added to its input ebit rate(s), *i.e.*, either $\bar{g}_{mn}$ or $\bar{f}^{mk}_{mn}$ and $\bar{f}^{kn}_{mn}$ for some k, augmented by the corresponding expected ratios of $1/q_{mn}$ or $1/q_k$ respectively. Based on Algorithm 1, a pflow can essentially be defined by its induced graph $\mathcal{G}$, and a single objective value $\eta_{\mathcal{G}}$ assigned to this pflow.

Crucially, the pflow abstraction leads to the following theorem (with the proof delegated to Appendix A), which generalizes the classical *flow decomposition theorem* [2] to the quantum network setting:

Theorem 2 (Eflow Decomposition): An eflow with $\eta_{st} > 0$ can be decomposed into a polynomial number of pflows. □

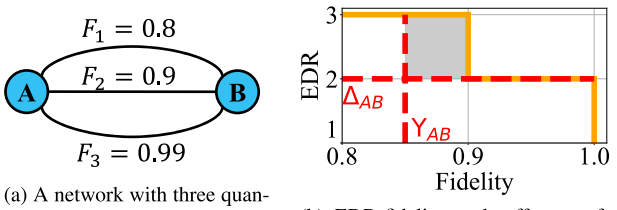
Theorem 2 enables an alternative pflow-based formulation to Program (5) in Definition 1. Let Ψ_{st} be the set of all possible pflows between s and t, and let $\eta(\psi) \geq 0$ be the pflow value assigned to $\psi \in \Psi_{st}$. Lemma 1 follows from Theorem 2:

Lemma 1 (Pflow-based EDR Characterization): The maximum expected EDR η^*_{st} in Eq. (6) can be computed by Program (7):

$$\begin{aligned} \eta^*_{st} = \max_{\eta} \quad & \sum\nolimits_{\psi \in \Psi_{st}} \eta(\psi) \\ \text{s.t.} \quad & \sum_{\psi \in \Psi_{st}: mn \in \psi} \bar{g}_{mn} \cdot \eta(\psi) \leq 1, \quad \forall mn \in L. \end{aligned} \tag{7}$$

Program (7) computes η^*_{st} by assigning values to pflows in Ψ_{st}, while making sure that no link $mn \in L$ is oversubscribed beyond a ratio of 1, i.e., being asked to generate more than c_{mn} ebits per unit time. Thus, the key observation is that each actual ebit is still generated along a single entanglement path. The fidelity of the ebit is precisely defined by the path along which it is generated based on Eq. (4). Assume an eflow can generate ebits all with fidelity no less than a given bound Υ_{st}. Following Lemma 1, the eflow can always be decomposed into a set of pflows, where each pflow generates ebits along a fixed path with fidelity lower bounded by Υ_{st} (some of the pflows may share the same path). This leads to Theorem 3.

Theorem 3 (Characterizing Worst-Case Fidelity): An eflow that generates ebits with minimum end-to-end fidelity Υ_{st} can



(a) A network with three quantum links and different fidelity. All links have unit capacity and 1.0 success probability.

(b) EDR-fidelity trade-off curve for SD pair AB, and a solution in the frontier.

Fig. 3. The EDR-fidelity Pareto frontier of a simple network. Shaded area denotes gap from a solution to the actual frontier.

be decomposed into a set of pflows, each along an st-path whose fidelity is at least Υ_{st}. □

Remark: The importance of Theorem 3 is not to characterize the maximum end-to-end fidelity for generating a single ebit. Such maximum fidelity can be easily computed by employing Dijkstra's algorithm and finding the highest-fidelity path following Eq. (4). Instead, the goal is to characterize the worst-case end-to-end fidelity for achieving an end-to-end EDR goal, or vice versa, utilizing as many paths/pflows as possible.

C. Trade-off Between EDR and Worst-Case Fidelity

Consider a quantum application having two performance requirements for remote entanglement distribution: 1) the long-term average EDR is at least Δ_{st}; 2) each generated ebit has fidelity no less than Υ_{st}. Having a higher EDR goal Δ_{st} means the network may need to utilize more paths for distribution, some maybe leading to lower end-to-end fidelity than others, which overall may lead to a lower Υ_{st} that can be satisfied.

Consider an SD pair A and B in Fig. 3(a) which are connected by three different quantum links, all with capacity 1 but different fidelity. When the end-to-end fidelity requirement increases, the achievable EDR will decrease as the number of feasible paths/pflows becomes less, and vice versa, as shown in Fig. 3(b). The trade-off can become more complicated when swapping probability and fidelity loss are taken into account.

We start to explore this trade-off from the motivating example, which is to simultaneously satisfy the expected EDR and fidelity as shown in Fig. 3(b). Therefore, we define the high-fidelity remote entanglement distribution (HF-RED) problem.

Definition 3 (HF-RED): Given a quantum network $G = (N, L)$ *and an SD pair st, let* $\Delta_{st} > 0$ *be the expected EDR bound and* $\Upsilon_{st} > 0$ *be the end-to-end fidelity bound. The* ***high-fidelity remote entanglement distribution*** *problem (denoted as* ***HF-RED****) is to seek a set of pflow* $\mathcal{P}^*_{st} \subseteq \Psi_{st}$*, which delivers end-to-end st-ebits satisfying that*

1) total expected EDR η_{st} *of all pflows is at least* Δ_{st}*, and*
2) each pflow has fidelity no less than Υ_{st}. □

Without loss of generality, we further define the optimization version of HF-RED as **OF-RED** to maximize the worst-case end-to-end fidelity subject to the expected EDR bound.

Definition 4 (OF-RED): Let $G = (N, L)$ *be an undirected graph with node set* N *and link set* L*. Let* s *be a source node and* t *be a destination node. Let* $\Delta_{st} > 0$ *be an expected entanglement distribution rate (EDR). The* ***optimal-fidelity remote entanglement distribution*** *problem (denoted as OF-RED) is to seek a set of pflow, which delivers end-to-end* (s, t)*-ebits satisfying that*

1) total expected EDR η_{st} *of all pflows is at least* Δ_{st}*, and*
2) the minimum end-to-end fidelity of all pflows is maximized.

We note that OF-RED is an important problem for characterizing the EDR-fidelity trade-off. Particularly, one can apply the ϵ-constraint method in multi-objective optimization [37] to find *weak Pareto optimal solutions*—solutions that cannot be improved on one of the metrics without hurting the other—by repetitively solving OF-RED with different bounds on the expected EDR. In Sec. V, we utilize this method depending on solving OF-RED efficiently to characterize the EDR-fidelity trade-off curve, which, nevertheless, is highly non-trivial.

D. Computational Complexity

Let $\mathcal{P}_{st} \subseteq \Psi_{st}$ be the set of st-pflows that are along paths with fidelity no lower than Υ_{st}. HF-RED can be easily formulated based on Program (7), by replacing Ψ_{st} with $\mathcal{P}_{st}$ in the formulation—this constrains the program to only use pflows satisfying the end-to-end fidelity constraint Υ_{st} when trying to achieve the EDR goal Δ_{st}. Notably, both Program (7) and this fidelity-aware version are linear programs (LPs), but with exponential sizes due to the potentially exponential number of possible pflows in Ψ_{st} (or $\mathcal{P}_{st}$). In fact, the following lemma demonstrates the computational complexity of this problem:

Lemma 2: HF-RED and OF-RED are NP-hard. □

Proof: We prove NP-hardness of HF-RED by a reduction from the *Multi-Path routing with Bandwidth and Delay constraints (MPBD)* problem, which is NP-complete [38]. Given a graph, an SD pair and two values $B, D > 0$, MPBD asks for a set of paths with delay upper bounded by D, and a network flow over these paths with total flow lower bounded by B. Given an MPBD instance, let us build an instance of HF-RED. First, we set all probabilities q_l and q_n to 1. Then we set $W_l = e^{-d_l}$ where $d_l > 0$ is the delay of link l, and $W_n = 1$ for $n \in N$. Note that since $d_l > 0$, $W_l \in (0, 1)$. The fidelity bound is $\Upsilon_{st} = (1 + 3 \cdot e^{-D})/4$. Capacity c_l is set as the bandwidth in MPBD, and EDR bound $\Delta_{st} = B$. Given this construction, any generated ebit represents a path p such that $\prod_{l \in p} W_l = e^{-\sum_l d_l} \geq e^{-D}$, which gives $\sum_l d_l \leq D$. Meanwhile, for any delay-feasible path in MPBD, generating end-to-end ebits along this path will satisfy the fidelity bound Υ_{st}. Since generation and swapping both have success probability 1, the EDR is exactly equal to the end-to-end st-flow value. Hence a solution to MPBD gives a feasible solution to HF-RED, and vice versa. HF-RED is thus NP-hard, and the NP-hardness of OF-RED follows. □

Remark (from fidelity to length): We utilize the above proof to transform end-to-end fidelity in Eq. (4) into an additive metric. Define *length* values $\zeta_l = -\log(W_l)$ and $\zeta_n = -\log(W_n)$ for link and node fidelity values, respectively. Consider end-to-end fidelity F^{E2E} of a path in Eq. (4). Define the path length as $Z = \sum_{i=1}^{X+1} \zeta_{l_i} \sum_{j=1}^{X} \zeta_{n_j}$, then $F^{\mathsf{E2E}} = \frac{1}{4} \cdot \left(1 + 3\, e^{-Z}\right)$. Since the above transformation is bijective, maximizing the worst-case fidelity is equivalent to minimizing the longest path length. Given a fidelity bound Υ_{st}, it is also easy to define an equivalent length bound $\mathbf{Z}_{st} = -\log\left(\frac{4\Upsilon_{st}-1}{3}\right)$, such that any path with length upper bounded by $\mathbf{Z}_{st}$ will have fidelity lower bounded by Υ_{st}, and vice versa. Note that using either W_l, W_n or ζ_l, ζ_n only differs in the *binary encoding* to represent the fidelity parameters. Because of the equivalence, we next focus on **minimizing the maximum path length** in OF-RED.

TABLE II
KEY NOTATIONS FOR ALGORITHM DESIGN

Parameters	Description
mn/z	extended enode mn with a path segment length of z
υ^*_{st}	optimal worst-case end-to-end fidelity between s and t
ζ_l, ζ_n	length values of link l and node n
ς	Boolean output of the approximate testing algorithm 2
ε	approximation accuracy parameter
θ	quantization factor of node/link lengths
$\zeta(p), \zeta^\theta(p)$	lengths of path p before & after quantization with θ
$\mathbf{Z}, Z$	path length bounds before and after quantization
$\mathbf{Z}^*, Z^*$	original optimal longest path length, and quantized value
$\mathbf{Z}^\theta$	optimal longest path length for quantized OF-RED
LB, UB	lower & upper bounds on optimal longest path length $\mathbf{Z}^*$
$Z_{\text{LB}}, Z_{\text{UB}}$	the quantized lower and upper bounds

V. FPTAS FOR OPTIMIZING FIDELITY UNDER EDR BOUND

The OF-RED problem aims to search for the highest worst-case fidelity υ^*_{st} (equivalently the minimum longest path length $\mathbf{Z}^*_{st}$) under minimum end-to-end EDR requirement Δ_{st} as in Fig. 3. The problem being NP-hard means no polynomial-time algorithm can solve the problem optimally unless P=NP. In this case, we propose a fully polynomial-time approximation scheme (FPTAS), which is theoretically the best polynomial-time algorithm one can hope for in this circumstance.

Definition 5 (FPTAS [59]): Given a minimization problem Λ, an algorithm $\mathcal{A}$ is a fully-polynomial approximation scheme (FPTAS) for Λ, iff for any instance of Λ with optimal objective value ξ^* and given an arbitrary constant factor a, the algorithm $\mathcal{A}$ could always find a feasible solution with objective value $\xi \geq (1+\varepsilon)\cdot\xi^*$, within time polynomial to input size and $1/\varepsilon$.

An FPTAS can achieve accuracy *arbitrarily close* to the optimal solution while incurring only a polynomial growth on time complexity over the accuracy parameter $1/\varepsilon$. In other words, it provides full flexibility to the accuracy-complexity trade-off, with each constant ε defining a polynomial-time constant-factor approximation algorithm. Our goal in this section is to design an FPTAS for the OF-RED problem, which can be used as a tool to characterize the approximate weak Pareto frontier of the EDR-fidelity trade-off. Our algorithm is designed as a non-trivial extension to existing QoS routing algorithms [27], [30], [59], with fundamentally different abstractions for routing paths (pflows) and end-to-end evaluation of the entanglement metrics.

Solution Overview: Our FPTAS includes four building blocks. Notations related to algorithms are summarized in Table II.

First, we design a pseudo-polynomial-time *Fidelity-aware Optimal Remote Entanglement Distribution (FORED)* program as an extension to Program (5). Under the restrictive condition that all length values are integers, the program outputs an eflow achieving maximum EDR with lower-bounded length (fidelity).

Our second building block, an *approximate testing algorithm*, uses the FORED program as a sub-routine to test if a specific real-valued length can be achieved with the EDR bound, subject to a small and bounded testing error.

Our third building block is a polynomial-time *sorting and trimming algorithm*, which finds a pair of close-enough lower and upper bounds for the optimal length value, to serve as the initial range in which the optimal value will be searched for.

Finally, a *two-stage bisection search algorithm* iteratively narrows down the initial range via approximate testing until a solution is found within a small approximation error of the optimal length (fidelity) value while satisfying the EDR bound.

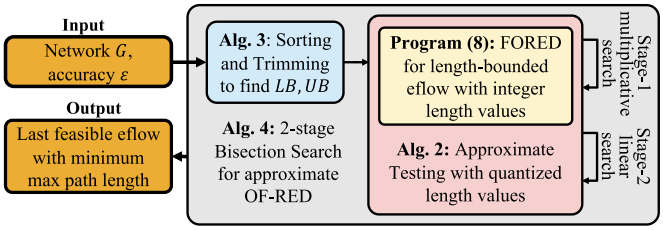


Fig. 4. The overall algorithmic framework of FENDI.

The overall algorithmic framework, named **FENDI**, is shown in Fig. 4. Given an approximation parameter $\varepsilon > 0$, our FPTAS can obtain a $(1+\varepsilon)$-approximation to the optimal longest path length in time polynomial to the network graph size $|N|$ and $1/\varepsilon$. Next, we design these building blocks one by one.

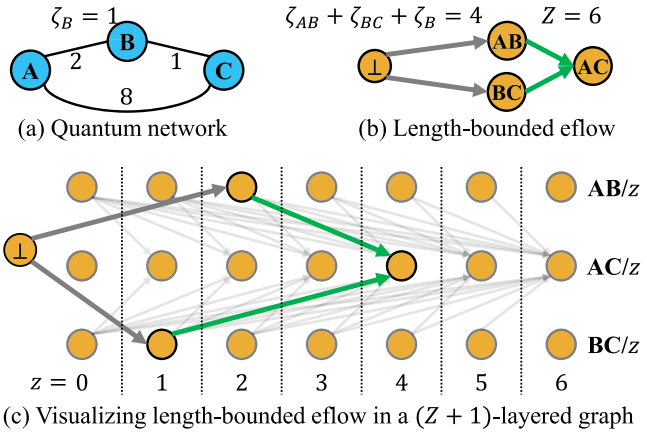
A. Fidelity-Aware Optimal Remote Entanglement Distribution

Summary: Our first building block aims to extend Program (5) into a new LP that maximizes expected EDR subject to an (integer) bound on the path lengths. Note that the objective (EDR) and constraint (fidelity) in this building block are reversed from the OF-RED definition, which is a necessary construction needed in later building blocks when searching for the approximately optimal length (fidelity).

Consider a given path length bound $\mathbf{Z}$, and assume all the length values ζ_l and ζ_n are *positive integers*. In this case, we assume the path length bound is also a positive integer without loss of generality, which we instead denote as Z to differentiate from a general, possibly non-integral path length $\mathbf{Z}$. We wish to find an eflow that maximizes the expected EDR, subject to the constraint that every pflow has its length bounded by Z.

Length-bounded eflow. The key to solving this "integral" problem optimally is to build the integer length values into the structure of the induced graph $\mathcal{G}$ of an eflow. Let $[Z] = \{0, 1, 2, \ldots, Z\}$. Consider two enodes mk and kn, whose ebits might be swapped to generate ebits between mn. Depending on how the mk- and kn-ebits are generated, we can divide the two enodes each into $Z+1$ copies, which we denote as *extended enodes* mk/z and kn/z, for $z \in [Z]$. Each enode mk/z denotes mk-ebits that are generated along a path with path length of exactly z. Because of the integer length bound Z, there are up to $Z+1$ different path length values (or equiv. $Z+1$ fidelity values) for ebits generated between each enode mn. When two enodes mk/z_1 and kn/z_2 swap, if the resulting length $z_1 + z_2 + \zeta_k > Z$, the resulting ebits will not satisfy the length/fidelity bound, and hence should be discarded. For elementary ebit generation, the initial enode is mn/ζ_{mn} if $\zeta_{mn} \leq Z$, reflecting the initial fidelity of the elementary ebits on link $mn \in L$.

Assume we have a three-node network shown in Fig. 5(a), and the goal is to establish AC-ebits either directly or through repeater B. Length values are marked beside nodes/links. Given a length bound $Z = 6$, direct generation along link



(a) Quantum network (b) Length-bounded eflow

(c) Visualizing length-bounded eflow in a $(Z+1)$-layered graph

Fig. 5. Example of a length-bounded eflow in a 3-node network in (a). AC is the SD pair. Length values ζ_l are marked beside links. Given a length bound $Z = 6$, there is one length-bounded eflow in (b). The eflow can be visualized in a $(Z+1)$-layered graph with all possible g and f variables as edges in (c).

AC would not be feasible with $\zeta_{AC} = 8$, and hence there is no extended enode $AC/8$ in Fig. 5(c). Meanwhile, the feasible eflow of swapping AB and BC to generate AC is visualized in Fig. 5(b)–(c), with the edges from extended enodes $AB/2$ and $BC/1$ to $AC/4$ given $\zeta_{AB} + \zeta_{BC} + \zeta_B = 2 + 1 + 1 = 4$.

FORED formulation. We extend ORED to FORED, whose solution (if feasible) is a length-bounded eflow achieving maximum expected EDR. We keep variables g_{mn} unchanged for $mn \in L$. For each f_{mn}^{mk}, we extend it to up to $O(Z^2)$ copies, denoted by $f_{mn/z}^{mk/z'}$, for $z' = [Z - \zeta_k]$, and $z = z' + \zeta_k, \ldots, Z$. In plain words, $f_{mn/z}^{mk/z'}$ denotes the *number of mk-ebits, with a path segment length of z', which contribute to swapping at node k to generate mn-ebits with a path segment length of z*. We then formulate FORED in Program (8):

$$\max_{f,g} \ \eta_{st}^{Z} \triangleq \sum\nolimits_{z=0}^{Z} I(st/z) \tag{8}$$

$$\text{s.t. } f_{mn/z}^{mk/z_1} = f_{mn/z}^{kn/z_2}, \quad \forall m,n,k \in N, \quad \forall z_1, z_2 \in [Z-\zeta_k], z = z_1 + z_2 + \zeta_k; \tag{8a}$$

$$I(mn/z) = \Omega(mn/z), \quad \forall z \in [Z], \ \forall m,n \in N, mn \neq st; \tag{8b}$$

$$I(st/z) = 0, \quad \forall z; \tag{8c}$$

where for $\forall m, n \in N$, $z \in [Z]$,

$$I(mn/z) \triangleq q_{mn} c_{mn} g_{mn} \cdot \mathbf{1}_{mn \in L, \zeta_{mn} = z} + \sum_{\substack{k \in N \\ \setminus \{m,n\}}} \sum_{z'=0}^{z-\zeta_k} \frac{q_k}{2} \left(f_{mn/z}^{mk/z'} + f_{mn/z}^{kn/(z-z'-\zeta_k)} \right), \tag{8d}$$

$$\Omega(mn/z) \triangleq \sum_{\substack{k \in N \\ \setminus \{m,n\}}} \left(\sum_{z'=z+\zeta_n}^{Z} f_{mk/z'}^{mn/z} + \sum_{z'=z+\zeta_m}^{Z} f_{kn/z'}^{mn/z} \right), \tag{8e}$$

and $\mathbf{1}_{mn \in L, \zeta_{mn} = z}$ denotes whether both $mn \in L$ and $\zeta_{mn} = z$.

Explanation: Objective (8) is to maximize the sum of end-to-end ebits generated over all paths of lengths up to the bound Z, represented by enodes st/z for $z \in [Z]$. Constraint (8a) considers the joint contribution to mn-ebits with a specific path length z, from a pair of mk- and kn-ebits with total path length $z - \zeta_k$. This accounts for the fact that a concatenated mn-path has a total length of the mk-segment and the kn-segment, plus the length ζ_k of node k. Constraint (8b) specifies flow conservation at each intermediate pair of nodes mn with each specific path length value z. Constraint (8d) is the definition of $I(mn/z)$ that includes all generated ebits between mn with a specific length z from either elementary ebit generation or intermediate swapping, minus all ebits contributed to further swapping. Constraint (8e) defines $\Omega(mn/z)$ that includes all the ebits between mn with a specific length z, which will be swapped to build ebits between other node pairs. Proof of the following theorem is delegated to Appendix B.

Theorem 4: Given integer link/node lengths $\zeta_i > 0$ for $i \in N \cup L$, and an integer length bound Z, Program (8) computes the maximum expected EDR between s and t, with all ebits generated along paths satisfying the length bound Z. □

Proposition 1: Program (8) can be solved optimally, in time polynomial to the input size and Z. □

Proof: Program (8) is an LP with $O(|N|^3 Z^2)$ variables, and can be solved in time polynomial to $|N|$ and Z [57]. □

Algorithm 2 Approximate Testing Procedure TEST($\mathbf{Z}, \varepsilon$)

Input: Network G, accuracy ε, non-quantized length bound $\mathbf{Z}$

Output: Test result $\varsigma \in \{true, false\}$

$\theta \leftarrow (2|N| - 3)/(\varepsilon \mathbf{Z})$, and $Z \leftarrow \lfloor \theta \mathbf{Z} \rfloor + (2|N| - 3)$;

Solve Program (8) with $\{\zeta_i^\theta\}$ and Z, and get η_{st}^Z;

return *((Program (8) is feasible) AND ($\eta_{st}^Z \geq \Delta_{st}$)).*

B. Approximate Testing Procedure

Summary: Our second building block aims to test if a (real-valued) length bound $\mathbf{Z}$ admits a feasible solution that has an expected EDR higher than the EDR goal in OF-RED, with a bounded testing error. This is achieved by first quantizing the original (real-valued) lengths into integers with a carefully designed quantization factor, and then calling Program (8) to derive the optimal EDR and compare it to the EDR goal.

Program (8) runs in pseudo-polynomial time and can be used to check, given any length bound Z, if there is a feasible length-bounded eflow with expected EDR bound Δ_{st}. This testing is limited by 1) the requirement in Program (8) that all length values must be positive integers and 2) the pseudo-polynomial running time. We design an *approximate testing* procedure that simultaneously addresses these two issues. Specifically, by designing a proper quantization scheme to transform any real length value into a positive integer within a polynomial scale, we can limit the size of the resulting LP in Program (8), and bound the quantization error of the transformation.

To start, we define a quantization of the length values $\mathcal{Z} \triangleq \{\zeta_i \,|\, i \in L \cup N\}$ with a factor $\theta > 0$, where the **quantized length** is denoted by $\zeta_i^\theta = \lfloor \theta \cdot \zeta_i \rfloor + 1$, for $i \in N \cup L$. This transformation ensures that the resulting value ζ_i is always a positive integer, which satisfies the requirement of Program (8).

Let $\zeta^\theta(p)$ be the length of an arbitrary path p in G with quantization factor θ. We have the following lemmas whose proofs can be found in Appendix C and Appendix D, respectively:

Lemma 3: $\theta \cdot \zeta(p) \le \zeta^\theta(p) \le \lfloor \theta \cdot \zeta(p) \rfloor + (2|N| - 3)$. □

Based on Lemma 3, we design the approximate testing procedure in Algorithm 2. Given an accuracy parameter $\varepsilon > 0$ and a non-quantized length bound $\mathbf{Z}$, and define quantization factor θ and corresponding quantized length bound Z in Line 1, the algorithm returns a test result $\varsigma \in \{true, false\}$, which indicates whether the network admits a feasible *length-bounded eflow* with expected EDR no lower than Δ_{st}. Let $\mathbf{Z}^*$ be the non-quantized length of the optimal solution of OF-RED. Lemma 4 shows a numerical relationship between the input length bound $\mathbf{Z}$ and the optimal $\mathbf{Z}^*$ given the testing outcome:

Lemma 4: Given any $\varepsilon > 0$ and $\mathbf{Z} > 0$, we have

$$\textit{TEST}(\mathbf{Z}, \varepsilon) = true \;\Rightarrow\; \mathbf{Z}^* \le (1+\varepsilon)\cdot \mathbf{Z};$$
$$\textit{TEST}(\mathbf{Z}, \varepsilon) = false \;\Rightarrow\; \mathbf{Z}^* > \mathbf{Z}.$$

□

Remark: The choice of the factor θ in Line 1 ensures both a polynomial size and bounded quantization error. On one hand, it ensures the quantized length bound Z is polynomial to $|N|/\varepsilon$ regardless of the value of the original length bound $\mathbf{Z}$. On the other hand, utilizing the maximum path length in the network, it ensures that the testing result has an error of at most $(1+\varepsilon)$.

The testing procedure is designed to enable a bisection search for the minimum longest path length $\mathbf{Z}^*$, if a reasonable initial range $[\text{LB}, \text{UB}]$ of $\mathbf{Z}^*$ is given. By repeatedly testing if a length bound $\mathbf{Z} \in [\text{LB}, \text{UB}]$ is feasible or not, the search can multiplicatively reduce the search space, and return a close-to-optimal feasible length bound $\mathbf{Z}$ within time logarithmic to the size of the initial search space. Since the time complexity of the search depends on the size of the search space, we next seek to find a pair of lower bound LB and upper bound UB on the optimal $\mathbf{Z}^*$ that are reasonably close to each other.

C. Sorting and Trimming Algorithm

Summary: By utilizing the approximate testing in Sec. V-B, we wish to search a range of possible length bounds and test each bound with approximate testing to find an approximately optimal length that can satisfy the EDR goal of OF-RED. However, before doing that, we need first to find a suitable (polynomially bounded) range of length values, defined by a lower and an upper bound, within which the optimal length resides. We do this by a sorting and trimming algorithm below.

We design a sorting and trimming algorithm in Algorithm 3 to find an initial pair of bounds LB, UB on $\mathbf{Z}^*$, such that $\text{LB} \le \mathbf{Z}^* \le \text{UB}$. Algorithm 3 sorts all node/link lengths in descending order and then iteratively finds a *critical length* $\zeta_{[i-1]}$ such that $G_{[i-1]}$ still admits a feasible solution to Program (5) with while $G_{[i]}$ does not. Here, $G_{[i-1]}$ includes all links and nodes with lengths up to $\zeta_{[i-1]}$ and Program (5) becomes infeasible or $\eta_{st}^{[i-1]} < \Delta_{st}$ when testing $\zeta_{[i]}$. If either condition is met, Algorithm 3 stops and returns LB to $\zeta_{[i-1]}$, which is the last feasible solution length. For UB, considering $G_{[i-1]}$ with N nodes, the maximum path length would be $(2|N|-3)$ with at most $|N|-1$ links and $|N|-2$ intermediate nodes, Therefore, UB can be $(2|N|-3)\cdot\zeta_{[i-1]}$ for the feasible solution in $G_{[i-1]}$.

Algorithm 3 Finding Lower and Upper Bounds on $\mathbf{Z}^*$

Input: Network G
Output: Lower and upper bounds (LB, UB) on $\mathbf{Z}^*$
1 Sort node/link lengths in $\{\zeta_l \,|\, l \in L\} \cup \{\zeta_n \,|\, n \in N\}$ in descending order as $\mathcal{Z} = (\zeta_{[1]}, \zeta_{[2]}, \dots)$;
2 **for** $\zeta_{[i]} \in \mathcal{Z}$ *in sorted order* **do**
3 Construct graph $G_{[i]}$ by pruning all nodes and links with lengths greater than $\zeta_{[i]}$ in G;
4 Solve Program (5) on $G_{[i]}$ for $\eta_{st}^{[i]}$;
5 **if** *Infeasible or* $\eta_{st}^{[i]} < \Delta_{st}$ **then break**;
6 **return** $(\text{LB} = \zeta_{[i-1]}, \text{UB} = (2|N|-3)\zeta_{[i-1]})$.

The following lemma states the gap between the so-found LB and UB, whose proof is delegated to Appendix E.

Lemma 5: Algorithm 3 finds LB and UB such that LB $\le \mathbf{Z}^* \le$ UB, and UB/LB $\in O(|N|)$.

D. Two-Stage Bisection Search Algorithm

Summary: With the three building blocks designed above, the last building block carries out an efficient bisection search on the range [LB, UB] that contains the optimal length value and finds an approximately optimal length bound that can satisfy the EDR goal in OF-RED. This search must be designed carefully, as below, to ensure a polynomial time complexity.

After finding LB and UB with Algorithm 3, we can apply a bisection search on the range $[\text{LB}, \text{UB}]$ to find an approximator of $\mathbf{Z}^*$. Each time we define a bound $\mathbf{Z} = (\text{LB} + \text{UB})/2$, and call $\text{TEST}(\mathbf{Z}, \varepsilon)$. If $\text{TEST}(\mathbf{Z}, \varepsilon)$ outputs $true$, we narrow the gap by setting $\text{UB} \leftarrow (1+\varepsilon)\mathbf{Z}$; otherwise, we set $\text{LB} \leftarrow \mathbf{Z}$. To achieve the desired accuracy, it takes at least $O(\log(\text{UB} - \text{LB})) = O(\log(|N|\zeta_{[i-1]}))$ search iterations (where $\zeta_{[i-1]}$ is the critical length in Algorithm 3), each making a call to $\text{TEST}(\mathbf{Z}, \varepsilon)$ which solves an LP of size $O(|N|^3(|N|/\varepsilon)^2)$.

In Algorithm 4, we propose an improved *2-stage search algorithm*, which reduces the asymptotic search complexity and sizes of the LPs solved in most search iterations. In Stage-1 (Lines 2–5), a *multiplicative bisection* (bisection in the logarithmic scale) is done on $[\text{LB}, \text{UB}]$, where each time an $\varepsilon = 1$ is used in approximate testing. By Lemma 4, $\text{TEST}(\mathbf{Z}, 1)$ returning $false$ means $\mathbf{Z}^* > \mathbf{Z}$ and hence LB is increased to $\mathbf{Z}$; $\text{TEST}(\mathbf{Z}, 1)$ returning $true$ means $\mathbf{Z}^* \le (1+\varepsilon)\cdot\mathbf{Z} = 2\mathbf{Z}$ and hence UB is decreased to $2\mathbf{Z}$. Stage-1 ends when LB and UB are within a constant factor of each other, such as $\text{UB}/\text{LB} \le 4$.

In Stage-2, instead of doing bisection directly on $[\text{LB}, \text{UB}]$, we do a bisection on the quantized bounds $[Z_{\text{LB}}, Z_{\text{UB}}]$. We fix the quantization factor $\theta = (2|N|-3)/(\varepsilon\text{LB})$, and only vary the quantized path length bound Z. The main purpose of this construction is to utilize quantization to naturally reduce the number of search iterations to achieve the desired accuracy defined by ε. Since LB and UB are within a constant ratio of each other, the quantized length bound $Z_{\text{UB}} \in O(|N|/\varepsilon)$, and hence $O(\log(|N|/\varepsilon))$ search iterations are needed to search all integers between Z_{LB} and Z_{UB}. This makes the search complexity no longer related to the critical length $\zeta_{[i-1]}$ as in the naive bisection search. Let $\mathbf{Z}^\theta$ be the minimum longest path length for *quantized OF-RED (QOF-RED)* with θ.

Theorem 5 states our main result with proof in Appendix F.

Algorithm 4 2-Stage Bisection for Approximate OF-RED

Input: Network G, search accuracy parameter ε
Output: Eflow with maximum path length $\mathbf{Z}^+$

Call Algorithm 3 to find LB and UB on $\mathbf{Z}^*$;
while $\text{UB} > 4 \cdot \text{LB}$ **do** // Stage-1
  $\mathbf{Z} = \sqrt{(\text{UB} \cdot \text{LB})/2}$;
  if $\text{TEST}(\mathbf{Z}, 1) = false$ **then** $\text{LB} \leftarrow \mathbf{Z}$;
  else $\text{UB} \leftarrow 2 \cdot \mathbf{Z}$;
$\theta \leftarrow \frac{2|N|-3}{\varepsilon \text{LB}}$, $Z_{\text{LB}} \leftarrow \lfloor \theta \text{LB} \rfloor$, $Z_{\text{UB}} \leftarrow \lfloor \theta \text{UB} \rfloor + (2|N|-3)$;
while $Z_{\text{UB}} > Z_{\text{LB}} + 1$ **do** // Stage-2
  $Z \leftarrow \lfloor (Z_{\text{LB}} + Z_{\text{UB}})/2 \rfloor$;
  Solve Program (8) with θ and Z, and get η_{st}^Z;
  if *Program (8) is feasible AND* $\eta_{st}^Z \geq \Delta_{st}$ **then** $Z_{\text{UB}} \leftarrow Z$;
  else $Z_{\text{LB}} \leftarrow Z$;
return *last feasible solution with max path length* $\mathbf{Z}^+$

Theorem 5: Given accuracy parameter ε, Algorithm 4 finds a $(1+\varepsilon)$-approximation of the optimal OF-RED path length value $\mathbf{Z}^*$, within time polynomial to $|N|$ and $1/\varepsilon$. □

E. Discussions

Reducing running time: Despite being polynomial-time, Algorithm 4 still has high complexity due to solving the large-size LPs. There are several methods to reduce running time: 1) setting a loose ε; 2) applying heuristic quantization that works empirically; 3) developing heuristic algorithms to solve the quantized LP. We will examine effect of the first method in our evaluation. Considering that a quantum network is designed for long-term operations, the overhead of offline optimization can often be negligible. For instance, by spending minutes or hours to compute a high-EDR and high-fidelity entanglement distribution plan for a quantum key distribution (QKD) application [42], the plan could be executed and deliver largely improved performance over a period of weeks or months before offline maintenance/re-optimization is needed.

Entanglement distribution protocol: While the goal of our algorithm is mainly to 1) compute theoretical upper bounds on the achievable EDR and worst-case fidelity and 2) characterize the EDR-fidelity trade-off, we note that the computed eflow can be implemented by a data plane protocol in Appendix G. To achieve the theoretical EDR and fidelity, quantum memories are required to perform post-selection and storage before further swapping. In evaluation, we will use this protocol to characterize the EDR-fidelity trade-off in a simulated quantum network, and evaluate the performance of several state-of-the-art protocols with respect to the characterized trade-off.

Entanglement purification and error correction: This paper regards purification or quantum error correction (QEC) as an independent process from the entanglement distribution process. Both purification and QEC require consuming multiple/many additional ebits or qubits in order to get one high-quality ebit. This may significantly reduce the achievable EDR. For example, assume we have a stream of entanglements with EDR η. After one rounding of entanglement purification with success probability q_p [20], [40], the EDR will immediately drop to $\frac{\eta}{2} \cdot q_p$. More purification rounds degrade the EDR exponentially, and QEC is generally even more costly than purification. Both operations also require idealized quantum memories not only for storage but also for local quantum computation, which are far more complicated to design and implement. With the abstractions developed in this paper, we wish to explore incorporating purification and QEC into end-to-end modeling in our future work.

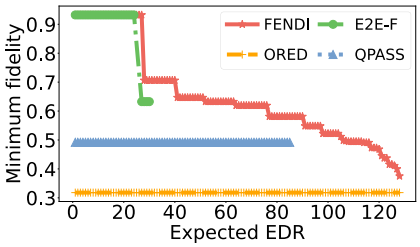


Fig. 6. The trade-off between worst-case fidelity and expected EDR for compared algorithms.

VI. Performance Evaluation

A. Evaluation Methodology

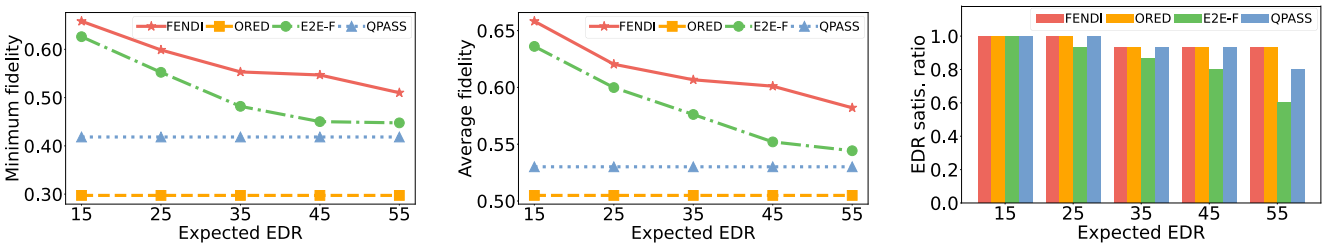
We developed a discrete-time quantum network simulator and carried out simulations on randomly generated topologies. We used random Waxman graphs [54] with parameters $\alpha = \beta = 0.8$. Each node or link had a success probability of 0.5 and 0.9, respectively, and fidelity was uniformly sampled from $[0.7, 0.95]$. Each link had a capacity uniformly sampled from $[26, 35]$. Parameters follow [64], except for the swapping success probability due to the limitation of the current BSM scheme with linear optics [3]. In each setting, we generated 5 graphs each with 15 nodes and 3 random SD pairs, except in Fig. 6 where we characterized the entire trade-off curve for one SD pair in a single graph. Results were averaged over all runs in the same setting to average out random noise.

Our time-slotted simulator is compatible with existing algorithms, though our data plane protocol (see Appendix G) does not require network-wide synchronization. Linear programs were solved by Gurobi [1]. Simulations were run on a Linux desktop with a 12-core 4GHz CPU and 256GB memory. In each simulation, we simulated entanglement generation, swapping and/or queuing for 1000 time slots based on the solution of algorithms. The following algorithms were compared:

- **FENDI:** Our FPTAS, with the solution executed with the post-selection-and-storage protocol in Appendix G.
- **ORED:** The fidelity-agnostic ORED algorithm, with a similar post-selection-and-storage protocol in [18].
- **E2E-F:** End-to-end fidelity-aware entanglement routing in [64], *without purification* for fair comparison.
- **QPASS:** Fidelity-agnostic entanglement routing in [48].

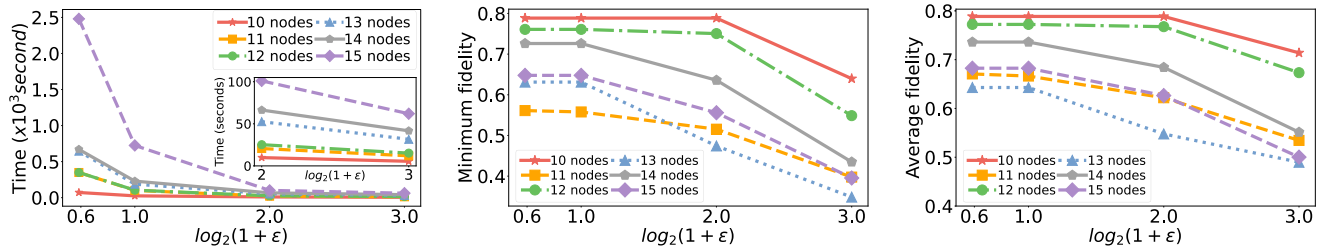
We set $\varepsilon = 0.5$ by default and the number of paths as 30 for QPASS and E2E-F. Since E2E-F and QPASS are *entanglement routing* algorithms for a bufferless quantum network, we adapted our simulator to discard all saved ebits in each slot.

The ***minimum fidelity*** and ***average fidelity*** measure the lowest and average fidelity values of all end-to-end entanglements. The ***EDR satisfaction ratio*** measures the fraction of simulation runs where the EDR bound is met. The ***running time*** measures the average time spent on running each *control plane* algorithm.



(a) Minimum end-to-end fidelity (b) Average end-to-end fidelity (c) EDR satisfaction ratio

Fig. 7. Comparison between FENDI and state-of-the-art algorithms.



(a) Running time (b) Minimum end-to-end fidelity (c) Average end-to-end fidelity

Fig. 8. Performance and running time of FENDI with varying ε and number of nodes.

B. Evaluation Results

1) Characterizing EDR-Fidelity Trade-off for Single SD Pair: We first investigate how FENDI can be used to characterize the EDR-fidelity trade-off curve for a single SD pair in a randomly generated 15-node graph, and the result is shown in Fig. 6. We applied the ϵ-constraint method [36], varying the expected EDR bound from 1 until the maximum value computed by ORED, and observed the maximum achievable worst-case fidelity given each expected EDR bound. A few key observations can be made: (i) Even in a 15-node network, there could be many (more than 20) paths between a pair of nodes, leading to many strongly Pareto optimal points in the frontier. (ii) FENDI was able to (approximately) characterize the entire frontier from one direction, presenting many different trade-off options for entanglement distribution—each could be implemented by the post-selection-and-storage protocol. (iii) None of the existing algorithms could characterize the trade-off well. Specifically, ORED could achieve the highest expected EDR, but the lowest fidelity due to using all possible paths in the network to maximize EDR. QPASS sought to maximize EDR, but could achieve neither the maximum EDR nor the highest fidelity. Both these methods are fidelity-agnostic, and hence could only optimize for one dimension but not the trade-off. The fidelity-aware E2E-F was able to trade-off EDR with fidelity, but only for a very small portion of the entire trade-off curve. The inefficacy comes from two aspects: 1) not being able to utilize all paths to achieve an arbitrary trade-off, and 2) not being able to provide guarantee for expected EDR. In fact, most (if not all) existing algorithms are designed to optimize for a single point in the area bounded by FENDI's trade-off curve, and mostly achieve a suboptimal point strictly within the boundary.

2) Achievable Fidelity Versus EDR: Fig. 7(a)–(b) shows the end-to-end worst-case and average fidelity with different expected EDRs in randomly generated networks. From Figs. 7(a)–(b), FENDI achieved the highest fidelity compared to all other algorithms. For any specific expected EDR bound, the two fidelity-aware algorithms (FENDI and E2E-F) achieved significantly higher fidelity than the fidelity-agnostic ones (ORED and QPASS), demonstrating *the crucial need for fidelity awareness in quantum networking*. With increasing EDR bounds, fidelity was sacrificed to meet the EDR requirement when lower-fidelity paths were utilized. Though both aimed to approach the optimal fidelity-EDR trade-off, the fidelity gap between FENDI and E2E-F generally increased with higher EDR bounds, demonstrating *importance of our approximation guarantee*. Note that for many tasks such as entanglement purification [4], entanglements are regarded as non-usable when fidelity drops below 0.5. Fig. 7(a) shows that to ensure minimum fidelity over 0.5, our algorithm could achieve significantly higher expected EDR, even compared to existing fidelity-aware algorithm such as E2E-F.

3) Capability to Satisfy EDR Requirement: From Fig. 7(c), FENDI achieved EDR satisfaction ratios on par with ORED. This is because both algorithms explore the same EDR feasibility region, and differ only by fidelity of paths (pflows) to meet a given expected EDR bound. Both FENDI and ORED achieved higher EDR satisfaction ratio than QPASS and E2E-F, even though E2E-F achieved similar (but still lower) fidelity compared to FENDI and higher fidelity than ORED. There are two reasons: 1) FENDI and ORED are *optimal* in terms of whether an expected EDR bound can be satisfied while E2E-F and QPASS have no such guarantee; 2) a buffered network can achieve higher long-term EDR than a bufferless network by storing instead of discarding unused intermediate ebits.

4) Performance Versus Running Time of FPTAS: Fig. 8 shows the trade-off between performance and running time for FENDI, with varying number of nodes and accuracy parameter ε. Note that despite ε, FENDI always achieved the same EDR satisfaction ratio as the same feasibility region of the problem was explored, and thus we omit the figure showing the EDR satisfaction ratio. From Fig. 8(a), the running time increased with number of nodes and decreased with ε. From Figs. 8(b) and 8(c), increasing ε led to fidelity reduction, matching our theoretical analysis. However, with a relatively loose ε, such as when $\varepsilon = 1$, the achieved fidelity was on par with when ε was set to a tight value such as 0.5. This shows that the theoretical

guarantee tends to be over-conservative in practice, and *it is reasonable to set a loose ε to achieve high time efficiency with reasonable performance*. The correlation between number of nodes and fidelity values of FENDI was weak. This could be because, on the one hand, a larger graph with more nodes could lead to more paths between each SD pair and hence increase fidelity; on the other hand, a larger graph also means it was more likely that two randomly picked nodes were further away in the graph, leading to degraded fidelity over long paths.

VII. Conclusion

In this paper, we studied how to characterize the entanglement distribution rate and fidelity trade-off in a general-topology quantum network with theoretical guarantee. We derived an end-to-end fidelity model with worst-case (isotropic) noise. We then formulated the HF-RED problem for maximizing the achievable fidelity under an expected EDR bound (modeled with an optimal entanglement flow abstraction), and proved its NP-hardness. With a novel decomposition theorem, we developed a *fully polynomial-time approximation scheme (FPTAS)* for the problem called FENDI. We also developed a discrete-time quantum network simulator for evaluation. Simulation results showed the superior performance of FENDI, compared to existing entanglement routing and distribution algorithms.

Appendix

A. Proof of Theorem 2

Proof: We find an induced graph $\mathcal{G}' \subseteq \mathcal{G}$ in which each enode $mn \in \mathcal{G}'$ has either $g_{mn} > 0$, or there is exactly one $k \in N$ such that $f_{mn}^{mk} = f_{mn}^{kn} > 0$. Such a subgraph must exist due to the constraint of $I(mn) - \Omega(mn) = 0$ for every $mn \neq st$, and $\eta_{st} > 0$ for the eflow. We then use Algorithm 1 to compute ebit generation ratios of the pflow in $\mathcal{G}'$. Let η^* be the maximally acceptable EDR of this pflow where $\eta^* \triangleq \min(\{f_{mn}^{mk}/\bar{f}_{mn}^{mk} \,|\, m,n,k \in N, \bar{f}_{mn}^{mk} > 0\} \cup \{g_{mn}/\bar{g}_{mn} \,|\, m,n \in N, \bar{g}_{mn} > 0\})$. We update the original eflow by deducting each variable by η^* times the corresponding ebit generation ratio in the pflow. Continue this process until $\eta_{st} = 0$, and we have a set of pflows with the sum of EDRs η_{st}. In the above process, either at least one g_{mn}, or at least one pair of $\{f_{mn}^{mk}, f_{mn}^{kn}\}$ variables with some k, becomes 0 after updating each pflow. Since there are in total $O(N^3)$ variables, this decomposition results in at most $O(N^3)$ pflows. □

B. Proof of Theorem 4

Proof: We call a mn/z by *enode* mn *at level* z. We first examine path length feasibility, *i.e.*, ebits generated between mn at level z has path length of exactly z. For any physical link $mn \in L$, the first term in Eq. (8d) ensures that g_{mn} only contributes to $I(mn/z)$ when $z = \zeta_{(mn)}$, *i.e.*, elementary ebits along mn are only counted at level $\zeta_{(mn)}$. Then, for any mn/z satisfying $f_{mn/z}^{mk/z_1} > 0$ and $f_{mn/z}^{kn/z_2} > 0$, we can see if ebits at mk/z_1 have path length of exactly z_1 and ebits at kn/z_2 have path length of exactly z_2, then ebits generated at mn/z by swapping at k exactly have path length of $z = z_1 + z_2 + \zeta_k^\theta$. By induction, any generated ebit at level z has path length of exactly z. Since there are at most Z levels, all ebits generated between st have path lengths bounded by Z.

Next we prove optimality of Program (8), by showing that every solution to Program (8) with objective value η_{st}^Z is a solution to HF-RED with EDR bound $\Delta_{st} = \eta_{st}^Z$ and fidelity bound $\Upsilon_{st} = \frac{1}{4} \cdot \left(1 + 3\, e^{-Z}\right)$, and vice versa. A feasible length-bounded eflow to Program (8) indicates a feasible eflow to Program (5), by summing up f variables and $I(\cdot)$ function values over all possible z. Combined with path length feasibility, the length-bounded eflow maintains worst-case fidelity above the fidelity threshold Υ_{st} and EDR bound Δ_{st} in the HF-RED problem. Now, we represent a feasible length-bounded eflow by a set of pflows with induced graphs $\{\mathcal{G}\}$ with $\{\eta_{\mathcal{G}}\}$. Each $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ would represent a path $p_{\mathcal{G}} \in G$ with path length bounded by Z. We can construct a feasible solution to Program (8) given each $\mathcal{G}$. For each enode $mn \in \mathcal{V}$, let ζ_{mn} be the length of the path segment in G between m and n (which can be computed for each mn in linear time). For each enode mn that has no in-coming link, we set $g_{mn} = \bar{g}_{mn} \cdot \eta_{\mathcal{G}}$. Then, for each $(mk, mn) \in \mathcal{E}$, we set $f_{mn/\zeta_{mn}}^{mk/\zeta_{mk}} = f_{mn/\zeta_{mn}}^{kn/\zeta_{kn}} = \bar{f}_{mn}^{mk} \cdot \eta_{\mathcal{G}}$. It can be checked that the constructed solution is feasible to Program (8) based on how $\{\bar{g}_{mn}, \bar{f}_{mn}^{mk}, \bar{f}_{mn}^{kn}\}$ are computed, how $\mathcal{G}$ is defined, and that each $\mathcal{G}$ represents a path with length bounded by Z. Summing up so-constructed solutions for all of $\{\mathcal{G}\}$ and $\{\eta_{\mathcal{G}}\}$, we get a feasible solution to Program (8), with the same objective value $\eta_{st}^Z = \sum_{\mathcal{G}} \eta_{\mathcal{G}}$. It follows that Program (8) outputs the maximum expected EDR among all feasible eflows satisfying the path length bound of Z. □

C. Proof of Lemma 3

Proof: The left side is trivial due to how lengths are quantized. The right side is because 1) each entanglement path in G has at most $|N|-1$ links and $|N|-2$ intermediate nodes whose lengths are counted (excluding source and destination), and 2) $\zeta^\theta(p)$ is an integer value due to quantization (and hence the floor over $\theta \cdot \zeta(p)$ on the right side). □

D. Proof of Lemma 4

Proof: If TEST$(\mathbf{Z}, \varepsilon) = true$, we have a feasible length-bounded eflow with maximum EDR $\eta_{st}^Z \geq \Delta_{st}$ and all paths satisfying Z, which means a feasible solution to OF-RED is bounded by Δ_{st}. Let p be the maximum-length path in the solution w.r.t. the original lengths $\mathcal{Z}$. Following Lemma 3, we have:

$$\zeta(p) \leq \zeta^\theta(p)/\theta \leq Z/\theta \leq (1+\varepsilon)\mathbf{Z}.$$

Since the solution is feasible to OF-RED, its maximum (non-quantized) path length is an upper bound on $\mathbf{Z}^*$, and hence we have $\mathbf{Z}^* \leq (1+\varepsilon)\mathbf{Z}$. This proves the first statement.

To prove the second statement, as long as there is a feasible OF-RED solution that has maximum path length bounded by $\mathbf{Z}$, then TEST$(\mathbf{Z}, \varepsilon)$ must return $true$. Consider such a solution for every path p satisfying $\zeta(p) \leq \mathbf{Z}$. By Lemma 3, we have:

$$\zeta^\theta(p) \leq \theta \cdot \zeta(p) + (2|N| - 3) \leq (2|N| - 3)/\varepsilon + (2|N| - 3).$$

Since $\zeta^\theta(p)$ must be an integer, this implies $\zeta^\theta(p) \leq \lfloor (2|N| - 3)/\varepsilon \rfloor + (2|N| - 3) = \lfloor \theta \mathbf{Z} \rfloor + (2|N| - 3) = Z$. By Theorem 4, this solution can be decomposed into a set of pflows with maximum quantized path length $\zeta^\theta(p)$ and satisfying $\eta_{st}^Z \geq \Delta_{st}$. In this case, TEST$(\mathbf{Z}, \varepsilon)$ must return $true$. Otherwise, it indicates there is no such feasible solution. □

E. Proof of Lemma 5

Proof: Algorithm 3 identifies a critical length $\zeta_{[i-1]}$ beyond which no feasible solution exists for Program 3 while maintaining feasibility for shorter lengths. This means at least one node/link length no less than $\zeta_{[i-1]}$ is needed to satisfy the EDR bound of Δ_{st}. Consequently, the optimal $\mathbf{Z}^*$ must be at least $\zeta_{[i-1]}$ as a lower bound. For the upper bound, since there is a feasible solution in $G_{[i-1]}$, and each path can have at most $|N|-1$ links and $|N|-2$ intermediate nodes, the feasible solution has a maximum path length of $(2|N|-3)\cdot\zeta_{[i-1]}$ as all nodes and links in $G_{[i-1]}$ have lengths at most $\zeta_{[i-1]}$. Therefore, the gap between the above pair of bounds is a multiplicative factor of $\text{UB}/\text{LB}=2|N|-3\in O(|N|)$. □

F. Proof of Theorem 5

Lemma 6: $\lfloor\theta\text{LB}\rfloor\le\mathbf{Z}^\theta\le\lfloor\theta\text{UB}\rfloor+(2|N|-3)$. □

Lemma 7: $\mathbf{Z}^\theta\le\theta\cdot(1+\varepsilon)\cdot\mathbf{Z}^*$. □

Proof: Note that a feasible solution to OF-RED indicates a feasible solution to QOF-RED, and vice versa. Given the optimal solution to original OF-RED with objective $\mathbf{Z}^*$, let p be its longest entanglement path such that $\zeta(p)=\mathbf{Z}^*$, and let p_θ be its longest entanglement path with quantization. By Lemma 3, $\zeta^\theta(p_\theta)\le\lfloor\theta\zeta(p_\theta)\rfloor+(2|N|-3)\le\lfloor\theta\zeta(p)\rfloor+(2|N|-3)\le\lfloor\theta\text{UB}\rfloor+(2|N|-3)$. This proves the right-hand side of Lemma 6, as $\mathbf{Z}^\theta$ is optimal and hence $\mathbf{Z}^\theta\le\zeta^\theta(p_\theta)$. Further, since $\zeta(p)=\mathbf{Z}^*$, we have $\zeta^\theta(p_\theta)\le\theta\zeta(p)+(2|N|-3)=\theta(\mathbf{Z}^*+(2|N|-3)/\theta)=\theta(\mathbf{Z}^*+\varepsilon\text{LB})\le\theta\cdot(1+\varepsilon)\cdot\mathbf{Z}^*$, and hence $\mathbf{Z}^\theta\le\zeta^\theta(p_\theta)\le\theta\cdot(1+\varepsilon)\cdot\mathbf{Z}^*$.

Now consider the optimal solution to QOF-RED, and let p'_θ and p' be its longest entanglement path with and without quantization. Since this solution is also feasible to OF-RED, we have $\theta\text{LB}\le\theta\mathbf{Z}^*\le\theta\zeta(p')$. By Lemma 3, we then have $\theta\zeta(p')\le\zeta^\theta(p')\le\zeta^\theta(p'_\theta)=\mathbf{Z}^\theta$. Hence $\mathbf{Z}^\theta\ge\lfloor\theta\text{LB}\rfloor$. □

The lemmas above show this quantized bisection is as effective as the bisection search on the original bounds $[\text{LB},\text{UB}]$. We then provide the proof of Theorem 5 as follows.

Proof: The approximation ratio directly comes from Lemma 7. Let $T(x)$ be the time for solving an LP with x variables. First, Algorithm 3 finds $[\text{LB},\text{UB}]$ on $\mathbf{Z}^*$ in up to $|\mathcal{Z}|=|N|+|L|$ iterations, each solving Program (5) with $O(|N|^3)$ variables in $O(T(|N|^3))$ time. For Stage-1 bisection of Algorithm 4, let $\pi_{[j]}$ be the ratio UB/LB after the j-th iteration. Initially $\pi_{[0]}=2|N|-3$ due to $[\text{LB},\text{UB}]$ bound by Algorithm 3. After each iteration j, $\pi_{[j]}=\sqrt{2\pi_{[j-1]}}$ based on how $\mathbf{Z}$ is computed. Let J be index of the last iteration, and apply the above recursively, then we have $\pi_{[J]}=2^{1/2+1/4+\cdots+1/2^J}\cdot\pi_{[0]}^{1/2^J}\le 2\cdot\pi_{[0]}^{1/2^J}=2\cdot(2|N|-3)^{1/2^J}$. As $\pi_{[J]}\le 4$ when Stage-1 ends, the total number of iterations is $O(\log\log|N|)$. Each iteration solves Program (8) with $\varepsilon=1$, and hence $Z\in O(|N|)$, resulting in $O(|N|^3Z^2)=O(|N|^5)$ variables. Thus each iteration takes $O(T(|N|^5))$ time. For Stage-2, the bisection is done on up to $Z_{\text{UB}}\in O(\frac{|N|}{\varepsilon})$ integers, with up to $O(\log\frac{|N|}{\varepsilon})$ search iterations. Each iteration solves Program (8) with $O(|N|^3Z_{\text{UB}}^2)=O(\frac{|N|^5}{\varepsilon^2})$ variables, and thus takes $O(T(\frac{|N|^5}{\varepsilon^2}))$ time. Summing up the above, the overall time complexity is $O(T(|N|^3)\cdot(|N|+|L|)+T(|N|^5)\cdot\log\log|N|+T(|N|^5/\varepsilon^2)\cdot\log\frac{|N|}{\varepsilon})$. Since an LP can be solved in polynomial time [57], the above time is polynomial to $|N|$ and $1/\varepsilon$. □

G. Data Plane Protocol for FENDI

Given a solution output by a central quantum network controller running Algorithm 4, we design an extension of the protocol in [18] to achieve the expected EDR and guarantee that all generated ebits have end-to-end fidelity of at least Υ_{st}.

Specifically, after the computation, the factor θ and the final quantized path length bound Z_{UB} are distributed to each repeater. For every enode mn, it maintains ***input buffers*** $\mathcal{E}_{mn/z}$ to store the ebits generated between mn with a specific range of fidelity represented by a quantized length z and ***output buffers*** $\mathcal{D}^{mn/z}_{mk/z'}$ for every $k\neq m,n$ to store the ebits that will be contributed to generate ebits between other pairs. Note that the number and sizes of buffers at each node may be dynamically adjusted by allocating the available quantum memories.

Each link $mn\in E$ will continuously generate $c_{mn}\cdot g_{mn}$ elementary ebits. Once successfully generated, these ebits are added to the buffer $\mathcal{E}_{mn/z}$ where $z=\zeta^\theta_{mn}=\lfloor-\log(W_{mn})\theta\rfloor+1$. Simultaneously, whenever an ebit is added to $\mathcal{E}_{mn/z}$ for any z, the two ends will jointly toss a random coin and move the ebit from $\mathcal{E}_{mn/z}$ to $\mathcal{D}^{mn/z}_{mk/z'}$ or $\mathcal{D}^{mn/z}_{kn/z'}$ with probabilities:

$$\Pr[\text{move to }\mathcal{D}^{mn/z}_{mk/z'}]=\frac{f^{mn/z}_{mk/z'}}{\sum_{z''}\sum_k(f^{mn/z}_{mk/z''}+f^{mn/z}_{kn/z''})};$$

$$\Pr[\text{move to }\mathcal{D}^{mn/z}_{kn/z'}]=\frac{f^{mn/z}_{kn/z'}}{\sum_{z''}\sum_k(f^{mn/z}_{mk/z''}+f^{mn/z}_{kn/z''})}.$$

Finally, each node k will check if for any mn, it satisfies

1) $z_1+z_2+\zeta^\theta_k=z_3$;
2) $f^{mk/z_1}_{mn/z_3}=f^{kn/z_2}_{mn/z_3}>0$; and
3) $\mathcal{D}^{mk/z_1}_{mn/z_3}\neq\emptyset$, and $\mathcal{D}^{kn/z_2}_{mn/z_3}\neq\emptyset$.

For each such a case, node k locally performs swapping between each pair of ebits in $\mathcal{D}^{mk/z_1}_{mn/z_3}$ and $\mathcal{D}^{kn/z_2}_{mn/z_3}$ respectively. Upon success, the ebit will then be added to $\mathcal{E}_{mn/z_3}$ by m and n. The source and destination will keep all ebits received in $\mathcal{E}_{st/z}$ for any z. All the above processes can be parallel and asynchronous. The strong network-wide synchronization requirement in traditional time-slotted entanglement routing protocols is thus relaxed. By an induction proof similar to the one in [18] which we omit due to page limit, this protocol is guaranteed to achieve a long-term EDR of at least Δ_{st} and an end-to-end fidelity of at least Υ_{st} output by the algorithm.

Remark: One implicit assumption not mentioned in [18] is that the proposed protocol requires perfect quantum memories to provide the guaranteed fidelity, and sufficiently large memories to achieve the full expected EDR. These assumptions are somewhat unrealistic under the current technologies. Hence, the computed EDR and fidelity both serve as upper bounds on the actual values that can be achieved by near-term devices. Though it is fairly well agreed that large-scale long-lived quantum memories will be an integral part of quantum networks in the future, especially with recent breakthroughs in optical memory devices with more than 1-hour coherence time [35].

We believe even establishing (tight) bounds on the achievable EDR and fidelity is still very useful for near-term quantum network design, such as when comparing different topologies and parameters or practical protocol design with these theoretical upper bounds. Furthermore, we have also preliminarily

tested the performance of the buffered protocol above with limited buffer space and found that it can still maintain an EDR close to the theoretical bound with a relatively small buffer size—such as equal to the capacity of each link. While out of the scope of the current paper which focuses on computing the theoretical bounds, we believe smart buffer management can further reduce the buffer size and increase achievable EDR and fidelity, which we will explore in our future work.

REFERENCES

[1] *Gurobi Optimizer*. Accessed: Jul. 25, 2022. [Online]. Available: http://www.gurobi.com/products/gurobi-optimizer

[2] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms and Applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[3] M. J. Bayerbach, S. E. D'Aurelio, P. van Loock, and S. Barz, "Bell-state measurement exceeding 50% success probability with linear optics," *Sci. Adv.*, vol. 9, no. 32, Aug. 2023, Art. no. eadf4080.

[4] C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher, "Concentrating partial entanglement by local operations," *Phys. Rev. A, Gen. Phys.*, vol. 53, no. 4, pp. 2046–2052, Apr. 1996.

[5] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theor. Comput. Sci.*, vol. 560, pp. 7–11, Dec. 2014.

[6] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, "Mixed-state entanglement and quantum error correction," *Phys. Rev. A, Gen. Phys.*, vol. 54, no. 5, pp. 3824–3851, Nov. 1996.

[7] K. A. G. Bonsma-Fisher et al., "Fiber-integrated quantum memory for telecom light," *Phys. Rev. A, Gen. Phys.*, vol. 108, no. 1, Jul. 2023, Art. no. 012606.

[8] C. E. Bradley et al., "Robust quantum-network memory based on spin qubits in isotopically engineered diamond," *NPJ Quantum Inf.*, vol. 8, no. 1, p. 122, Oct. 2022.

[9] M. Caleffi, A. S. Cacciapuoti, and G. Bianchi, "Quantum internet: From communication to distributed computing!" in *Proc. ACM NANOCOM*, 2018, pp. 1–4.

[10] J. Calsamiglia and N. Lütkenhaus, "Maximum efficiency of a linear-optical bell-state analyzer," *Appl. Phys. B, Lasers Opt.*, vol. 72, no. 1, pp. 67–71, Jan. 2001.

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

[12] A. Chang and G. Xue, "Order matters: On the impact of swapping order on an entanglement path in a quantum network," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2022, pp. 1–6.

[13] L. Chen et al., "Q-DDCA: Decentralized dynamic congestion avoid routing in large-scale quantum networks," *IEEE/ACM Trans. Netw.*, vol. 32, no. 1, pp. 368–381, Feb. 2024, doi: 10.1109/TNET.2023.3285093.

[14] L. Chen et al., "A heuristic remote entanglement distribution algorithm on memory-limited quantum paths," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7491–7504, Nov. 2022.

[15] C. Cicconetti, M. Conti, and A. Passarella, "Resource allocation in quantum networks for distributed quantum computing," 2022, *arXiv:2203.05844*.

[16] T. Coopmans, S. Brand, and D. Elkouss, "Improved analytical bounds on delivery times of long-distance entanglement," *Phys. Rev. A, Gen. Phys.*, vol. 105, no. 1, Jan. 2022, Art. no. 012608.

[17] A. Dahlberg et al., "A link layer protocol for quantum networks," in *Proc. ACM SIGCOMM*, 2019, pp. 159–173.

[18] W. Dai, T. Peng, and M. Z. Win, "Optimal protocols for remote entanglement distribution," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2020, pp. 1014–1019.

[19] W. Dai, T. Peng, and M. Z. Win, "Optimal remote entanglement distribution," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 3, pp. 540–556, Mar. 2020.

[20] S. Das, M. S. Rahman, and M. Majumdar, "Design of a quantum repeater using quantum circuits and benchmarking its performance on an IBM quantum computer," *Quantum Inf. Process.*, vol. 20, no. 7, pp. 1–17, Jul. 2021.

[21] O. Davidson, O. Yogev, E. Poem, and O. Firstenberg, "Fast, noise-free atomic optical memory with 35-percent end-to-end efficiency," *Commun. Phys.*, vol. 6, no. 1, p. 131, Jun. 2023.

[22] W. Dür, H.-J. Briegel, J. I. Cirac, and P. Zoller, "Quantum repeaters based on entanglement purification," *Phys. Rev. A, Gen. Phys.*, vol. 59, no. 1, pp. 169–181, Jan. 1999.

[23] C. Elliott, "Building the quantum network," *New J. Phys.*, vol. 4, no. 1, p. 46, 2002.

[24] F. Ewert and P. van Loock, "3/4-efficient bell measurement with passive linear optics and unentangled ancillae," *Phys. Rev. Lett.*, vol. 113, no. 14, Sep. 2014, Art. no. 140403.

[25] R. P. Feynman, "Simulating physics with computers," *Int. J. Theor. Phys.*, vol. 21, nos. 6–7, pp. 467–488, Jun. 1982.

[26] W. P. Grice, "Arbitrarily complete bell-state measurement using only linear optical elements," *Phys. Rev. A, Gen. Phys.*, vol. 84, no. 4, Oct. 2011, Art. no. 042331.

[27] R. Hassin, "Approximation schemes for the restricted shortest path problem," *Math. Oper. Res.*, vol. 17, no. 1, pp. 36–42, Feb. 1992.

[28] T. Kilmer and S. Guha, "Boosting linear-optical bell measurement success probability with predetection squeezing and imperfect photon-number-resolving detectors," *Phys. Rev. A, Gen. Phys.*, vol. 99, no. 3, Mar. 2019, Art. no. 032302.

[29] M. Koashi and N. Imoto, "No-cloning theorem of entangled states," *Phys. Rev. Lett.*, vol. 81, no. 19, pp. 4264–4267, Nov. 1998.

[30] T. Korkmaz and M. Krunz, "Multi-constrained optimal path selection," in *Proc. IEEE INFOCOM Conf. Comput. Commun. 20th Annu. Joint Conf. IEEE Comput. Commun. Soc.*, vol. 2, May 2001, pp. 834–843.

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

[32] Y. Lee, E. Bersin, A. Dahlberg, S. Wehner, and D. Englund, "A quantum router architecture for high-fidelity entanglement flows in quantum networks," *NPJ Quantum Inf.*, vol. 8, no. 1, p. 75, Jun. 2022.

[33] J. Li, Q. Jia, K. Xue, D. S. L. Wei, and N. Yu, "A connection-oriented entanglement distribution design in quantum networks," *IEEE Trans. Quantum Eng.*, vol. 3, pp. 1–13, 2022.

[34] J. Li et al., "Fidelity-guaranteed entanglement routing in quantum networks," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 6748–6763, Oct. 2022.

[35] Y. Ma, Y.-Z. Ma, Z.-Q. Zhou, C.-F. Li, and G.-C. Guo, "One-hour coherent optical storage in an atomic frequency comb memory," *Nature Commun.*, vol. 12, no. 1, pp. 1–6, Apr. 2021.

[36] G. Mavrotas, "Effective implementation of the ε-constraint method in multi-objective mathematical programming problems," *Appl. Math. Comput.*, vol. 213, no. 2, pp. 455–465, Jul. 2009.

[37] K. Miettinen, *Nonlinear Multiobjective Optimization*. Cham, Switzerland: Springer, 1999.

[38] S. Misra, G. Xue, and D. Yang, "Polynomial time approximations for multi-path routing with bandwidth and delay constraints," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 558–566.

[39] S. Muralidharan, L. Li, J. Kim, N. Lütkenhaus, M. D. Lukin, and L. Jiang, "Optimal architectures for long distance quantum communication," *Sci. Rep.*, vol. 6, no. 1, p. 20463, Feb. 2016.

[40] J.-W. Pan, C. Simon, C. Brukner, and A. Zeilinger, "Entanglement purification for quantum communication," *Nature*, vol. 410, no. 6832, pp. 1067–1070, Apr. 2001.

[41] M. Pant et al., "Routing entanglement in the quantum internet," *NPJ Quantum Inf.*, vol. 5, no. 1, pp. 1–9, Mar. 2019.

[42] M. Peev et al., "The secoqc quantum key distribution network in Vienna," *New J. Phys.*, vol. 11, no. 7, 2009, Art. no. 075001.

[43] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, "Fundamental limits of repeaterless quantum communications," *Nature Commun.*, vol. 8, no. 1, p. 15043, Apr. 2017.

[44] S. Pouryousef, N. K. Panigrahy, and D. Towsley, "A quantum overlay network for efficient entanglement distribution," 2022, *arXiv:2212.01694*.

[45] P. Promponas, V. Valls, S. Guha, and L. Tassiulas, "Maximizing entanglement rates via efficient memory management in flexible quantum switches," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 7, pp. 1749–1762, Jul. 2024.

[46] M. Sasaki et al., "Field test of quantum key distribution in the Tokyo QKD Network," *Opt. Exp.*, vol. 19, no. 11, pp. 10387–10409, 2011.

[47] E. Schoute, L. Mancinska, T. Islam, I. Kerenidis, and S. Wehner, "Shortcuts to quantum network routing," 2016, *arXiv:1610.05238*.

[48] S. Shi and C. Qian, "Concurrent entanglement routing for quantum networks: Model and designs," in *Proc. ACM SIGCOMM*, 2020, pp. 62–75.

[49] A. Singh, K. Dev, H. Siljak, H. D. Joshi, and M. Magarini, "Quantum internet—Applications, functionalities, enabling technologies, challenges, and research directions," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2218–2247, 4th Quart., 2021.

[50] L. J. Stephenson et al., "High-rate, high-fidelity entanglement of qubits across an elementary quantum network," *Phys. Rev. Lett.*, vol. 124, no. 11, Mar. 2020, Art. no. 110501.

[51] R. V. Meter and J. Touch, "Designing quantum repeater networks," *IEEE Commun. Mag.*, vol. 51, no. 8, pp. 64–71, Aug. 2013.

[52] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the stochastic analysis of a quantum entanglement distribution switch," *IEEE Trans. Quantum Eng.*, vol. 2, pp. 1–16, 2021.

[53] M. Victora, S. Krastanov, A. S. de la Cerda, S. Willis, and P. Narang, "Purification and entanglement routing on quantum networks," 2020, *arXiv:2011.11644*.

[54] B. M. Waxman, "Routing of multipoint connections," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 9, pp. 1617–1622, Jan. 1988.

[55] Y. Xia, W. Li, W. Clark, D. Hart, Q. Zhuang, and Z. Zhang, "Demonstration of a reconfigurable entangled radio-frequency photonic sensor network," *Phys. Rev. Lett.*, vol. 124, no. 15, Apr. 2020, Art. no. 150502.

[56] L. Yang, Y. Zhao, L. Huang, and C. Qiao, "Asynchronous entanglement provisioning and routing for distributed quantum computing," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2023, pp. 1–10.

[57] Y. Ye and P. M. Pardalos, "A class of linear complementarity problems solvable in polynomial time," *Linear Algebra Appl.*, vol. 152, pp. 3–17, Jul. 1991.

[58] J. Yin et al., "Satellite-based entanglement distribution over 1200 kilometers," *Science*, vol. 356, no. 6343, pp. 1140–1144, Jun. 2017.

[59] R. Yu, G. Xue, and X. Zhang, "QoS-aware and reliable traffic steering for service function chaining in mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2522–2531, Nov. 2017.

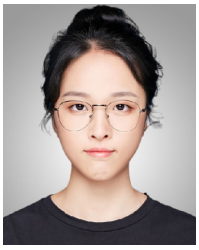
[60] Y. Zeng, J. Zhang, J. Liu, Z. Liu, and Y. Yang, "Multi-entanglement routing design over quantum networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2022, pp. 510–519.

[61] C. Zhan, H. Gupta, and M. Hillery, "Optimizing initial state of detector sensors in quantum sensor networks," *ACM Trans. Quantum Comput.*, vol. 5, no. 2, pp. 1–25, Jun. 2024.

[62] Y. Zhao and C. Qiao, "Redundant entanglement provisioning and selection for throughput maximization in quantum networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2021, pp. 1–10.

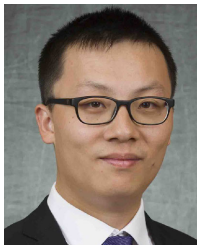
[63] Y. Zhao, Y. Wang, E. Wang, H. Xu, L. Huang, and C. Qiao, "An asynchronous transport protocol for quantum data networks," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 7, pp. 1885–1899, Jul. 2024.

[64] Y. Zhao, G. Zhao, and C. Qiao, "E2E fidelity aware routing and purification for throughput maximization in quantum networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2022, pp. 480–489.



Huayue Gu (Graduate Student Member, IEEE) received the B.E. degree in computer science from Nanjing University of Posts and Telecommunications, Jiangsu, China, in 2019, and the M.S. degree in computer science from the University of California at Riverside, CA, USA, in 2021. She is currently pursuing the Ph.D. degree with the Computer Science Department, North Carolina State University. Her research interests include quantum networking, quantum communication, and data analytics.

Zhouyu Li (Graduate Student Member, IEEE) received the B.E. degree from Central South University, Changsha, China, in 2019, and the M.S. degree from Georgia Institute of Technology, Atlanta, USA, in 2020. He is currently the Ph.D. degree in computer science with North Carolina State University. His research interests include privacy, cloud/edge computing, and network routing.



Ruozhou Yu (Senior Member, IEEE) received the Ph.D. degree in computer science from Arizona State University, USA, in 2019. He is currently an Assistant Professor of computer science with NC State University, USA. His interests include quantum networking, edge computing, algorithms and optimization, distributed learning, and security and privacy. He received the NSF CAREER Award in 2021. He has served on the organizing committees of IEEE INFOCOM from 2022 to 2025 and IEEE IPCCC from 2020 to 2024, the TPC Track Chair for IEEE ICCCN 2023, and the TPC Member of IEEE INFOCOM from 2020 to 2025 and ACM Mobihoc in 2023 to 2024. He is an Area Editor for *Computer Networks* (Elsevier).

Xiaojian Wang (Graduate Student Member, IEEE) received the B.E. degree from Taiyuan University of Technology, China, in 2017, and the joint M.S. degree in computer science from the University of West Florida, FL, USA, and Taiyuan University of Technology in 2020. She is currently pursuing the Ph.D. degree with the Department of Computer Science, College of Engineering, North Carolina State University. Her research interests include payment channel networks, security, and blockchain.

Fangtong Zhou (Graduate Student Member, IEEE) received the B.E. degree in electrical engineering and automation from Harbin Institute of Technology, Harbin, China, in 2018, and the M.S. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2020. She is currently pursuing the Ph.D. degree with the School of Computer Science, North Carolina State University. Her research interests include machine learning in computer networking, federated learning, and reinforcement learning for resource provisioning.

Jianqing Liu (Member, IEEE) received the B.S. degree from the University of Electronic Science and Technology of China in 2013 and the Ph.D. degree from the University of Florida in 2018. He is currently an Assistant Professor of computer science with NC State University. His research interests include wireless communications and networking, security, and privacy. He received the U.S. NSF CAREER Award in 2021. He also received several best paper awards, including the 2018 Best Journal Paper Award from IEEE TCGCC.

Guoliang Xue (Fellow, IEEE) is currently a Professor of computer science with Arizona State University. His research interests include wireless networking, security and privacy, and optimization. He received the IEEE Communications Society William R. Bennett Prize in 2019. He is the Steering Committee Chair of IEEE INFOCOM. He has served as the VP-Conferences of the IEEE Communications Society and an Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING and IEEE/ACM TRANSACTIONS ON NETWORKING.