Mi-Co: Models and Algorithms for Cost-efficient Entanglement Distribution in the Quantum Internet

Huayue Gu*, Zhouyu Li*, Xiaojian Wang, Dejun Yang, Guoliang Xue, Ruozhou Yu

***Abstract*—Operating a quantum network incurs high capital and operational expenditures, which are expected to be compensated by the high value of enabled quantum applications. However, existing mechanisms mainly focus on maximizing the entanglement distribution rate and neglect the cost incurred on users. This paper aims to address how to utilize quantum network resources in a cost-efficient manner while sustaining high-quantity entanglement distribution. We first consider how to establish a steady stream of entanglements between remote nodes with the minimum cost. Utilizing a recent flow-based abstraction and a novel graph representation, we design an optimal algorithm for min-cost remote entanglement distribution. Our algorithm provides rigorous tools for supporting high-performance quantum network applications with financial consideration and offers long-term data transmission in remote distances. Extensive simulation results validate the advantageous performance compared to existing solutions and heuristics.**

***Index Terms*—Quantum network, entanglement distribution, cost**

I. Introduction

Long-distance entanglement distribution is a cornerstone for many promising applications, including quantum key distribution (QKD) [4], remote quantum sensing [14], and distributed quantum computing (DQC) [10]. Recently, notable progress has been made in lab-scale quantum networks [18], hinting at potential large-scale deployment in the future [42].

The primary challenges of large-scale quantum networking are distance and noise. Entangled photons undergo high loss when transmitted in media over long distances, which cannot be compensated by classical amplification due to the no-cloning theorem [8]. Meanwhile, noisy quantum channels and imperfect quantum operations lead to decoherence of the entangling correlation, which cannot be corrected by classical error correction [32]. To counter these effects, the first-generation quantum repeaters emphasize the use of near-term feasible quantum operations, including entanglement generation and swapping, to distribute entanglements between remote nodes. Extensive study has been conducted on maximizing the quantity (rate) of entanglement distribution, based on the unique characteristics of these quantum operations [12], [37], [39].

* Both authors contributed equally to this research. Gu, Li, Wang, Yu ({hgu5, zli85, xwang244, ryu5}@ncsu.edu) are with NC State University, Raleigh, NC 27606, USA. Yang (djyang@mines.edu) is with Colorado School of Mines, Golden, CO 80401, USA. Xue (xue@asu.edu) is with Arizona State University, Tempe, AZ 85281, USA. Gu, Li, Wang, and Yu were supported in part by NSF grants 2045539 and 2350152. Yang was supported by CNS-2008935 and CNS-2414522. The research 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 number ERKJ432, as part of the PiQSci quantum networking project, and by NSF grant 2007083. The information reported herein does not reflect the position or the policy of the funding agencies.

Compared to entanglement distribution rate, other factors are less considered yet also critical. While lab-scale testbeds are feasibility-driven and largely cost-agnostic, a commercial quantum network service (such as EBP quantum network [17]) must be built atop a sustainable business model considering the capital and operational expenditures of the service and revenue. For example, deploying a single repeaterless quantum node costs over $200,000 today [17], and the cost for future quantum repeaters with transduction device, quantum memory, and purification is destined to multiply. This is not to mention the costs for workforce development, maintenance and operation. These costs shall be compensated by the high value of quantum-enabled applications. Correspondingly, like classical Internet nowadays, a real quantum internet would be built as a shared infrastructure serving as diverse quantum applications and as many quantum users as possible. Quantum network operators peer with each other to provide global quantum connectivity and decide their service prices based on market trends driven by supply and demand. Users choose among operators based on prices and service quality, and are charged based on services they receive or resources they consume. Overall, as the business paradigms of quantum technologies evolve, cost management will eventually become an integral part of the entire quantum ecosystem.

While the actual cost of operating a large-scale commercial quantum network is still hard to decide due to technology pre-maturity, we believe it is proper to start investigating cost management in quantum networking from an *algorithmic perspective*. Consider a user requesting quantum connectivity service in a quantum network. The joint interest of the user and the operator is that the service is fulfilled with the most cost-efficient resources in the network. Nevertheless, because of unique quantum characteristics, the use of different resources (e.g. fiber bandwidth, quantum storage, and swapping) has a profound impact on both the cost and the quantity of the entanglement distribution service. In this case, exploring the most cost-efficient way of serving a user request, becomes a highly non-trivial problem. As a first-step towards combining quantum network and internet economics, this paper focuses on studying the problem of how to find the *minimum-cost* way of establishing a stream of remote entanglements with a steady expected rate from an algorithmic perspective. Our main contributions are as follows:

- We formulate the minimum cost entanglement distribution problem (Mi-Co) in a general quantum network.
- We develop a Generation and Swapping Graph (GSG) abstraction, based on which we design the optimal algorithm for the min-cost distribution of a steady stream of entanglements between remote nodes.

- We conduct extensive simulations and demonstrate that the proposed algorithms achieve high cost efficiency when achieving the same rate as existing works.

Organization: §II presents the background on classical and quantum internet economics. §III shows quantum preliminaries and network model. §IV presents an optimal algorithm for min-cost entanglement distribution. §V presents evaluation results. §VI surveys the related work. §VII and §VIII presents the discussion and conclusion.

II. Classical and Quantum Internet Economics

A. Classical Internet Economics

The economic model of the classical Internet serves as an excellent example of how the quantum internet ecosystem can evolve. Here, we briefly review the cost and profit model of a classical Internet Service Provider (ISP).

ISP cost model: An ISP incurs both *capital expenditure (CapEx)* and *operational expenditure (OpEx)*. The costs for deploying network devices and materials (wire/fiber), land usage, construction, labor and permitting, and research and engineering development constitute major CapEx of an ISP. Further, operating a network infrastructure incurs continuous energy, maintenance, upgrade, device/material/space renting, customer support, as well as costs associated with peering with other ISPs to provide global connectivity to its customers, which constitute the OpEx. Both costs can be substantial on a yearly basis due to the ISP's need to continuously upgrade and operate the infrastructure for market competitiveness. For instance, the US broadband providers incurred over $102.4 billion in CapEx, marking a 19% year-to-year increase [1].

ISP profit model: ISPs subsidize the costs and make profit by charging for users' usage of Internet resources. A residential user is charged monthly based on a pre-defined service category (best-effort service with data or data rate cap). For business users, pricing is negotiated case-by-case. Specifically, many business users (such as banking, gaming, or critical services) demand guaranteed service quality such as data rate and/or delay, which must be satisfied by reserved or prioritized resources such as link bandwidth dedicated to each user. In this case, pricing is typically based on the amount of resources that a user's service consumes. In a market with multiple ISPs, a user would choose the most economical service that satisfies its demand. Thus to maximize revenue, an ISP competes by delivering the least price to satisfy a user's demand that can sustain its business, for instance, by leveraging the most cost-efficient resources and optimizing resource utilization. This has motivated extensive research on cost-efficient network optimization in the Internet, such as traffic engineering [44], resource allocation [22], [26], and so on.

B. Quantum Internet Economics

We expect the future quantum internet ecosystem to be driven by supply and demand similar to those in the classical Internet. Quantum network operators build and maintain expensive quantum network infrastructures such as optical switching centers, repeater facilities or even satellite constellations. These infrastructures will support diverse quantum applications and users to maximize the operators' profit. Users seek competitive prices on demanded services, and operators seek to optimize network operations to reduce cost and improve revenue. Nevertheless, the fundamental differences between classical and quantum networks introduce unique features of the cost management problem in the quantum internet.

Quantum network service model: A quantum network provides the basic entanglement distribution service between two or more remote end-points based on demands from quantum users. Entanglement is a basic resource for all quantum applications. The quantity (i.e., rate and success probability) of entanglement distribution fundamentally decides the practicality, usability, and performance of supported applications.

In the near future, the quantum network will likely only serve business-level customers and support high-value applications such as high-stake secure communication, distributed quantum computing centers, or highly sensitive sensing in scientific and military domains. These applications typically have high performance requirements regarding entanglement distribution rate. For instance, distributed quantum computing typically requires abundant entanglement resources to be constantly available between two or more remote quantum computers so that remote quantum operations (such as CNOT gates) can be executed as soon as possible before qubit decoherence [19], [30]. With near-term quantum technology, this is best achieved by maintaining a steady stream of entanglements between any two or more computers, such that sufficient entanglement resources in short-lived quantum memories can be ensured for other quantum operations [9], [15]. Similarly, sensors in a quantum sensor network need to maintain steady entanglement connectivity for continuous monitoring on a target quantity [47], and entanglement-based QKD relies on a stream of entanglements between key distribution centers to continuously generate enough keys for local users [35].

To provide the performance guarantee, the quantum network infrastructure can act as a virtualization platform, reserving resources to satisfy the need of each user [34], [43]. Each user's service is implemented on dedicated or prioritized resources to guarantee the rate. Correspondingly, users are charged based on **overall resource consumption** in the network for their services, such as fiber bandwidth and quantum (optical) processing at nodes. Each resource may be priced differently based on factors such as current utilization level, deployment/maintenance costs, energy consumption, and so on. To serve a user with the minimum cost/price, it is thus important to **select the most cost-efficient resources** that can satisfy the user's demand.

Unique features of quantum network cost management: As we model in detail in the next section, the characteristics of quantum networks make cost management drastically different from the classical Internet.

First, compared to the classical Internet, where bandwidth is a uniform resource across links, quantum resources such as fiber bandwidth exhibit drastic heterogeneity in resource (cost) efficiency for realizing the same service quality. Specifically, factors such as distance may significantly affect the achievable rate of entanglements using the same optical bandwidth. The success probability of key operations like generation and swap-

TABLE I: Notation Table of Network Model

Parameters	Description
$G=(V,E)$	quantum network with nodes V and links E
mn, M	an enode, and set of enodes $M=\{mn \mid m,n \in V\}$
q_e, q_n	ebit generation & swapping success probabilities
c_e^{gen}, c_n	ebit generation & swapping costs
p	a pflow $\{f_{mn}^{mk}, g_{mn}, \kappa_{mn}\}$ following Definition 1
c_p	expected cost of a pflow p
I_{mn}	total ebit rate generated between node pair mn
Ω_{mn}	total ebit rate contributed by mn to swapping
λ_{st}, c_{st}	expected EDR and cost of an SD pair in Eqs. (2)–(4)

ping may also differ across nodes, as decided by the physical environment and device heterogeneity.

Second, each unit resource may result in different rate when distributing between different end points, as decided by the different generation and swapping processes to implement the end-to-end distribution. A link that can generate 20 entanglements per second (eps) may support only 10 end-to-end eps if one swapping step with success probability 50% is applied, or only 5 end-to-end eps if two swapping steps are applied. This means the length of a path can result in not just linear accumulation of node and link cost but also multiplicative increase in each node/link's resource consumption to provide a certain entanglement rate.

Combined with the non-trivial modeling of processes such as generation and swapping themselves, the overall resource cost minimization problem becomes highly complex, and fundamentally different from cost minimization in classical networks. A new algorithmic framework is needed to enable efficient cost management in the quantum internet.

III. Quantum Preliminaries

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

A. Entanglement Distribution Process

We consider the fundamental task of distributing two-qubit (bipartite) entangled states between remote end points. A *maximally entangled bipartite state* (called a Bell state or an *ebit*) is a fundamental resource in quantum communication, which can be used to construct arbitrary multi-partite entangled states and support distributed quantum applications. We focus on distributing one of the four Bell states and use $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle+|11\rangle)$ as the example[1]. Below, we introduce the physical processes to implement remote entanglement distribution.

Entanglement generation: Assume two quantum nodes intend to exchange quantum information. An entanglement source can continuously generate entangled photon pairs via spontaneous parametric down-conversion (SPDC) [11] or four-wave mixing [25] and send each photon of the pair to two different quantum nodes via a quantum channel such as optical fiber or free space. Alternatively, a dedicated entanglement source can reside in-between the two nodes, and both photons of a generated pair can be transmitted to the two nodes respectively along the fiber. This is entanglement generation, and ebits generated and distributed along a physical channel are *elementary ebits*.

[1]All four Bell states—$|\Phi^{\pm}\rangle$ and $|\Psi^{\pm}\rangle$— are symmetric under local operations and classical communication (LOCC).

As a non-linear optical operation, the efficiency of the SPDC-based entanglement source is very low with current technologies, typically in the order of 10^{-6} ebits per input pump photon [6]. Meanwhile, photon transmission over fiber incurs the exponential loss of photons, typically in the order of 0.1dB/km, leading to excessive loss over tens of kilometers [13]. One way to improve the rate is to increase the pump power [5], [20]. However, this also leads to multi-photon coincidence events—generating more than one entangled pair in one photon detection period that cannot be distinguished—which may even degrade the generation rate [28]. Since the entangled photons will get lost during transmission due to channel attenuation and other environmental factors, the entanglement generation can be viewed as a probabilistic process [33]. Following the existing work [13], we use $q_e \in (0,1]$ to denote the entanglement success probability as

$$q_e = 1 - \left[1 - p_{succ}\left(1 - p_{tloss}\right)\right]^{N_{attp}}, \tag{1}$$

where p_{succ} is the generation efficiency of entanglement source, p_{tloss} is the probability of transmission loss, and N_{attp} is the number of attempts in unit slot. Here the transmission loss probability increases exponentially with the link length.

Entanglement swapping: Quantum repeaters are designed to alleviate the excessive loss of entangled photons over long distances via entanglement swapping. Assume two remote nodes want to exchange quantum information, instead of directly transmitting entangled photons, both nodes may each establish an ebit with an intermediate repeater, who then concatenates the two ebits into one end-to-end ebit between the two nodes. Each quantum swapping consists of a two-qubit operation for locally entangling the two ebits at the repeater, a one-qubit operation for correction, and a Bell state measurement. However, each entanglement swapping operation has a chance to fail due to the protocol design and the imperfections in the swapping circuit [3]. Therefore, we use $q_n \in (0,1]$ to denote the swapping success probability of each repeater n.

B. Cost Economics for Entanglement Distribution

As discussed in Sec. II, our pricing model is motivated by the unprecedented success of Internet economics, where *traffic* is the primary commodity sold by ISPs and eventually consumed by Internet users [27], [29], [38]. The bandwidth on links, along with possible processing on nodes (such as middleboxes), is priced as *resource costs* to the users. In a quantum network, users will similarly be charged based on the amount of resources consumed on nodes and links, including entanglement source capacity, optical bandwidth, and swapping operations. Below, we define the resource cost for each physical process.

Entanglement generation: Consider $c_e^{\mathsf{gen}} > 0$ as the cost for *attempting* to generate one entanglement per unit time along a link e. The cost includes using the entanglement source capacity and occupying the optical bandwidth on this link. Considering the success probability of entanglement generation q_e, if the target expected rate of generated entanglements over e is λ_e, then the expected cost to achieve this rate is $c_e^{\mathsf{gen}} \cdot \lambda_e / q_e$, factoring in the success probability on this link.

Entanglement swapping: Swapping consumes memories and quantum operations on a repeater. Because swapping also suc-

ceeds only with a probability, if an expected rate of successful swapping at node n is λ_n, then the actual cost to achieve this rate is $c_n \cdot \lambda_n / q_n$, given unit swapping cost $c_n > 0$.

As discussed in Sec. II-B, purchasing different "bundles" of resources has a much more profound impact on the cost-performance trade-off of a user than on the Internet. If a user's service is implemented along paths that result in low rate, then achieving the satisfactory rate (explained in the next subsection) may cost much more resources than picking some better paths. How to pick the best suite of resources, and utilize them to achieve the best performance is thus an important algorithmic problem faced by a quantum user and/or the network operator. As the future quantum internet is expected to serve many users (SD pairs) with sufficient capacity, this paper focuses on a single SD pair, with possible extension to the multi-SD pair case discussed in Sec. VII.

C. Quantum Network Model

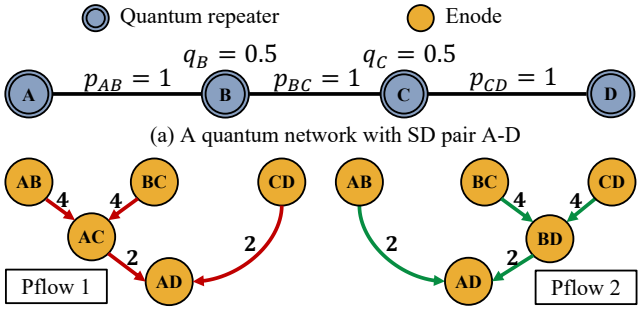
We model a quantum network as an undirected graph $G = (V, E)$, where V is the set of quantum nodes, and E is the set of physical channels (links) between nodes. Each link $e \in E$ is equipped with an entanglement generation probability $q_e \in (0, 1]$ and a unit cost $c_e^{\text{gen}} \in \mathbb{R}^+$ for attempting to generate an elementary ebit. Each repeater $v \in V$ is equipped with a swapping success probability $q_v \in (0, 1]$ and a unit cost $c_v \in \mathbb{R}^+$ for swapping. We also use mn to denote an *unordered* node pair $\{m, n\}$ for $m, n \in V$, and hence $mn = nm$. Each pair mn is called an *enode* denoting a pair of nodes between which entanglement may be established. The set of all enodes is $M = \{mn \,|\, m, n \in V\}$, and note that E is a subset of M. Following the conventional notation, we call a pair of communicating nodes st a source-destination (SD) pair without explicitly defining which node is the source or destination.

We adopt a time-slotted system model following [45], [46], while all our definitions and algorithms can be trivially extended to continuous-time asynchronous operations [41]. We also assume a central controller controls all operations in the quantum network as [37], [46] to monitor network status, allocate resources, and decide the costs of different operations. In each time slot, the following phases are carried out in order:

1) **Entanglement generation:** For a pair of nodes $mn \in E$ with a direct link, they will attempt to generate elementary mn-ebits at a pre-defined rate.
2) **Entanglement swapping:** When ebits are available between enode mk and kn sharing the repeater k, repeater k can attempt to perform swapping between pairs of mk- and kn-ebits to create ebits between nodes mn. The source mk- and kn-ebits are not required to be elementary.

D. Satisfying User Demand

As motivated in Sec. II-B, a user submits its service demand as the expected rate of a steady entanglement stream between an SD pair st for an extended period of time. To satisfy this demand, we utilize (and later extend) an abstraction originally developed in [12] to describe how a user demand can be implemented in a quantum network.



(a) A quantum network with SD pair A-D

(b) Two pflows for generating 1-unit ebit between A-D

Fig. 1: Pflows in a quantum network. Values on each pflow's arcs represent the required entanglement generation rates to get one end-to-end EDR between AD.

let $\lambda_{st}(\tau)$ be the number of generated st-ebits in time slot τ. Let $\mathbb{E}[\cdot]$ denote expectation. The long-term entanglement distribution rate (EDR) is defined as

$$\lambda_{st} = \liminf_{T \to \infty} \frac{1}{T} \sum_{\tau=1}^{T} \mathbb{E}[\lambda_{st}(\tau)]. \tag{2}$$

To achieve an EDR, a remote entanglement distribution (RED) algorithm needs to decide how to generate and swap entanglements in the network. The *(primitive) entanglement flow (pflow)* abstraction in [21] lends us a tool to define and devise an RED algorithm that achieves a long-term EDR λ_{st}.

Definition 1 (Pflow)**.** *Given quantum network G and SD pair $st \in M$, a pflow with EDR λ_{st} is defined by variables $\{g_{mn} \geq 0 | \, mn \in M\} \cup \{f_{mn}^{mk} \geq 0 \,|\, m, k, n \in V\}$, such that*
(i) ***Generation:*** *$g_{mn} = 0$ for $\forall mn \notin E$;*
(ii) ***Swapping:*** *$f_{mn}^{mk} = f_{mn}^{kn}$ for $\forall m, k, n \in V$;*
(iii) ***Conservation:*** *$I_{mn} - \Omega_{mn} = 0$ for $\forall mn \neq st$, and $I_{st} - \Omega_{st} = \lambda_{st}$, where $I_{mn} = g_{mn} \cdot q_{mn} + q_k \sum_{k \neq m,n} f_{mn}^{mk}$ and $\Omega_{mn} = \sum_{k \neq m,n} (f_{mk}^{mn} + f_{kn}^{mn})$;*
(iv) ***Primitiveness:*** *For any $mn \in M$, either $g_{mn} > 0$, or there exists at most one k such that $f_{mn}^{mk} > 0$, but not both.* □

Explanation: Here, g_{mn} denotes the rate of entanglement generation attempts on each link $mn \in E$, and f_{mn}^{mk} denotes the contribution of mk-ebits to generate mn-ebits via swapping. Condition (i) ensures the generation of elementary ebits only on physical links. Condition (ii) ensures that swapping consumes equal amounts of ebits from two source enodes. Condition (iii) maintains entanglement flow conservation by having $I_{mn} = \Omega_{mn}$ at any non-SD pair enode mn (*i.e.*, all generated mn-ebits contribute to further swapping), and defines the EDR λ_{st} of SD pair st, where I_{mn} and Ω_{mn} are the input and output ebit rates at enode mn respectively. Note that we slightly abuse the notation and define $q_{mn} = 0$ for $mn \notin E$ for notation simplicity. Finally, condition (iv) ensures that each pflow defines a *unique* way of generating end-to-end st-ebits by dictating that ebits between any enode mn can only be generated in exactly one way: either via elementary ebit generation ($g_{mn} > 0$) or via swapping at an intermediate repeater k ($f_{mn}^{mk} = f_{mn}^{kn} > 0$ for at most one k), but not both.

Example: Fig. 1 shows a repeater chain with four nodes A, B, C, D and SD pair AD. While there is only one path, Fig. 1(b) shows that there are two different orders of swapping

along this path. After entanglement generation along physical links, one way is to swap AB- and BC-ebits first, and then swap the successfully generated AC-ebits with CD-ebits. In the second way, BC- and CD-ebits are swapped first, and the successful ones are then swapped with AB-ebits. Crucially, these two ways will result in different entanglement generation and swapping rates on different links/nodes to satisfy the same EDR. Assume link generation probabilities are all 1, and swapping probabilities are uniformly 0.5. In the first way, maintaining $\lambda_{AD} = 1$ would require generation rates $g_{AB} = g_{BC} = 4$ and $g_{CD} = 2$, and swapping rates $f_{AC}^{AB} = f_{AC}^{BC} = 4$ and $f_{AD}^{AC} = f_{AD}^{CD} = 2$. But in the second way, maintaining the same λ_{AD} would instead need generation rates $g_{AB} = 2$ and $g_{BC} = g_{CD} = 4$, and swapping rates $f_{BD}^{BC} = f_{BD}^{CD} = 2$ and $f_{AD}^{AB} = f_{AD}^{BD} = 2$. There are different ways to obtain the same EDR, leading to different generation and swapping rates at links and nodes. When the links and nodes have different costs, it is worthwhile to investigate the impact of how entanglements are established on the total cost of the entanglement users. Specifically, the pflow abstraction has interesting properties, which can be utilized to design efficient algorithms for remote entanglement distribution:

1) Each pflow corresponds to exactly one path in G.
2) Given a pflow p with EDR λ_{st}, the expected cost c_p for generating an end-to-end ebit is well-defined as follows:

$$c_p \triangleq \frac{1}{\lambda_{st}} \left(\sum_{m,n,k \in p} c_k f_{mn}^{mk} + \sum_{e \in E \cap p} \frac{c_e^{\mathsf{gen}}}{q_e} g_e \right). \tag{3}$$

where c_e^{gen}/q_e is the unit generation cost on a link scaled by its generation success probability's reciprocal

To achieve an expected EDR of λ_{st}, observe that a pflow specifies a deterministic rate of entanglement generation g_{mn} on each link in each time slot; however, the actual rates of generation and swapping both depend on the rate of *successful* generation, which is a random variable. Let $c_{st}(\tau)$ be the total cost of performing all operations within one time slot. Following a given pflow p, the long-term expected cost is

$$c_{st} = \liminf_{T \to \infty} \frac{1}{T} \sum_{\tau=1}^{T} \mathbb{E}[c_{st}(\tau)] = c_p \cdot \lambda_{st}. \tag{4}$$

This shows that we can focus on minimizing or bounding the expected cost of maintaining a unit end-to-end EDR of 1, and then scale the cost by the actual desired EDR λ_{st} of the SD pair. Our next focus is an algorithm for cost minimization.

IV. MI-CO: MINIMUM-COST ENTANGLEMENT DISTRIBUTION

In this section, we start with the problem of minimizing the cost and propose the optimal minimum cost pflow algorithm.

A. Problem Description

We define the following min-cost remote entanglement distribution (Mi-Co) problem based on the pflow definition above:

Definition 2 (Mi-Co)**.** *Let $G = (V, E)$ be a quantum network with a source-destination pair st. The **minimum-cost remote entanglement distribution** problem is to find an st-pflow p_{st} such that c_p is minimized, with $\lambda_{st} = 1$.*

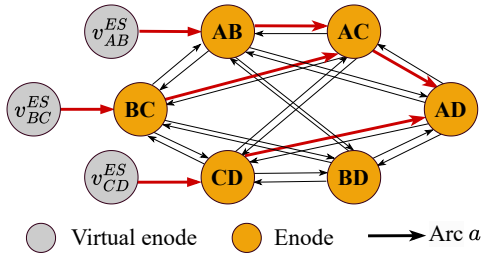


Fig. 2: An example GSG for the network in Fig. 1. Pflow 1 in Fig. 1 is shown as the red bold arcs.

The min-cost pflow represents the most cost-efficient way of generating a steady stream of ebits between st via probabilistic quantum operations. In the following, we design an optimal algorithm to solve this problem, with the help of a new abstraction called a *Generation and Swapping Graph (GSG)*.

B. Generation and Swapping Graph (GSG)

We define an auxiliary graph structure to visualize a pflow and design an efficient combinatorial algorithm for Mi-Co.

Definition 3 (GSG)**.** *Given a quantum network G, its GSG is a directed multi-graph $G_{\mathsf{GSG}} = (M \cup \mathcal{V}_E, \mathcal{A})$, where each vertex in G_{GSG} is either an enode $m \in M$ or a virtual enode $v_{mn}^{ES} \in \mathcal{V}_E$ for $mn \in E$. The edge set consists of two types of edges (arcs): $\mathcal{A} = \mathcal{A}_{\mathsf{src}} \cup \mathcal{A}_{\mathsf{swp}}$. The set $\mathcal{A}_{\mathsf{src}}$ contains arcs $a_{mn} = (v_{mn}^{ES}, mn)$ denoting the generation of ebits on each physical link $mn \in E$; each arc has cost $c(a_{mn}) = c_{mn}$. The set $\mathcal{A}_{\mathsf{swp}}$ contains arcs $a = (mk, mn)$ for any $m, k, n \in V$, denoting swapping contribution from mk to mn with swapping cost $c(mk, kn) = c_k/2$ so that we have $c(mk, mn) + c(kn, mn) = c_k$ which denotes the full unit swapping cost at k.*

A GSG covers every possible enode between which ebits can be generated, and includes all possible pflows for generating these ebits as subgraphs. Specifically, a pflow is described by a *binary tree* in the GSG with special structural requirements. As in Fig. 2, a quantum network in Fig. 1 can be transformed into a GSG. The pflow in Fig. 1(b) is depicted in Fig. 2 with red bold arcs showing generation and swapping operations in the pflow. A pflow has all arcs pointing to the root enode st, and each enode in the binary tree has either one incoming arc from a virtual enode by entanglement generation or two matching incoming arcs from two enodes by swapping but not both.

A key advantage of the GSG is its **polynomial** size, compared to the possibly exponential number of different pflows in the network. Specifically, the GSG of a quantum network G has $|V|(|V|-1)+|E|$ enodes plus virtual enodes and $O(|V|^3+|E|)$ arcs. We highlight that while we utilize the GSG for assisting understanding and algorithm design, in implementation it does not need to be explicitly built. Instead, the graph structure can be implicitly incorporated into the algorithm to reduce time complexity and save memory.

C. Minimum-Cost Pflow Algorithm Design

With the GSG abstraction, we design Algorithm 1 to solve the Mi-Co problem optimally. Line 1 initializes visited flag $vis(\cdot)$, the parent enode(s) $pa(\cdot)$ and the unit generation cost $cost(\cdot)$

Algorithm 1: Optimal algorithm for min-cost pflow (Mi-Co)

Input: GSG $G_{\mathsf{GSG}} = (M \cup \mathcal{V}_E, \mathcal{A})$, SD pair st
Output: Pflow p between SD pair st

1 $vis(v) \leftarrow 0, pa(v) \leftarrow \perp, cost(v) \leftarrow \infty, \forall v \in M \cup \mathcal{V}_E$; $cost(v) \leftarrow 0, \forall v \in \mathcal{V}_E$;
2 **while** $\exists vis(v) = 0$ **do**
3 $\quad v \leftarrow \arg\min_{v \in M \cup \mathcal{V}_E}\{cost(v) \mid vis(v) = 0\}$;
4 $\quad vis(v) \leftarrow 1$;
5 $\quad$ **if** $v = v_{mn}^{ES}$ *is virtual enode in* $\mathcal{V}_E$ **then**
6 $\quad\quad cost(mn) \leftarrow cost(v) + \min_K\{c_{mn}(K)\}$;
7 $\quad\quad pa(mn) \leftarrow v_{mn}^{ES}$;
8 $\quad$ **else**
9 $\quad\quad$ **for** $k \in V$ **do**
10 $\quad\quad\quad$ **if** $vis(mk) = 0$ **then**
11 $\quad\quad\quad\quad$ **if** $cost(mk) > (cost(mn) + cost(nk) + c(mn, mk) + c(nk, mk))/q_n$ **then**
12 $\quad\quad\quad\quad\quad cost(mk) \leftarrow (cost(mn) + cost(nk) + c(mn, mk) + c(nk, mk))/q_n$;
13 $\quad\quad\quad\quad\quad pa(mk) \leftarrow (mn, kn)$;
14 $\quad\quad\quad$ **if** $vis(nk) = 0$ **then**
15 $\quad\quad\quad\quad$ **if** $cost(nk) > (cost(mk) + cost(mn) + c(mn, nk) + c(nk, mk))/q_m$ **then**
16 $\quad\quad\quad\quad\quad cost(nk) \leftarrow (cost(mk) + cost(mn) + c(mn, nk) + c(nk, mk))/q_m$;
17 $\quad\quad\quad\quad\quad pa(nk) \leftarrow (mn, mk)$;
18 Backtrack from enode st to get the min-cost pflow p;
19 **return** *min-cost pflow* p.

for all enodes in the GSG. In each iteration, the algorithm starts from the enode with the least cost and then marks it as visited in Lines 3–4. In Lines 5–17, the algorithm checks if the newly visited enode leads to a more cost-efficient way of generating ebits for another (neighbor) enode of the visited. If a virtual enode v_{mn}^{ES}, then the only neighbor enode mn of v_{mn}^{ES} is updated with direct generation in Line 6. If mn is visited, then swapping is attempted for any unvisited enode mk or kn, to see if mn-nk or mn-mk can lead to less cost for mk or kn, respectively. Note how the $cost(mk)$ is updated based on the costs of source enodes $cost(mn)$ and $cost(nk)$, plus the swapping cost c_k, discounted by the success probability q_n as in Line 12, and similarly for $cost(nk)$ in Line 16. This cost update precisely computes the cost c_p as defined in Eq. (3) along with Definition 1, where p is the min-cost pflow found so far for enode mk (and respectively nk). The algorithm ends when all enodes are visited. Then, based on the parent and cost sets, the st-pflow with minimum cost can be constructed by backtracking from the enode st through the parent set $pa(\cdot)$.

Theorem 1. *Algorithm 1 computes the minimum-cost pflow for generating ebits between s and t in polynomial time.* □

Proof. We prove the optimality of Algorithm 1 by induction. In the **base** case, for all enodes mn whose optimal way of generation is by generating elementary ebits via direct links, their costs are correctly calculated by their first updates via the virtual enodes v_{mn}, which all have cost 0 and hence are visited before any other enodes. In the **inductive** case, consider mn as the next unvisited enode with minimum cost, and let M_{mn} be all enodes that have been visited so far whose minimum costs have been correctly computed by the induction assumption. Because mn has the minimum cost among all unvisited nodes, its current way of generation (pflow rooted at mn) must be optimal, otherwise, there exists at least one other enode which must have a lower cost than mn's current cost to contribute to a way of generating mn with less cost. It follows that when any enode mn is visited, its current cost must be minimum. For the time complexity, since there are at most $O(|V|^2)$ enodes and $O(|V|^3)$ arcs in the GSG, the overall complexity is $O(|V|^3 \log |V|)$ using a Fibonacci heap for Line 3. □

V. Performance Evaluation

In this section, we perform simulation-based evaluation to validate our theoretical results. We use random Waxman graphs [40] with parameters $\alpha = \beta = 0.5$. A typical inter-city quantum network has no more than 10 quantum nodes right now [2], hence we set the default graph size to 20. Graph nodes are randomly located in a synthetic 10km-by-10km area to represent the near-term city-scale quantum network setup [2]. Following existing work [12], [21], each node has an entanglement swapping success probability uniformly sampled from $[0.5, 0.75]$. Our quantum link parameters are set according to [13] where p_{succ} is 10^{-4}. Due to the lack of real cost data, we aim to validate theoretical analysis and show our algorithms' performance with fair comparison to existing works and baselines. We infer the cost based on available data to model the entanglement distribution process in a quantum network. Link generation costs are assumed to be proportional to the distance, which aligns with the existing cost for renting dark fiber, factoring in other costs like maintenance, device operations, and energy consumption [16]. Specifically, we set the unit cost for attempting entanglement generation as ξl_e, where ξ is the entanglement generation cost coefficient, representing the expenses of generating one entanglement on a 1-km quantum link. By default, we set $\xi = 5$. The cost for performing one swapping is 3, considering the passive quantum operations, which include one-qubit, two-qubit operations, and BSM, each costing 1 per operation. In each setting, we generate 10 graphs, each with 5 random SD pairs. Results are averaged over 5 runs in the same setting to average out random noise.

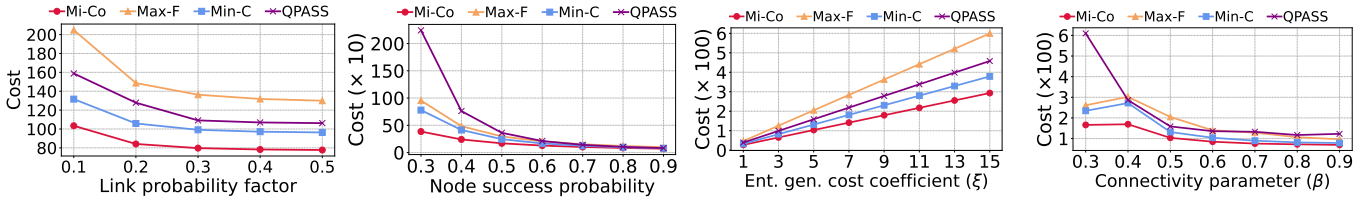
The following algorithms are compared:

- **Mi-Co:** Algorithm 1 with minimum cost.
- **Max-F:** Modified Dijkstra algorithm inspired by [24] to find the path between the source and sink with maximum fidelity without purification.
- **Min-C:** Modified Dijkstra algorithm inspired by [24] to find the path between the source and sink with minimum cost without purification.
- **QPASS:** Path-based entanglement routing in [37] modified to minimize cost.

We use ***cost*** to measure the expected cost for generating an end-to-end entanglement with the selected path or pflow.

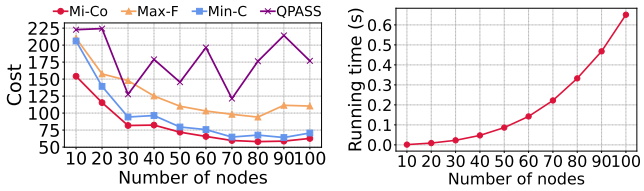
A. Evaluation Results

1) Expected cost comparison: Fig. 3 shows the expected cost of Mi-Co and other algorithms with varying link success



(a) Impact of link success probability (b) Impact of node success probability (c) Impact of generation cost factor (ξ). (d) Impact of network connectivity.

Fig. 3: Expected cost between Mi-Co and compared algorithms



(a) Cost comparison (b) Running time comparison

Fig. 4: Impact of network size.

probability factors, node success probabilities, entanglement generation cost coefficient l_e, and network connectivity parameter β. Mi-Co, being an optimal algorithm, consistently outperforms heuristic solutions in cost minimization. We observe that Mi-Co maintains the lowest costs across different link success probability factors, as shown in Fig. 3(a). Additionally, in Fig. 3(b), the expected cost decreases as node success probability increases. This indicates that leveraging more accurate devices on quantum nodes, while leading to higher infrastructure setup costs, significantly reduces runtime expenses. Fig. 3(c) illustrates the cost of Mi-Co and other algorithms with varying generation cost coefficient l_e. The cost increases linearly with the coefficient. In a quantum network modeled as a Waxman graph, quantum links between node pairs are established with probability $\beta e^{-\frac{d}{\alpha l_e}}$. Increasing the connectivity parameter β linearly raises the probability of forming quantum links between node pairs, thus enhancing overall connectivity. Fig. 3(d) demonstrates that costs decrease as network connectivity increases. This suggests that building more links among quantum nodes can reduce the operational expenses of a quantum network.

2) Impact of network size: Figs. 4(a)–(b) illustrate the cost and algorithm running time for varying numbers of quantum nodes in a fixed-size area. As the number of quantum nodes increases, the cost decreases in Fig. 4(a), while the algorithm running time increases in Fig. 4(b). In Fig. 4(a), Mi-Co consistently provides the optimal solution, outperforming baseline algorithms across all network sizes. In Fig. 4(b), although Mi-Co's running time increases with the number of deployed quantum nodes, it remains under one second on a commodity-off-the-shelf desktop for networks with up to 100 nodes, a scale not anticipated in near-future inter-city quantum networks.

To summarize, using the pflow abstraction, **Mi-Co can achieve the optimal cost** for distributing entanglement, lower than the lowest cost achievable by any path-based algorithm;

VI. Related Work

Early work in quantum networking focused on feasibility demonstration in ideal situations. Schoute *et al.* [36] presented efficient routing algorithms in ring and sphere topologies. Pant *et al.* [33] developed and analyzed entanglement routing in a square-grid topology. The limitation of these is the focus on special network topologies that are hardly realistic in practice.

Recent works have focused on general quantum networks. Shi *et al.* [37] proposed Q-PASS and Q-CAST for entanglement routing to maximize throughput. Zhao *et al.* [45] designed a redundant entanglement provisioning and selection algorithm to maximize throughput. Dai *et al.* [12] proposed the first optimal remote entanglement distribution (ORED) protocol to achieve the maximum end-to-end throughput with long-term quantum memories. The above works ignore the critical cost factor in protocol design, which is the main driving force behind the Internet. Some works have considered the energy efficiency of quantum communication. Galve *et al.* [20] investigated the energy cost of entanglement production, which scales exponentially with the amount of entanglement in the harmonic chain. Meier *et al.* [31] proposed a general framework for the energy consumption in quantum and classical computation. Jaschke *et al.* [23] theoretically showed the potential quantum advantage in energy efficiency from the quality of quantum gates and the entanglement generated in the quantum processing unit. However, these are neither comprehensive nor directly related to quantum networking design. Meanwhile, they ignore the service price from end users instead of service providers.

Our work seeks to fill this gap by introducing the optimal algorithm to minimize the remote entanglement distribution cost from users' perspective. Our approach is inspired by ORED above and FENDI in [21]. However, both of them do not have any cost-related considerations. This paper builds upon the abstractions in ORED and FENDI and extends them to a new graph-based abstraction considering cost as a primary constraint due to its practical importance, in addition to entanglement rate. This consideration makes the problem that we study significantly more challenging than ORED or FENDI.

VII. Discussion

Tackling multiple SD pairs: This paper focuses on optimizing the cost for fulfilling a single user request (an SD pair), from the view of a cost-minimizing network operator or the user. If the network has enough resources to handle many SD pairs without bottleneck, then our algorithms can be easily extended to the multi-SD pair case by solving for each SD pair independently. The case with resource contention is more complex as have been studied in many works discussed in Sec. VI. We highlight one direction where our algorithm may contribute to multi-SD pairs entanglement distribution. By setting costs of links and nodes dynamically based on the current resource contention, one may utilize our algorithms

to develop competitive online algorithms for the network operator to maximize the total number of user requests that can be accepted, following existing online algorithm design frameworks [7]. We leave the exploration of such algorithms and other cost-aware networking problems as future work.

VIII. Conclusions

In this paper, we designed cost-efficient entanglement distribution protocols in a general optical quantum network. We started with the problem of establishing a stream of entanglements with minimum cost and designed an optimal algorithm based on a new abstraction called the Generation and Swapping Graph (GSG). Extensive simulation results showed the superior performance of our algorithms in terms of cost efficiency compared to existing algorithms and heuristic solutions.

References

[1] "USTelecom 2022 Broadband Capex Report." URL: https://www.ustelecom.org/research/2022-broadband-capex/

[2] G. Avis, F. Ferreira da Silva, T. Coopmans, A. Dahlberg, H. Jirovská, D. Maier, J. Rabbie, A. Torres-Knoop, and S. Wehner, "Requirements for a processing-node quantum repeater on a real-world fiber grid," *NPJ Quantum Information*, vol. 9, no. 1, p. 100, 2023.

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

[4] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theoretical Computer Science*, vol. 560, pp. 7–11, 2014.

[5] C. Bény, C. T. Chubb, T. Farrelly, and T. J. Osborne, "Energy cost of entanglement extraction in complex quantum systems," *Nature Communications*, vol. 9, no. 1, p. 3792, 2018.

[6] M. Bock, A. Lenhard, C. Chunnilall, and C. Becher, "Highly efficient heralded single-photon source for telecom wavelengths based on a ppln waveguide," *Optics Express*, vol. 24, no. 21, pp. 23 992–24 001, 2016.

[7] N. Buchbinder and J. (Seffi) Naor, "The Design of Competitive Online Algorithms via a Primal—Dual Approach," *Foundations and Trends® in Theoretical Computer Science*, vol. 3, no. 2–3, pp. 93–263, 2009.

[8] V. Bužek and M. Hillery, "Quantum copying: Beyond the no-cloning theorem," *Physical Review A*, vol. 54, no. 3, p. 1844, 1996.

[9] M. Caleffi, M. Amoretti, D. Ferrari, J. Illiano, A. Manzalini, and A. S. Cacciapuoti, "Distributed quantum computing: a survey," *Computer Networks*, vol. 254, p. 110672, 2024.

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

[11] C. Couteau, "Spontaneous parametric down-conversion," *Contemporary Physics*, vol. 59, no. 3, pp. 291–304, 2018.

[12] W. Dai, T. Peng, and M. Z. Win, "Optimal protocols for remote entanglement distribution," in *IEEE ICNC*, 2020, pp. 1014–1019.

[13] ——, "Quantum queuing delay," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 3, pp. 605–618, 2020.

[14] G. M. D'Ariano, P. L. Presti, and M. G. Paris, "Using entanglement improves the precision of quantum measurements," *Physical Review Letters*, vol. 87, no. 27, p. 270404, 2001.

[15] M. G. Davis, J. Chung, D. Englund, and R. Kettimuthu, "Towards distributed quantum computing by qubit and gate graph partitioning techniques," in *IEEE QCE*, vol. 1, 2023, pp. 161–167.

[16] D. Dehlinger and M. Mitchell, "Entangled photon apparatus for the undergraduate laboratory," *American Journal of Physics*, vol. 70, no. 9, pp. 898–902, 2002.

[17] D. D. Earl, "A scalable quantum cryptography network for protected automation communication," Qubitekk, Inc, Tech. Rep., 2022.

[18] C. Elliott, "Building the quantum network," *New Journal of Physics*, vol. 4, no. 1, p. 46, 2002.

[19] M. Fellous-Asiani, "The resource cost of large scale quantum computing," *arXiv preprint arXiv:2112.04022*, 2021.

[20] F. Galve and E. Lutz, "Energy cost and optimal entanglement production in harmonic chains," *Physical Review A*, vol. 79, no. 3, p. 032327, 2009.

[21] H. Gu, Z. Li, R. Yu, X. Wang, F. Zhou, J. Liu, and G. Xue, "Fendi: Toward high-fidelity entanglement distribution in the quantum internet," *IEEE/ACM Transactions on Networking*, 2024.

[22] L. Gu, D. Zeng, S. Guo, A. Barnawi, and Y. Xiang, "Cost efficient resource management in fog computing supported medical cyber-physical system," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, no. 1, pp. 108–119, 2015.

[23] D. Jaschke and S. Montangero, "Is quantum computing green? an estimate for an energy-efficiency quantum advantage," *Quantum Science and Technology*, vol. 8, no. 2, p. 025001, 2023.

[24] H. Leone, N. R. Miller, D. Singh, N. K. Langford, and P. P. Rohde, "Qunet: Cost vector analysis & multi-path entanglement routing in quantum networks," *arXiv preprint arXiv:2105.00418*, 2021.

[25] X. Li, P. L. Voss, J. E. Sharping, and P. Kumar, "Optical-fiber source of polarization-entangled photons in the 1550 nm telecom band," *Physical Review Letters*, vol. 94, no. 5, p. 053601, 2005.

[26] Z. Li, M. Chen, G. Li, X. Lin, and Y. Liu, "Map-driven mmwave link quality prediction with spatial-temporal mobility awareness," *IEEE Transactions on Mobile Computing*, 2024.

[27] R. E. Litan and A. M. Rivlin, "Projecting the economic impact of the internet," *American Economic Review*, vol. 91, no. 2, pp. 313–317, 2001.

[28] X.-s. Ma, S. Zotter, J. Kofler, T. Jennewein, and A. Zeilinger, "Experimental generation of single photons via active multiplexing," *Physical Review A*, vol. 83, no. 4, p. 043814, 2011.

[29] J. K. MacKie-Mason and H. Varian, "Economic faqs about the internet," *Journal of Economic Perspectives*, vol. 8, no. 3, pp. 75–96, 1994.

[30] Y. Mao, Y. Liu, and Y. Yang, "Qubit allocation for distributed quantum computing," in *IEEE INFOCOM*, 2023, pp. 1–10.

[31] F. Meier and H. Yamasaki, "Energy-consumption advantage of quantum computation," *arXiv preprint arXiv:2305.11212*, 2023.

[32] M. A. Nielsen, "The entanglement fidelity and quantum error correction," *arXiv preprint quant-ph/9606012*, 1996.

[33] M. Pant, H. Krovi, D. Towsley, L. Tassiulas, L. Jiang, P. Basu, D. Englund, and S. Guha, "Routing entanglement in the quantum internet," *npj Quantum Information*, vol. 5, no. 1, pp. 1–9, 2019.

[34] S. Pelletier, R. Yu, G. Rouskas, and J. Liu, "Qubit recycling in entanglement distillation," in *IEEE QCE*, vol. 1, 2023, pp. 32–38.

[35] G. Ribordy, J. Brendel, J.-D. Gautier, N. Gisin, and H. Zbinden, "Long-distance entanglement-based quantum key distribution," *Physical Review A*, vol. 63, no. 1, p. 012309, 2000.

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

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

[38] R. Srivastava, I. Choi, T. Cook, N. U. E. Team *et al.*, "The commercial prospects for quantum computing," *Networked Quantum Information Technologies*, pp. 2018–10, 2016.

[39] Y. Wang, Y. Zhao, L. Huang, and C. Qiao, "Routing and wavelength assignment for entanglement swapping of photonic qubits," in *IEEE INFOCOM*, 2024, pp. 1431–1440.

[40] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, 1988.

[41] L. Yang, Y. Zhao, L. Huang, and C. Qiao, "Asynchronous entanglement provisioning and routing for distributed quantum computing," in *IEEE INFOCOM*, 2023.

[42] J. Yin, Y. Cao, Y.-H. Li, S.-K. Liao, L. Zhang, J.-G. Ren, W.-Q. Cai, W.-Y. Liu, B. Li, H. Dai *et al.*, "Satellite-based entanglement distribution over 1200 kilometers," *Science*, vol. 356, no. 6343, pp. 1140–1144, 2017.

[43] Y. Zeng, J. Zhang, J. Liu, Z. Liu, and Y. Yang, "Multi-entanglement routing design over quantum networks," in *IEEE INFOCOM*, 2022.

[44] G. Zhao, J. Wang, H. Xu, Z. Yu, and C. Qiao, "Coin: Cost-efficient traffic engineering with various pricing schemes in clouds," in *IEEE INFOCOM*, 2023, pp. 1–10.

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

[46] Y. Zhao, G. Zhao, and C. Qiao, "E2E fidelity aware routing and purification for throughput maximization in quantum networks," in *IEEE INFOCOM*, 2022.

[47] Q. Zhuang, Z. Zhang, and J. H. Shapiro, "Distributed quantum sensing using continuous-variable multipartite entanglement," *Physical Review A*, vol. 97, no. 3, p. 032329, 2018.