

# Resource Allocation for Rate and Fidelity Maximization in Quantum Networks

SHAHROOZ POURYOUSEF<sup>1,2</sup>, HASSAN SHAPOURIAN<sup>2</sup>, ALIREZA SHABANI<sup>2</sup>, RAMANA KOMPPELLA<sup>2</sup> AND DON TOWSLEY<sup>1</sup> (Member, IEEE)

<sup>1</sup>University of Massachusetts Amherst USA (email: {shahrooz,towsley}@cs.umass.edu)

<sup>2</sup>Cisco Research, San Jose, CA (email: {hshapour,ashabani,Kompella}@cisco.com)

Corresponding author: Shahrooz Pouryousef (email: shahrooz@cs.umass.edu).

**ABSTRACT** Existing classical optical network infrastructure cannot be immediately used for quantum network applications due to photon loss. The first step towards enabling quantum networks is the integration of quantum repeaters into optical networks. However, the expenses and intrinsic noise inherent in quantum hardware underscore the need for an efficient deployment strategy that optimizes the allocation of quantum repeaters and memories. In this paper, we present a comprehensive framework for network planning, aiming to efficiently distributing quantum repeaters across existing infrastructure, with the objective of maximizing quantum network utility within an entanglement distribution network. We apply our framework to several cases including a preliminary illustration of a dumbbell network topology and real-world cases of the SURFnet and ESnet. We explore the effect of quantum memory multiplexing within quantum repeaters, as well as the influence of memory coherence time on quantum network utility. We further examine the effects of different fairness assumptions on network planning, uncovering their impacts on real-time network performance.

**INDEX TERMS** Quantum Networks, Network Planning, Optimization, Utility, Repeater Allocation

## I. INTRODUCTION

The advent of the quantum Internet holds immense potential for realizing a wide array of transformative quantum applications, including quantum key distribution (QKD) [4], [12], [20], [26], [30], quantum computation [7], quantum sensing [10], clock synchronization [16], and quantum-enhanced measurement networks [12], among others [15]. One of the primary challenges to realizing such a large-scale quantum network lies in the transmission of quantum information through optical fiber over long distances, as photon loss increases exponentially with distance. To overcome this limitation, the concept of a quantum entanglement distribution network has been introduced [1], [5], [19]. The basic idea behind a quantum network is to strategically position a series of repeater stations along the transmission path [8], [22]. By leveraging the concept of entanglement swapping, long-range entangled qubits (in the form of Einstein-Podolsky-Rosen (EPR) pairs) between a pair of end users can be established. This process involves performing Bell state measurements at each intermediate node to effectively combine elementary link entanglements between adjacent repeaters. Once entanglement is established, quantum information can

be transmitted through quantum teleportation. Therefore, the successful execution of quantum Internet applications demands the development of novel protocols and the integration of quantum hardware, all aimed at establishing and maintaining reliable and high-fidelity entanglement across long distances in a quantum network [9], [18].

How do we ensure optimal performance of quantum networks in reliably delivering entanglement to the end users? Addressing this question requires a systematic approach, starting with quantum network planning. Similar to its classical counterpart, efficient resource management is crucial in quantum networks. In particular, quantum resources such as quantum repeaters and links must be carefully placed and optimized to meet the specific requirements of user pairs in real-time scenarios. To achieve effective quantum network planning, several key questions need to be addressed. First, determining the optimal number of quantum repeaters and their placement is essential to maximizing the success probability of end-to-end entanglement while maintaining fidelity. Additionally, allocating quantum memories at repeaters to users is crucial in achieving network fairness and ensuring efficient utilization of available resources. Furthermore, the

coherence times of quantum memories at both the end-user nodes and repeaters should be accounted for in network planning, as they impose upper bounds on the time frame available for classical communications.

In this paper, we formulate quantum network planning as an optimization problem. In short, the objective function is a quantum network utility and repeater locations as decision variables. Our utility function includes the rate and fidelity of the generated entanglement between end-users. The concept of network utility for classical networks made its debut in the influential research conducted by Kelly [13], [14]. There is a huge amount of research on network utility maximization in classical networks. Analog to classical network utility, the idea of quantum network utility maximization has been proposed in works such as [17], [27]. Given a quantum network, Vardoyan et al. [27] solve an optimization problem for finding the rates and link fidelities in order to maximize the utility function of a set of user pairs. However, in this paper, we start with planning the network for utility maximization.

We study how the following network parameters affect the optimal solution to our network planning optimization problem: number of end-user pairs, distance between network nodes (which can potentially be used as repeaters), repeater capacity (i.e., maximum number of quantum memories per repeater), and quantum memory coherence time. We use a quantum memory multiplexing approach [2], [24] to achieve higher end-to-end entanglement rates and treat memory allocation to different end-user pairs as part of our optimization problem. We find that the impact of multi-user demands on the end-to-end entanglement rate becomes more significant as the node distance is increased, while more end users may not necessarily imply the need for more repeaters. We observe that the requirement imposed on coherence time is much less restrictive for repeater memories than it is for end-node memories. Finally, we examine the planned network (i.e., the output of our optimization problem) at run-time given a random network traffic and show that its average performance is comparable to an unachievable upper bound.

The rest of the paper is organized as follows: In Sec. II, we introduce our network model and entanglement distribution protocol, and how to characterize the quantum network performance and utility in terms of rate and signal quality. We further explain what is the output of our network planning framework. In Sec. III, we present our network planning framework as an optimization problem and elucidate two ways of formulating the problem. We discuss why the optimization problem is nonlinear by definition and how we make it linear at the cost of neglecting some effects or introducing extra overhead. Sec. IV is devoted to several experiments where we apply our framework to various network topologies. Finally, we conclude in Sec. VI with some closing remarks and future directions. The derivation of the end-to-end entanglement generation rate in the presence of memory multiplexing and some additional optimization results are provided in three appendices.

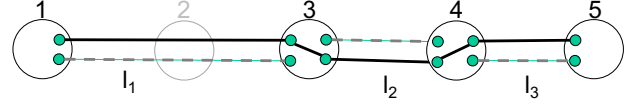


FIGURE 1: An example of quantum network planning for a linear chain with 3 potential locations for repeaters and a maximum of two memories per path. An instance of end-to-end entanglement generation is shown, where solid (dashed) lines represent successful (failed) attempts on links. The line connecting two memories inside a repeater indicates a successful Bell state measurement.

## II. NETWORK MODEL

We consider a quantum network represented by a graph  $G = (V, E)$ , where  $V$  is the set of nodes and  $E$  is the set of optical communication links. There are two types of nodes in the network: A set of nodes corresponding to the end users denoted by  $Q \subset V$  and a set of nodes denoted by  $R \subset V$  which provides potential locations for placing repeaters ( $V = Q \cup R$ ). We assume when a repeater is placed at a location with a degree of more than two, it acts as a quantum switch although we call it a repeater in this paper. There are  $|Q|/2$  number of user pairs that we want to maximize their utility function with respect to using at most  $N_{\max} \leq |R|$  repeaters. The unused nodes in  $R$  then operate as optical routers. We refer to the Einstein-Podolsky-Rosen (EPR) pairs or entanglement bits (ebits) generated along such links as link-level entanglement (or ebit). An end-to-end entangled state between a pair of users can be established using a process called *entanglement swapping*, that is to connect link-level ebits via Bell state measurements (BSMs) at the repeaters.

For example, suppose two nodes 1 and 3 in Fig. 1 share an ebit  $|\psi_{13}^+\rangle$ , and node 3 shares another pair  $|\psi_{34}^+\rangle$  with node 4. Then, node 3 can create an ebit  $|\psi_{14}^+\rangle$  between 1 and 4 by performing a BSM followed by a classical communication exchange. This operation is known as *entanglement swapping*. The process can be repeated to create ebits between distant parties 1 and 5. Table 1 shows the notations used in this paper.

### A. ENTANGLEMENT DISTRIBUTION PROTOCOL

Our entanglement distribution protocol is sequential based on the spatial multiplexing of quantum memories. A path between two users is said to have a width  $W$  when each end-user has  $W$  quantum memories, and each repeater node is equipped with  $2W$  quantum memories. The memories can be processed in parallel, and a BSM can be performed on any pair of quantum memories within each repeater. We assume BSMs are probabilistic with success probability  $q_s = 1/2$ .

The protocol starts with the sender who tries to prepare  $W$  EPR pairs and sends one end of each EPR pair through the optical link to the first repeater on the path to the receiver. Upon receiving the qubits from the sender, the first repeater sends an acknowledgment signal to the sender (which con-

$Q$	Set of user pairs
$R$	Set of potential locations for repeaters
$N_{\max}$	Number of repeaters budget
$D_u$	Number of memories at repeater $u$
$W_E$	Number of memories at end nodes
$T_{RM}$	Memory coherence time for repeaters
$T_{EM}$	Memory coherence time for end nodes
$q_s$	Success probability of Bell-state measurement
$p_l(l)$	Transmission probability of link with length $l$
$\tau_l(l)$	Returns the transmission delay on link with length $l$
$\tau_{e2e}(p)$	Returns the delay time to deliver end-to-end ebit on path $p$

TABLE 1: List of quantum network parameters.

tains the indices of qubits successfully received), prepares  $W$  EPR pairs, and sends one qubit of each EPR pair to the second repeater on the path. The second repeater similarly sends an acknowledgment signal to the first repeater and sends the first qubit of each  $W$  EPR pair to the third repeater. As soon as the first repeater receives the acknowledgment signal from the second receiver (which contains the indices of successful EPR pairs between the first and second repeaters), the first repeater makes BSM and releases the outcomes to the neighboring nodes. Then, the second repeater performs BSM after receiving the acknowledgment signal from the third repeater and the outcome of BSM in the first repeater. This process continues with the next repeaters until we reach the receiver on the other end. Finally, the receiver sends a sweeping acknowledgment signal to the sender. Figure 1 shows an instance of our protocol for a path with  $W = 2$ . The solid (dashed) lines indicate successful (failed) EPR trials and BSM is performed on successful links to generate an end-to-end EPR pair.

We assume a hard cut-off for the coherence time of quantum memories beyond which the memory is erased. As a result, the end-to-end entanglement may not be established due to the short coherence time of the repeater memories or those of end nodes. The optical link propagation time is calculated by  $\tau_l(l_{uv}) = l_{uv}/c$  where  $l_{uv}$  is the graph distance between the two nodes  $u$  and  $v$ . The time required to generate an end-to-end entanglement is denoted by  $\tau_{e2e}(\cdot)$  which includes the classical messages exchange between consecutive repeaters on a given path as explained above.

Consider a path with  $h$  links (corresponding to  $h - 1$  repeaters) and width  $W$ , where the success probability of link-level EPR pair on the  $i$ -th link is  $p_i = p_l(l_i)$  where  $i = 1, 2, \dots, h$  and

$$p_l(x) = 10^{-\alpha x}. \quad (1)$$

Here,  $x$  is the length of optical link (as an optical fiber) in km, and  $\alpha = 0.02$  is the signal attenuation rate in optical fiber (using 0.2 dB/km at Telecomm wavelength). The average end-to-end ebit generation rate (or throughput in short) can then be computed using a recurrence relation [24] (see

Appendix VIII-A for details). The recurrence relation leads to nonlinear equations characterizing the end-to-end rate which makes the objective function for our optimization problem nonlinear. Although there are ways to make our problem linear (as we explain in Sec. III), solving the recurrence relation for each path is time-consuming and takes longer for longer paths (with larger  $h$ ). They can easily add up to increase the overall time to compute the optimal solution. As a result, we approximate the average throughput by

$$R_{e2e}(p, W) = q_s^{h-1} \cdot W \cdot p_{\min}, \quad (2)$$

where  $p_{\min} = \min(p_1, p_2, \dots, p_h)$  is the minimum link-level success probability on the path. As we explain in Appendix VIII-A, this approximation is valid in the regime where  $W p_{\min} \gg 1$ . For reference, the end-to-end ebit rate associated with temporal or frequency multiplexing in a multimode memory corresponding to a path with  $W = 1$  is given by

$$R_{e2e} = q_s^{h-1} \prod_{i=1}^h (1 - (1 - p_i)^M), \quad (3)$$

where  $M$  is the multiplexing factor (see e.g., [1], [25] and references therein).

The quality of end-to-end ebits is often characterized by their fidelity. We assume that the link-level ebits are in the form of Werner states with fidelity  $F_L$ ; as a result, the end-to-end fidelity is

$$F_{e2e} = \frac{1}{4} + \frac{3}{4} \left( \frac{P_2(4\eta^2 - 1)}{3} \right)^{h-1} \left( \frac{4F_L - 1}{3} \right)^h, \quad (4)$$

where  $P_2$  is the two-qubit gate fidelity and  $\eta$  is the measurement fidelity of the swapping operation [5]. We assume  $P_2 = \eta = 1$  in our experiments.

Regarding the scheduling of link-level entanglement generation, one may consider a parallel protocol where the main difference with our sequential protocol above is that all repeaters on a path start generating link-level entanglement simultaneously. Such a parallel protocol gives the same end-to-end success probability as Eq. (2) while it can reduce  $\tau_{e2e}$ , ultimately leading to larger ebit rate per unit time,  $R_{e2e}/\tau_{e2e}$ . This is however at the expense of longer run times for repeater memories since regardless of the link-level synchronization protocol the BSMs must be performed sequentially from sender to receiver. In other words, a given repeater needs to know the indices of successful BSMs in previous steps to determine which quantum memories of theirs are entangled with the sender's memories. We imagine a future quantum network to have lower-quality memories (with shorter coherence time) inside repeaters (i.e., network core) and high-quality memories (with longer coherence time or possibly fault-tolerant) at the end users (i.e., network edge). Therefore, we adopt the sequential protocol as it imposes a less strict requirement on the coherence time of repeater memories. To increase the end-to-end ebit rate per unit time, we can increase  $R_{e2e}$  by increasing the path width  $W$  (c.f. Eq. (2)).

## B. OBJECTIVE

The objective of our network planning optimization problem is to maximize the aggregate utility of the set  $Q$  of user pairs. The quantum utility function of a user pair is defined as

$$U(R_{e2e}, F_{e2e}) = \log_2(R_{e2e} \cdot f(F_{e2e})), \quad (5)$$

in terms of the end-to-end rate  $R_{e2e}$  and fidelity  $F_{e2e}$  of the EPR pairs delivered to them. This is the *negativity* quantity proposed in [28] to quantify the degree of entanglement in composite systems. Other possible utility functions are distillable entanglement [23] and secret key rate [3] as explored in [27]. Here, the functional form of  $f(F_{e2e})$  depends on the application and takes different forms for computing [17] and networking, or secret sharing [3]. In this paper, we use the following formula based on the entanglement negativity [28],

$$f(F) = F - \frac{1}{2}, \quad (6)$$

as a proxy for the quality of the end-to-end ebits, since it is an upper bound on the distillable entanglement [21]. The utility function based on *negativity* is preferred as it is concave and one can use convex optimization techniques to efficiently find the optimal value [27].

We note that the utility function defined in (5) does not necessarily favor more repeaters. This is not only because the end-to-end fidelity (4) decreases as we add more links (or increase  $h$ ) but also because the overall swapping success probability decreases in the end-to-end rate (2). Therefore, even if we set  $F_L = 1$  (which implies  $F_{e2e} = 1$ ) and neglect the impact of fidelity the optimal solution may only use a fraction of potential locations for repeaters. Based on this observation, we omit the fidelity from the utility function so that we can reduce our link-based formulation to an integer linear programming.

## C. PLANNING OUTPUT

The output of the optimization problem provides four results: (1) the number of repeaters to be used, (2) where to place them in the network, (3) the paths for each user pair, and (4) the assigned quantum memories at the repeaters to different paths. In some of our experiments, we also estimate the coherence time required for repeaters and user node memories separately.

### 1) Definition of a path

A path is a sequence of repeaters' locations. Two consecutive locations on the path are not necessarily two consecutive nodes on the actual graph. We assume there is a direct physical link between each two locations with the length of the shortest path between them. This allows us to have more than one path between two end nodes even on a repeater chain. For example, in Figure 1, we can have paths  $[1, 5]$ ,  $[1, 3, 5]$ ,  $[1, 2, 3, 4, 5]$  and  $[1, 3, 4, 5]$  etc. The path  $[1, 5]$  means none of the locations have been decided to become a repeater and that means no repeater would be used. In this figure, we have the path  $[1, 3, 4, 5]$  which means there is a

$P^q$	Set of all paths for user pair $q$ in $G$
$K$	Number of allowed paths for each user pair
$r_u$	Indicates whether node $u \in R$ is used as a repeater node or not
$x_p^q$	Indicates whether path $p$ is used for user pair $q$ or not
$w_p^q$	Width of path $p$ for user pair $q$

TABLE 2: List of variables used in the path-based formulation.

direct physical link between node 1 and node 3. The length of this link is the summation of the link that connects node 1 to 2 and node 2 to node 3.

Figure 1 illustrates an example of our optimization. It is a linear chain with nodes 1 and 5 as users and 3 potential places for repeater placement: nodes 2, 3, and 4. The optimal solutions places two repeaters at nodes 3 and 4. Note that node 2 is grayed out which means this node will not be used as a repeater but rather an optical router providing an optical link between 1 and 3. In this example, since there is only one user pair, the optimal solution is to assign both memories to this user pair to maximize the end-to-end ebit rate.

Before closing this section, let us make a few remarks on related previous work. A similar idea for network planning but with one multi-mode memory per channel (c.f., Eq. (3)) has been proposed in [22]. Our work is similar to their work as we also use the preexisting infrastructure for network planning. However, our goal is to maximize the network utility, which favors short paths with few repeaters. In addition, we consider a different type of quantum memory scheme using spatial multiplexing (c.f., Eq. (2)) and analyze the effect of the finite coherence time of quantum memories. In contrast to Ref. [22] which uses equally-distanced repeaters to estimate the end-to-end entanglement rate of a given path (regardless of the repeater positions), we evaluate the entanglement rate for each path specifically based on the exact location of the repeaters.

## III. OPTIMIZATION PROBLEM

In this section, we present two equivalent ways of formulating the quantum network planning problem and discuss how we turn them into integer (binary) linear programs.

### A. PATH-BASED FORMULATION

Equations (7)-(16) define our path-based network planning optimization problem. We assume each user pair uses only one path but our approach can be easily extended to multiple paths per user pair. Let  $P^q$  denote the set of paths for user pair  $q$  and  $r_u \in \{0, 1\}$  be whether node  $u$  is used as a repeater node or not. Given path  $p$ , the end-to-end throughput  $R_{e2e}(\cdot)$  and fidelity  $F_{e2e}(\cdot)$  are computed using Eqs. (2) and (4), respectively. Decision variables are  $x_p^q$  and  $w_p^q$  which indicate the path that should be used for user pair  $q$  and the width of path  $p$ , e.g., the number of memories to deploy on the entire path  $p$  (width of path  $p$ ) for source-destination pair  $q$ . This



in turn implies which nodes to be used as repeaters where  $r_u = 1$ .

**Problem 1 (path-based problem formulation)**

$$\max_{r_u, x_p^q, w_p^q} \sum_{q \in Q} \sum_{p \in P^q} U(Re_{2e}(p, w_p^q), Fe_{2e}(p)) x_p^q \quad (7)$$

s.t.

$$\sum_{\substack{q \in Q \\ p \in P^q | u \in p}} w_p^q \cdot x_p^q \leq D_u r_u \quad \forall u \in R \quad (8)$$

$$\sum_{p \in P^q | u \in p} x_p^q \leq K \quad \forall q \in Q \quad (9)$$

$$\sum_{p \in P^q | u \in p} w_p^q \cdot x_p^q \leq W_E \quad \forall q \in Q \quad (10)$$

$$\sum_{u \in R} r_u \leq N_{\max} \quad (11)$$

$$2\tau_l(l_e) \cdot x_p^q \leq T_{RM} \quad \forall q \in Q, \forall p \in P^q, \forall e \in p \quad (12)$$

$$\tau_{e2e}(p) \cdot x_p^q \leq T_{EM} \quad \forall q \in Q, \forall p \in P^q \quad (13)$$

$$x_p^q \in \{0, 1\}, \quad \forall q \in Q, p \in P^q \quad (14)$$

$$r_u \in \{0, 1\}, \quad \forall u \in R \quad (15)$$

$$w_p^q \in \{1, 2, 3, \dots, \min(D, W_E)\}, \quad \forall q \in Q, p \in P^q \quad (16)$$

Constraint (8) is about using at most  $2D_u$  memories at network node  $u$  which is selected as a repeater. We assume the maximum number of memories for all deployed repeaters is identical, i.e.,  $D_u = D$ . Constraint (9) ensures that only  $K$  paths be used for each user pair. Constraint (10) enforces the memory limit of end nodes and constraint (11) checks that we use at most  $N_{\max}$  repeaters in the network.  $w_p^q$  indicates the width of the path  $p$  for user pair  $q$ . Constraint (12) ensures that for each path decided to be used in the network and for all optical links on that path, the time required to generate entanglement and receive the acknowledgment signal for ebit generation must be less than or equal to the memory coherence time of the repeaters. Constraint (13) forces the time required for end-to-end entanglement generation to be less than the end-node memory coherence time for a selected path.

The above problem formulation has two drawbacks. First, the objective function as defined in Sec. II-A is nonlinear which makes the problem an integer non-linear program. We also have a product of two decision variables in memory constraints (8) and (10). We resolve this issue by enumerating all the versions of each path (including possible values for the path width) and computing the nonlinear utility function. Second, it is not practical to enumerate all paths for large networks (which implies  $|Q| \min(D, W_E) |R|!$  variables for

$x_p^q$ ) since this scales exponentially with the number of nodes  $|R|$ . In order to resolve this issue, we note that we may not need to enumerate all the paths in the network to obtain either the optimal or a near-optimal solution. Instead, we use the algorithm proposed in [31] to find the first  $k$  shortest paths and run our path-based optimization algorithm (7) on the reduced set. As we show in the evaluation section, we can reach the optimal solution by limiting the number of decision variables to  $|Q| \min(D, W_E) k$  where  $k = 1000 - 4000$  for random networks with  $|R| \leq 50$  (Appendix VIII-B) and dumbbell topology with  $|R| \leq 10$  (Sec. IV-A). Alternatively, as we explain next, one can formulate a link-based version of the same problem where the number of decision variables scales polynomially with the number of network nodes.

## B. LINK-BASED FORMULATION

Here, we present a link-based formulation of the quantum network utility maximization problem. For each user pair  $q = (s, t)$ , we define an array of binary variables  $x_{uv}^q$  associated with each directed link  $(u, v) \in \mathcal{E}_q$  where the set of links is defined as

$$\mathcal{E}_q = \{(u, v) | u \in R \cup \{s\}, v \in R \cup \{t\}, u \neq v\}. \quad (17)$$

An end-to-end path is described by a subset of  $x_{uv}^q$  which are non-zero. Constraint (19) is the flow continuity equation (similar to the maximum flow problem) to ensure that there is a directed path between the sender  $s$  and receiver  $t$ . For instance, the solution in Figure 1 for  $q = (1, 5)$  corresponds to  $x_{13}^q = x_{34}^q = x_{45}^q = 1$  with other entries being zero.

$\mathcal{E}_q$	Set of all pairs of potential repeater nodes and pairs of repeater nodes and end users $q$ in $G$
$x_{uv}^{q,w}$	Indicates whether the link $(u, v)$ with width $w$ is used as part of a path for $q$ user pair or not
$d_q$	Longest link for a path connecting user pair $q = (s, t) \in Q$
$\beta_{q,w}$	Indicates whether path with width $w$ is used for user pair $q$ or not

TABLE 3: List of variables used in the link-based formulation.

$$\max \sum_{q=(s,t)} [\log_2(q_{sw}) \left( \sum_{w,(u,v) \in \mathcal{E}_q} x_{uv}^{q,w} - 1 \right) + \sum_w \beta_{q,w} \log_2(w) - \alpha_2 d_q] \quad (18)$$

$$\text{s.t.} \quad \sum_{v,w} x_{uv}^{q,w} - x_{vu}^{q,w} = \begin{cases} 1, & \text{if } u = s \\ -1, & \text{if } u = t \\ 0, & \text{if } u \in R \end{cases} \quad (19)$$

$$\sum_w x_{uv}^{q,w} \leq 1 \quad \forall (u, v) \in \mathcal{E}_q, \forall q \in Q \quad (20)$$

$$\beta_{q,w} = x_{st}^{q,w} + \sum_v x_{sv}^{q,w} \quad \forall w, \forall q \in Q \quad (21)$$

$$d_q \geq l_{uv} x_{uv}^{q,w} \quad \forall (u, v) \in \mathcal{E}_q, \forall q \in Q \quad (22)$$

$$\sum_{w,q,v} w x_{uv}^{q,w} \leq D_u r_u \quad \forall u \in R \quad (23)$$

$$\sum_{w,v} w x_{sv}^{q,w} \leq W_E \quad \forall q \in Q \quad (24)$$

$$\sum_u r_u \leq N_{\max} \quad \forall u \in R \quad (25)$$

$$2\tau_l(l_{uv})x_{uv}^{q,w} \leq T_{RM} \quad \forall w, \forall (u, v) \in \mathcal{E}_q, \forall q \in Q \quad (26)$$

$$3 \sum_{(u,v)} \tau_l(l_{uv})x_{uv}^{q,w} \leq T_{EM} \quad \forall w, \forall q \in Q \quad (27)$$

$$x_{uv}^{q,w} \in \{0, 1\} \quad \forall w, \forall (u, v) \in \mathcal{E}_q \quad (28)$$

$$r_u \in \{0, 1\} \quad \forall u \in R \quad (29)$$

$$\beta_{q,w} \in \{0, 1\} \quad \forall w, \forall q \in Q \quad (30)$$

The objective function (18) is the aggregate utility (5) where we rewrite the end-to-end ebit rate Eq. (2) using the decision variables  $x_{uv}^{q,w}$ . We do not consider fidelity in the objective function for link-based formulation. We recall that

the role of the fidelity term is to penalize overusing repeaters, and we still have another term, namely, the overall swap success probability in the end-to-end rate (2) to enforce that. Since the dependence of the utility function on the path width  $w$  is nonlinear (i.e.,  $\log_2 w$ ), we cannot use  $w$  as a decision variable and maintain a linear programming problem. Hence, we introduce  $W_E$  copies of  $x_{u,v}^{q,w}$  and include  $w = 1, \dots, W_E$  as a superscript and auxiliary variable  $\beta_{q,w}$  defined in (21) is an array of size  $W_E$  where the only non-zero element determines which value of  $w$  is used. We impose constraint (20) to ensure that only one path (out of  $W_E$ ) will be chosen. The summand in the objective is  $\log_2 R_{e2e}$  as defined in (2) and can be understood as follows: the first term is  $\log_2 q_{sw}^{h-1}$  where we rewrite the number of active links as a sum over all entries of  $x_{u,v}^{q,w}$ . The second term accounts for which value of path width is used and the last term is the minimum success probability (1) on a path after taking the logarithm, i.e.,  $\log_2 p_{\min} = \log_2 10^{-\alpha d_q} = -\alpha_2 d_q$  where  $\alpha_2 = \alpha \log_2 10$ , and  $d_q$  gives the longest link on the path (calculated via constraint (22)).

Let us now discuss the remaining constraints in the link-based formulation. Constraints (23), (24), and (25) are identical to constraints (8), (10), and (11) in the path-based formulation which imposes repeater memory, end-user memory, and a maximum number of repeater constraints, respectively. Constraint (26) is analogous to (12) in the path-based approach and does not allow links where the signal round trip takes longer than the repeater memory coherence time. Lastly, constraint (27) ensures the end-to-end entanglement distribution process does not take longer than the memory coherence time at the end users. We note that (27) has the benefit of being a linear constraint at the cost of being more stringent than (13) in the path-based formalism.

We note that the link-based formulation reduces the problem size (i.e., number of entries in  $x_{uv}^{q,w}$ ) to  $|Q| \min(W_E, D)[|R|(|R| + 1) + 1]$  which is significantly smaller than the path-based approach.

Although we do not use entanglement distribution protocols based on frequency or time multiplexing in our simulations, it is worth noting that the utility function associated with the rate in this case (3) can also be written as a linear function,

$$\log_2 R_{e2e} = \log_2(q_{sw}) \left( \sum_{(u,v) \in \mathcal{E}_q} x_{uv}^q - 1 \right) + \sum_{(u,v) \in \mathcal{E}_q} x_{uv}^q \log_2 (1 - (1 - p_l(l_{uv}))^M), \quad (31)$$

where superscript  $w$  is dropped since this multiplexing scheme assumes one memory per channel.

### C. SCALE INVARIANCE AND EQUIVALENCE BETWEEN THE TWO FORMULATIONS

One way to prove the equivalence of the two formulations is to show that an optimal solution in each of the formulations maps to a valid solution in the other formulation [6], [22]. It

is easy to see why. Suppose the optimal utility function for path-based and link-based schemes are denoted as  $U_p$  and  $U_l$ . An optimal path  $p$  in the path-based formulation consists of some links connecting the two end users, which in the link-based formulation corresponds to setting those entries in  $x_{uv}^{q,w}$  one and keeping the rest zero. The path  $p$  is then a valid solution to the link-based formulation since all the constraints in either formulation are equivalents. Hence, we have  $U_p \leq U_l$ . Similarly, the optimal set of activated links given in terms of the array  $x_{uv}^{q,w}$  can be viewed as a path where only  $u, v$  nodes with  $x_{uv}^{q,w} = 1$  are being used. Therefore, we can write  $U_l \leq U_p$ . The two inequalities have to be satisfied simultaneously which implies  $U_l = U_p$ , i.e., the two optimal solutions are identical.

We note that either formulation of the problem enjoys a scale invariance property as follows: The problem does not change as we rescale repeater capacity  $D \rightarrow \lambda D$ , end user capacity  $W_E \rightarrow \lambda W_E$ , and  $w \rightarrow \lambda w$  by a scaling factor  $\lambda$ . This is because the network capacity constraints (23), (24), (8), and (10) remain the same after the rescaling and the objective function is shifted by a constant  $|Q| \log_2 \lambda$  (which can be removed). Therefore, the optimal solution remains the same and the optimal number of memories for each pair scales the same way  $w_{\text{opt}} \rightarrow \lambda w_{\text{opt}}$ . This means that only relative ratios are relevant, i.e., which portion of repeater memories  $\frac{w_{\text{opt}}^q}{D}$  are assigned to user pair  $q$ . For instance, if we have two user pairs and the optimal solution for  $D = 10$  is  $w_{\text{opt}}^{q_1} = w_{\text{opt}}^{q_2} = 5$ , it means that if we solve the problem for  $D = 1000$ , then we simply have  $w_{\text{opt}}^{q_1} = w_{\text{opt}}^{q_2} = 500$ .

The scale invariance is an important property of our formulation for the following reason: Suppose we run an optimization problem for a small value of repeater capacity  $D = 10$  so that the problem size is small and manageable and find the optimal path with  $w_{\text{opt}}^q = 3$  to have the longest link of length 100km. This solution violates our approximation for the end-to-end ebit rate (2) which requires  $w_{\text{opt}}^q p_{\text{min}} \gg 1$  while we have  $w_{\text{opt}}^q p_{\text{min}} = 0.03$ . Thanks to the scale invariance property, we can say our solution,  $\frac{w_{\text{opt}}^q}{D} = 0.3$ , is still valid for  $D \gtrsim 1000$  which implies the minimum number of memories to be  $w_{\text{opt}}^q = 300$ . We use this fact when we run quantum network planning for the ESnet.

#### IV. EVALUATION

In this section, we report some insights from our experiments. We use a synthetic (dumbbell-shape geometry) and two real-world topologies (SURFnet and ESnet) as our physical topologies. We use IBM CPLEX solver to solve the linearized version of our optimization problem 7 and 18. We assume the entanglement swapping success probability is 0.5, and the fidelity of a link-level ebit is 0.95 unless stated otherwise.

##### A. SYNTHETIC TOPOLOGY

Our synthetic topology is a dumbbell-shape geometry shown in Figure 2(a). In this topology, there are  $n$  symmetric user

pairs connected through a backbone link. In Figure 2(a) node 1 is paired with node 2, node 3 is paired with node 4 and so on. The length of the link connecting each node to the closer end of the backbone link is 1km. We vary the length of the backbone link in this experiment.

We use our path-based formulation (7) in this experiment as it includes the end-to-end fidelity in the utility function. As mentioned in the previous section, here we use  $k$  shortest paths algorithm and consider  $|P^q| = 4,000$  paths for each user pair and  $K$  is one. Please note that we can have more than one path in a repeater chain based on our definition of a path in section II-C1. We set  $W_E = D = 100$ , and do not impose constraints on the memory coherence time of repeaters or end nodes in this experiment.

##### 1) Utility vs. Distance between repeaters

As the first experiment, we show how the utilities of user pairs change as we increase the distance between potential places for repeaters. For this, we consider  $|R| = 10$  locations for repeaters at equal distances  $L/(|R| + 1)$  along the backbone link with length  $L$  as shown in Figure 2(a). The distance between the potential repeater locations is increased uniformly by increasing  $L$ , and we solve the optimization problem for each value of  $L$ .

Figure 2(b) shows how the optimal end-to-end entanglement rate for each user pair varies as we increase the backbone link distance. When there is only one user pair in the network, all memories available on the repeaters are assigned to that user pair and it receives a high rate compared to cases where we have more than one user pair. In the presence of more user pairs, user pairs share repeaters and receive fewer memories to maximize the aggregate utility function Eq. (5).

##### 2) Fidelity/Number of used repeaters vs. Distance between repeaters

Figures 2(c) and 2(d) show the optimal end-to-end fidelity and the number of repeaters used in the network out of our 10 repeater budget as a function of backbone link length. There is an inverse correlation between the number of repeaters and the average end-to-end user pairs fidelity. This is expected based on Eq. (4) as end-to-end fidelity on a path decreases as more repeaters are used. When the backbone link length is small (less than 40km), no repeaters are used and there will be a direct link between the end nodes. As we increase the backbone link length, link-level ebit generation success probability decays exponentially and more repeaters are allocated to increase link-level generation success probabilities. As is evident from the plot, the optimal solution never utilizes all 10 available repeater locations in the network.

##### B. ESNET TOPOLOGY

In this experiment, we use the ESnet topology [11] and examine how repeater placement on this network affects utility maximization. We have derived the geographical locations of the nodes from [11] and estimated link lengths in terms of their geodesic distances. We focus on the East Coast and

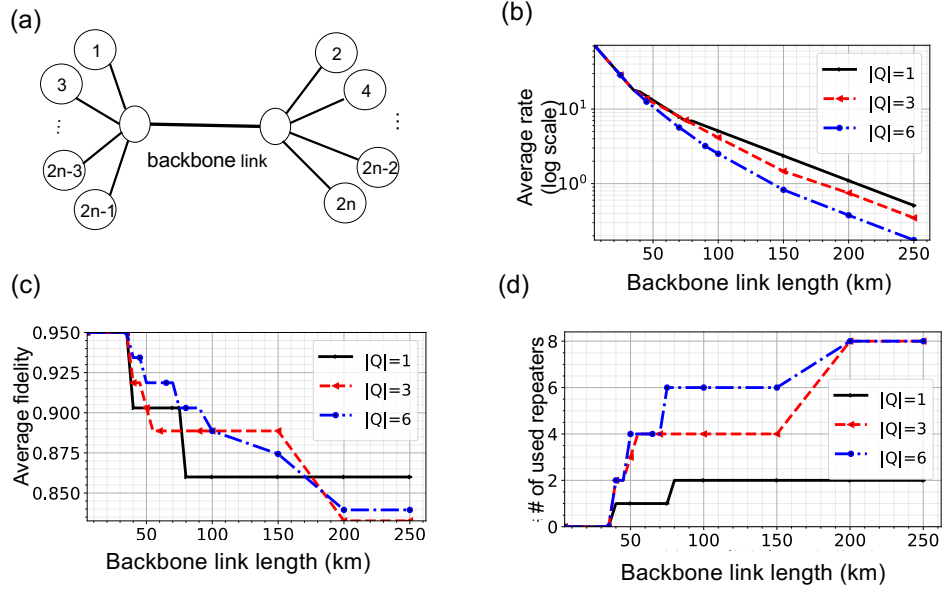


FIGURE 2: (a) Dumbbell topology with  $|Q| = n$  user pairs, (b) optimal ebit rate per user pair, (c) optimal end-to-end fidelity, and (d) number of used repeaters as a function of the backbone link length.

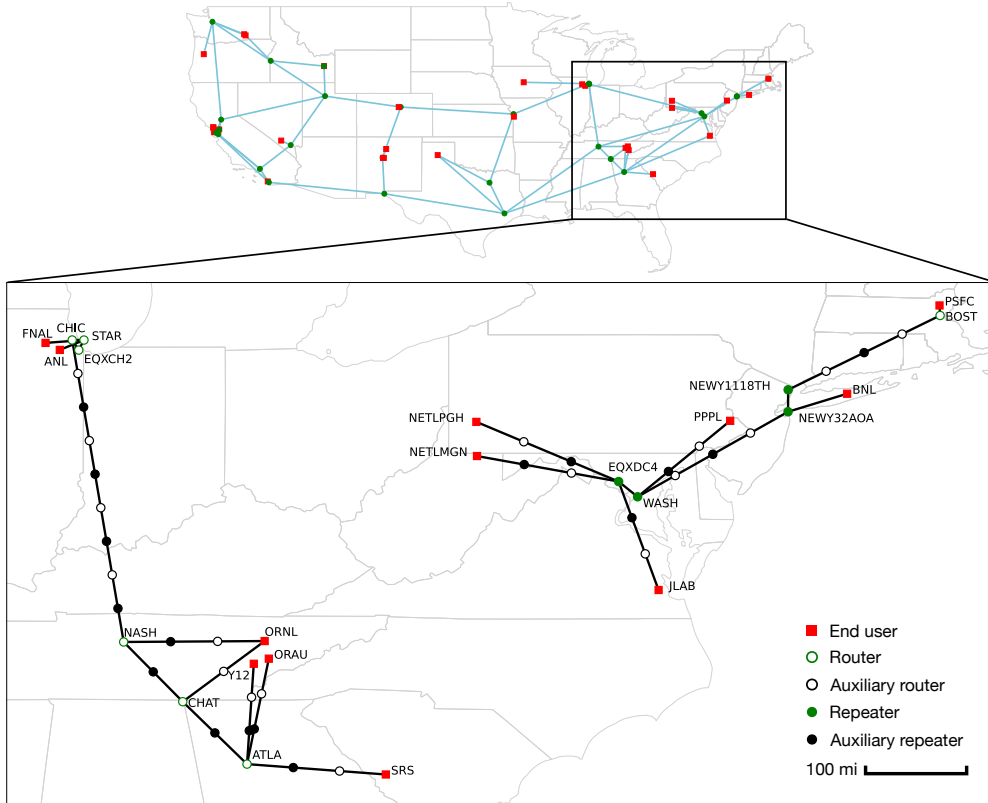


FIGURE 3: Optimal locations of repeaters for the augmented subgraph of the ESnet including nodes in the East Coast and the Midwest. The black circles (open and filled) denote the auxiliary nodes placed to make the longest link 100km long. The optimization solution is shown as filled circles which indicate the locations of nodes turned into repeaters while open circles are not used. Some end nodes are shifted to improve readability.



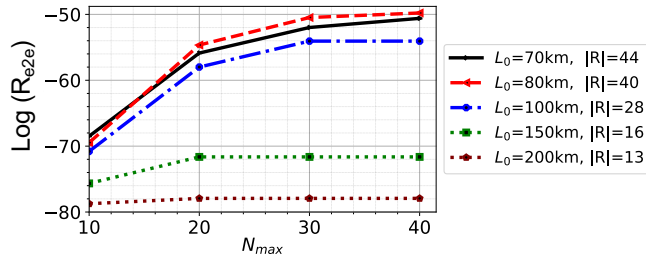


FIGURE 4: ESnet network planning on the ESnet augmented network graph where we place additional repeaters to upper bound the maximum link length  $L_0$  (see main text for details). The legend also shows the number of potential repeater locations after the augmentation. Here, we set the memory capacity of repeaters and end users to be  $W_E = D = 10$ .

the Midwest shown in Figure 3 and consider three user pairs in each region. The ESnet core and edge nodes are shown as red squares and green circles in the upper panel of this figure. Since the original links are long (greater than few hundred kilometers), we have augmented the network graph by adding auxiliary nodes so that no link is longer than  $2L_0$  (to be specified for each experiment). We achieve this in the following way: Given a link of length  $\ell > 2L_0$ , we place  $n_\ell = \lfloor \frac{\ell}{L_0} \rfloor - 1$  repeaters.

Figure 4 shows how the optimal aggregate utility changes as we increase the repeater budget  $N_{\max}$  for different values of  $L_0$ . We observe that increasing  $N_{\max}$  for a fixed value of  $L_0$  initially improves utility but eventually saturates. This illustrates competition between repeater spacing and number of repeaters in the optimal solution (c.f. Eq. (2)) where adding more repeaters may increase link-level success probabilities but the end-to-end ebit rate decreases overall due to the lower swap success probabilities. The fact that the saturation occurs for small values of  $N_{\max}$  depends on the details of the network topology. We further see that decreasing  $L_0$  from 200km to 100km increases the optimal aggregate utility but the improvement diminishes as we further decrease  $L_0$  below 100km.

The lower panel of Figure 3 shows an example of the optimal solution where the graph is augmented with repeater locations no more than  $L_0 = 70$  km apart (added nodes are shown as filled and open black circles). With this value of  $L_0$ , we observe that the longest link length is 100km. The result of optimization for the following set of user pairs are shown: (SRS, ORAU), (Y12, FNAL), (ORNL, ANL), (NETLPGH, PSFC), (NETLMGN, PPPL), and (BNL, JLAB). Here, we use our link-based LP formulation (18) with link fidelity equal to one. In this case, after the augmentation, we have  $|R| = 44$ , and we set  $N_{\max} = 10$  for each region. We further show the individual paths for each user pair explicitly in Appendix VIII-C.

### C. SURFNET TOPOLOGY

In this experiment, we show how quantum memory coherence time and memory capacity at repeaters and end nodes affect the optimal quantum utility of the network. We use the SURFnet topology (Figure 5(a)) in the next two experiments. We consider a set of 4 user pairs randomly selected in the network as a workload. In this experiment, we choose user pairs with distances in the range of 200 and 250 km from each other. For Figure 5(b) we plot the average of 100 different workloads and for Figure 5(c) and 5(d) we consider only one workload. We assume we can use  $N_{\max} = 10$  repeaters each with  $D = 100$  memories across the network. Each node in the SURFnet topology is a potential location for a repeater.

#### 1) Utility vs. Memory coherence time

Figure 5(b) shows the aggregate utility of the user pairs in SURFnet topology as a function of the memory coherence times of repeaters and end nodes. The  $x$ -axis is end nodes memory coherence time and the  $y$ -axis is repeaters memory coherence time both in milliseconds ( $ms$ ). When end node memory coherence time is less than  $3.2$   $ms$ , using repeaters with high-quality memories (memories with a long coherence time for qubits) does not further improve the utility. This is because end node memory coherence time does not support holding qubits for entanglement generation and receiving the heralding signal across any path (even the shortest path). In this experiment, we set the aggregate utility to  $-50$  when there is no solution for our optimization problem. When end node memory coherence time is above  $3.5$   $ms$ , as we increase the coherence time of memories at repeaters, we can handle longer links which could be favored by the solver over shorter links since such paths have fewer links leading to larger end-to-end fidelity.

#### 2) Utility vs. Memory capacity

Here, we show how increasing the memory capacities of end nodes or repeaters in the network affects the aggregate utility of the user pairs in the workload. Figure 5(c) shows as we increase the memory capacity of the repeaters in the network (core nodes), the utility increases. However, it will not affect the results after we reach 100 memories per end node in Figure 5(c)). The same observation is true for the case the repeater nodes' memory capacity is fixed  $D = 100$  and we increase the memory size at the end nodes (Figure 5(d)).

#### 3) Planning assumptions

In this part, we conduct an experiment to show how different assumptions at the network planning stage can affect the performance of the network at runtime (e.g., operation time). We can plan the network based on different assumptions about the network workload at runtime. Each workload is a set of user pairs. We first plan the network for a specific workload. We call this workload the planning workload. Then, at runtime, we assume a time-slotted model where at each time slot different workloads are considered in the network. We call these workloads runtime workloads. The

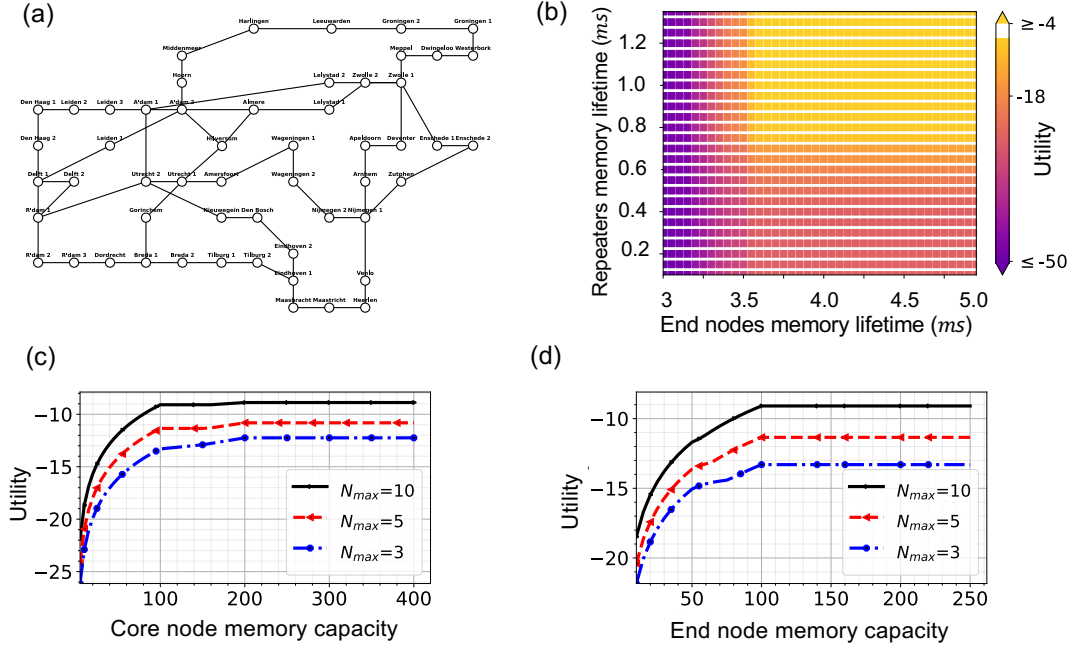


FIGURE 5: (a) SURFnet topology, (b) utility as a function of memory coherence time for 100 different sets of  $|Q| = 4$  randomly selected user pairs, (c) utility as a function of core node memory capacity with user memory fixed  $W_E = 100$ , and (d) utility as a function of end node memory capacity with fixed repeater capacity  $D = 100$  for  $|Q| = 4$  user pairs in SURFnet.

user pairs in the workloads of runtime are from the user pairs in the planning workload. The probability of each user pair from the planning workload in each workload of runtime depends on their probability.

Here in this experiment, we consider two different scenarios. In the first scenario, we assume the probability of each user pair in the planning workload is equal. In this case, the probability of each user pair in the planning workload to be in the run time workloads is equal. The second scenario is when the probability of user pairs of planning workload is different. We use a weight for each user pair to indicate the probability of having that user pair in the runtime workloads. The objective function in this experiment for the network planning optimization problem is maximizing the weighted aggregate of the utilities.

In this experiment, we assume after we plan the network, the place of the repeaters and the paths that connect each user pair are fixed and we will use this setting for the network operation time or runtime. The only thing that can be modified at the network operation stage is the number of assigned memories to each path for each user pair.

For each scenario, we simulate 300 events of a Poisson process with different values of the mean for the distribution. The mean value indicates the average number of user pairs in the workload. The list of user pairs is 10 and we repeat this experiment for different planning workloads. For each event among 300 events, we choose the user pairs based on the assumption about their weight. For example, if the weight of all the user pairs is equal, we randomly choose that many numbers of user pairs from the set of our user

pairs in the planning workload. When the probability of user pairs is different, we choose the user pairs from the planning workload in proportion to their weight. In this case, the network would be planned based on the user pairs with higher weights and the repeaters may be placed in between user pairs that have higher weight (probability).

Figure 6 shows the aggregate utility in the network for the two different scenarios. The mean value (x-axis in Figure 6) indicates the mean number of user pairs per time unit (time slot). The blue line indicates the case that we perform network planning for each received workload at any given time. We use this scheme as a reference indicating the upper bound performance, although it is unrealistic to imagine a network topology change in real-time based on the network workloads. While this approach is not practical, it shows the optimal aggregate utility that we can have for each set of user pairs at each given time if we plan the network instantaneously for the workload at that time. The black lines correspond to the quantum network utility evaluated as the output of the optimization problem. As we see in Fig. 6 (a) and (b), the numerical values are different as the optimization involves different assumptions about the probability of planning workload user pairs to be in the run-time workload. The red lines show the aggregate utility of the network by simulating how the demands are handled on a static network design based on the solution of the optimization problem with optimal locations for repeaters and paths for the user pairs. In both cases, the planned network performance is comparable with the upper bound. For the first scenario, there is a gap between the aggregate utility at network operation time and

# Planning assumptions

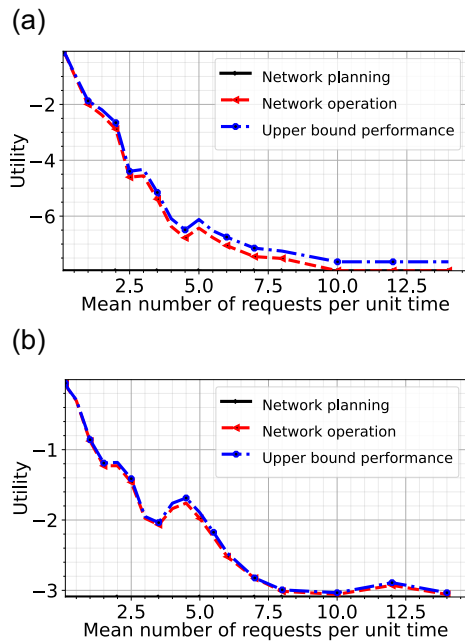


FIGURE 6: The effect of assumptions at the network planning stage on the network performance at the operation time.

the upper bound value, whereas we do not have this gap for the second scenario. The reason is that in the second scenario, the resources have been planned to serve the user pairs with higher weight and since the probability of having a user pair with higher weight is high, there is no gap for the case we plan the network considering different weights for different user pairs.

## V. RELATED WORK

In this section, we overview the current state of research on quantum network planning and quantum network utility maximization.

**Repeater placement:** Rabbie et al. [22] proposed the idea of quantum network design using the preexisting infrastructure. They formulate the problem of satisfying a certain rate and fidelity threshold for a set of user pairs with a minimum number of repeaters in the network as an optimization problem. A similar idea to our approach for network planning but with one multi-mode memory per channel has been proposed in [8]. Our work is similar to their work as we also use the preexisting infrastructure for network planning. However, our goal is to maximize the network utility which favors shortest paths with fewer repeaters. In addition, we consider a different type of quantum memory scheme using spatial multiplexing and analyze the effect of finite coherence time of quantum memories. In contrast to Ref. [22] which uses equally-distanced repeaters to estimate the end-to-end entanglement rate of a given path (regardless of the repeater positions), we evaluate the entanglement rate for each path specifically based on the exact location of the repeaters. We further study the performance of our planned network at

runtime.

**Quantum network utility:** The idea of quantum network utility maximization has been introduced in [27]. Vardoyan et al. [27] solve an optimization problem for finding the rates and link fidelities in order to maximize the utility function of a set of user pairs in a quantum network. They assume a centralized solver knows the topology and the location of each repeater as well as the utility function of user pairs. Their result elucidates a trade-off between the end-to-end entanglement generation rate and the fidelity. Lee et al. [17] introduce a framework to quantify the performance and capture quantum networks' social and economic value. They develop an example of an aggregate utility metric for distributed quantum computing that extends the quantum volume from single quantum processors to a network of quantum processors. Although we use a similar formula for the network utility, we are addressing a separate issue (that is the network planning). Furthermore, our approach of modeling the quantum network is quite different. Both references model the entanglement distribution network in terms of entanglement flows along the network elementary links, which can be justified in the regime where there are infinite number of memories per channel and/or memories have infinite coherence time. In contrast, we use a physical model based on spatial multiplexing of quantum memories where the link-level entanglement generation rate can be derived explicitly based on the number of memories and the channel transmission rate. This approach in turn lets us simulate the network dynamics in an actual real-time scenario.

## VI. CONCLUSIONS

This paper introduces a comprehensive network planning framework designed to efficiently distribute quantum hardware within the existing infrastructure, aiming to maximize the utility of the quantum network. We investigate the impact of memory coherence time at the repeaters and end nodes on network planning strategies. Additionally, we analyze the influence of different fairness assumptions made during the network planning stage on the network's performance during runtime. Our findings reveal that the coherence time requirement for quantum memories is significantly less restrictive for repeater memories compared to those of end users.

Our optimization results on real-world examples suggest that spatial multiplexing would lead to reasonably a high end-to-end ebit rate while not imposing a huge demand on quantum memory coherence time (e.g. sub 10ms). A promising technology to this end is on-chip quantum memory candidates such as vacancy color centers [29].

In the context of optimization problems, there are several avenues for future research. We consider a quantum network utility function based on entanglement negativity as the objective function in our optimization problem. It would be interesting to consider other objective functions for different purposes such as distributed quantum computing [17] or quantum key distribution [27] and see how the optimal solution depends on the choice of the objective

function. The objective function in terms of quantum network utility is a nonlinear function in general, and to make it a linear programming we had to either drop terms or treat some variables as indices which introduces extra overhead (i.e., increases the number of decision variables). Thus, along the lines of efficiently solving the network planning problem while keeping all terms in the objective function, exploring nonlinear solvers, or reformulating the problem as a semidefinite programming could be worth pursuing. We should however note that either integer linear-programming or nonlinear-programming are NP-hard and our framework is only applicable to quantum networks up to a certain size.

There are also new directions to explore in network modeling and protocols. We used an asynchronized sequential scheme for the entanglement distribution protocol. A possible direction would be to formulate the network planning for other protocols (synchronous or asynchronous) and compare the optimal solutions across different protocols in terms of the overall network throughput and required resources. On another note, we used a simplified model for the quantum memory decoherence in terms of a hard cutoff. It would be interesting to incorporate other decoherence models possibly with a continuous behavior.

## VII. ACKNOWLEDGEMENTS

The authors acknowledge insightful discussions with Bing Qi, Stephen DiAdamo, Matthew Turlington and Lee Sattler.

## REFERENCES

- [1] Koji Azuma, Sophia E Economou, David Elkouss, Paul Hilaire, Liang Jiang, Hoi-Kwong Lo, and Ilan Tzitrin. Quantum repeaters: From quantum networks to the quantum internet. *arXiv preprint arXiv:2212.10820*, 2022.
- [2] Koji Azuma, Kiyoshi Tamaki, and Hoi-Kwong Lo. All-photon quantum repeaters. *Nature communications*, 6(1):6787, 2015.
- [3] CH BENNET. Quantum cryptography: Public key distribution and coin tossing. In *Proc. of IEEE Int. Conf. on Comp. Sys. and Signal Proc.*, Dec. 1984, 1984.
- [4] Charles H Bennett and Gilles Brassard. Quantum cryptography: Public key distribution and coin tossing. *arXiv:2003.06557*, 2020.
- [5] H-J Briegel, Wolfgang Dür, Juan I Cirac, and Peter Zoller. Quantum repeaters: the role of imperfect local operations in quantum communication. *Physical Review Letters*, 81(26):5932, 1998.
- [6] Kaushik Chakraborty, David Elkouss, Bruno Rijsman, and Stephanie Wehner. Entanglement distribution in a quantum network: A multicommodity flow-based approach. *IEEE Transactions on Quantum Engineering*, 1:1–21, 2020.
- [7] J Ignacio Cirac, AK Ekert, Susana F Huelga, and Chiara Macchiavello. Distributed quantum computation over noisy channels. *Physical Review A*, 59(6):4249, 1999.
- [8] Francisco Ferreira da Silva, Guus Avis, Joshua A Slater, and Stephanie Wehner. Requirements for upgrading trusted nodes to a repeater chain over 900 km of optical fiber. *arXiv preprint arXiv:2303.03234*, 2023.
- [9] Axel Dahlberg, Matthew Skrzypczyk, Tim Coopmans, Leon Wubben, Filip Rozpedek, Matteo Pompili, Arian Stolk, Przemysław Pawełczak, Robert Knegjens, Julio de Oliveira Filho, et al. A link layer protocol for quantum networks. In *Proceedings of the ACM SIGCOMM*, pages 159–173, 2019.
- [10] G Mauro D’Ariano, Paolopiacido Lo Presti, and Matteo GA Paris. Using entanglement improves the precision of quantum measurements. *Physical review letters*, 87(27):270404, 2001.
- [11] ESnet. Energy Sciences Network <https://my.es.net/>. [Online; accessed 2-January-2023].
- [12] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum-enhanced measurements: beating the standard quantum limit. *Science*, 306(5700):1330–1336, 2004.
- [13] Frank Kelly. Charging and rate control for elastic traffic. *European transactions on Telecommunications*, 8(1):33–37, 1997.
- [14] Frank P Kelly, Aman K Maulloo, and David Kim Hong Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49:237–252, 1998.
- [15] H Jeff Kimble. The quantum internet. *Nature*, 453(7198):1023–1030, 2008.
- [16] Peter Komar, Eric M Kessler, Michael Bishof, Liang Jiang, Anders S Sørensen, Jun Ye, and Mikhail D Lukin. A quantum network of clocks. *Nature Physics*, 10(8):582–587, 2014.
- [17] Yuan Lee, Wenhan Dai, Don Towsley, and Dirk Englund. Quantum network utility: A framework for benchmarking quantum networks. *arXiv preprint arXiv:2210.10752*, 2022.
- [18] Seth Lloyd, Jeffrey H Shapiro, Franco NC Wong, Prem Kumar, Selim M Shahriar, and Horace P Yuen. Infrastructure for the quantum internet. *ACM SIGCOMM Computer Communication Review*, 34(5):9–20, 2004.
- [19] William J. Munro, Koji Azuma, Kiyoshi Tamaki, and Kae Nemoto. Inside quantum repeaters. *IEEE Journal of Selected Topics in Quantum Electronics*, 21(3):78–90, 2015.
- [20] Momtchil Peev, Christoph Pacher, Romain Alléaume, Claudio Barreiro, Jan Bouda, W Boxleitner, Thierry Debuisschert, Eleni Diamanti, Mehrdad Dianati, JF Dynes, et al. The secoqc quantum key distribution network in vienna. *New Journal of Physics*, 11(7):075001, 2009.
- [21] Martin B Plenio. Logarithmic negativity: a full entanglement monotone that is not convex. *Physical review letters*, 95(9):090503, 2005.
- [22] Julian Rabbie, Kaushik Chakraborty, Guus Avis, and Stephanie Wehner. Designing quantum networks using preexisting infrastructure. *npj Quantum Information*, 8(1):5, 2022.
- [23] Eric M Rains. A semidefinite program for distillable entanglement. *IEEE Transactions on Information Theory*, 47(7):2921–2933, 2001.
- [24] Shouqian Shi and Chen Qian. Concurrent entanglement routing for quantum networks: Model and designs. In *Proceedings of the Annual conference of the ACM SIGCOMM*, pages 62–75, 2020.
- [25] Neil Sinclair, Erhan Saglamyurek, Hassan Mallahzadeh, Joshua A Slater, Mathew George, Raimund Ricken, Morgan P Hedges, Daniel Oblak, Christoph Simon, Wolfgang Sohler, et al. Spectral multiplexing for scalable quantum photonics using an atomic frequency comb quantum memory and feed-forward control. *Physical review letters*, 113(5):053603, 2014.
- [26] Damien Stucki, Matthieu Legre, Francois Buntschu, B Clausen, Nadine Felber, Nicolas Gisin, Luca Henzen, Pascal Junod, Gérald Litzistorf, Patrick Monbaron, et al. Long-term performance of the swissquantum quantum key distribution network in a field environment. *New Journal of Physics*, 13(12):123001, 2011.
- [27] Gayane Vardoyan and Stephanie Wehner. Quantum network utility maximization. *arXiv preprint arXiv:2210.08135*, 2022.
- [28] Guifré Vidal and Reinhard F Werner. Computable measure of entanglement. *Physical Review A*, 65(3):032314, 2002.
- [29] Hanfeng Wang, Matthew E Trusheim, Laura Kim, Hamza Raniwala, and Dirk R Englund. Field programmable spin arrays for scalable quantum repeaters. *Nature Communications*, 14(1):704, 2023.
- [30] Shuang Wang, Wei Chen, Zhen-Qiang Yin, Hong-Wei Li, De-Yong He, Yu-Hu Li, Zheng Zhou, Xiao-Tian Song, Fang-Yi Li, Dong Wang, et al. Field and long-term demonstration of a wide area quantum key distribution network. *Optics express*, 22(18):21739–21756, 2014.
- [31] Jin Y Yen. Finding the k shortest loopless paths in a network. *management Science*, 17(11):712–716, 1971.

## VIII. APPENDIX

### A. DERIVATION OF AVERAGE END-TO-END ENTANGLEMENT GENERATION RATE

In this appendix, we derive the end-to-end ebit rate for a path with spatial multiplexing and explain our approximate formula.

Consider a path with  $h$  links and width  $W$  where the success probability for link-level entanglement generation is given by  $p_k$  with  $1 \leq k \leq h$ . Let  $Q_k^i$  be the probability of the  $k$ -th link on the path having  $w$  successful ebits given by the



binomial distribution  $B(W, p)$  as in

$$\text{Prob}(i_k = w) = \binom{W}{w} p^w (1 - p_k)^{W-w}, \quad (32)$$

where  $0 \leq w \leq W$ . Let  $P_k^i$  be the probability of each of the first  $k$  links of the path having at least  $i$  successful ebits, which obeys a recurrence relation as follows

$$P_k^i = P_{k-1}^i \cdot \text{Prob}(i_k \geq i) + \text{Prob}(i_k = i) \cdot \sum_{l=i+1}^W P_{k-1}^l, \quad (33)$$

where  $\text{Prob}(i_k \geq w) = 1 - \Phi_k(w)$  and  $\Phi_k(w)$  is the CDF of the probability distribution of the  $k$ -th link. The initial condition is set by the first link that is  $P_1^i = \text{Prob}(i_1 = w)$ . The average throughput can be computed by

$$R_{e2e} = q_{sw}^{h-1} \sum_{w=1}^W i \cdot P_h^w. \quad (34)$$

The above expression can be computed iteratively. Alternatively, the average throughput can be written as

$$R_{e2e} = q_{sw}^{h-1} \sum_{w=1}^W w \sum_{k=1}^h \text{Prob}(i_k = w) \prod_{j=1, j \neq k}^h \text{Prob}(i_j \geq w). \quad (35)$$

The binomial distribution (32) in the limit  $Wp_k \gg 1$  can be well approximated by the normal distribution  $\mathcal{N}(Wp_k, Wp_k(1 - p_k))$  which sharply peaks at  $Wp_k$ . The average throughput can then be approximated by the bottleneck link (call it  $\ell$ -th link) with smallest peak at  $Wp_{\min}$ . As a result, the dominant term in the above sum corresponds to  $k = \ell$  such that  $\text{Prob}(i_j \geq w) \approx 1$  and  $\sum_{w=1}^W w \text{Prob}(i_\ell = w) = Wp_\ell$ . Hence, we arrive at Eq. (2).

## B. ANALYSIS OF PATH-BASED FORMULA

In this appendix, we show our path-based formulation with a reasonable number of shortest paths is able to find the optimal solution that the full optimization problem (the link-based formulation) can find for different random topologies with different numbers of nodes. We choose  $|Q| = 6$  user pairs randomly in each topology.

Figure 7(a) shows that the aggregate utility of the user pairs reaches the optimal value above a certain number of enumerated paths. This is expected because if we enumerate paths from shortest to longest, the paths after a certain point will be so enough that the end-to-end rate using them drops significantly. In addition, longer paths would most likely have a larger number of links and that affects the end-to-end fidelity and the expected throughput (due to swaps). For these reasons, we set the number of enumerated paths as an input to our optimization problem in all our experiments to 4000. Note that we also consider different versions of a path each with a different width.

Figure 7(b) shows the processing time in seconds for solving the path-based formulation as we increase the number of

## Utility and Processing time as # of enumerated paths

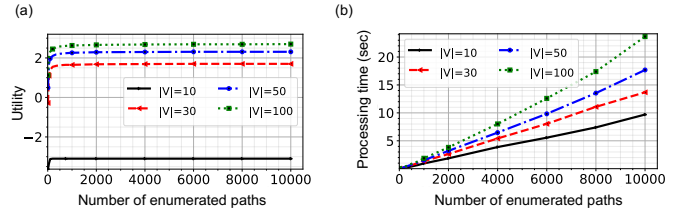


FIGURE 7: Utility and the processing time as a function of the number of enumerated paths.  $|V|$  is the number of nodes in random topologies.

paths in the input of the optimization problem. For topologies in the size of SURFnet (with 50 nodes), we have processing time of 25 seconds when we enumerate 10k paths. We have shown in 7(a) that enumerating only 2000 paths is enough for topologies with 50 nodes.

## C. ESNET PATHS

Figure 8 shows the optimal paths for the user pairs on the ESnet along with the number of memories for each path in terms of repeater capacity  $D$ .



SHAHROOZ POURYOUSEF received the B.S. degree in Information Technology from Shahid-Madani University of Azarbayjan, Iran, in 2013 and the M.S. degree in computer engineering from Sharif University of Technology, Iran, in 2015 and another M.S. degree in the College of Information and Computer Science at the University of Massachusetts Amherst USA. He is currently pursuing a Ph.D. degree in the College of Information and Computer Science at the University of Massachusetts Amherst USA. His research interest includes quantum networks and distributed quantum computing.



HASSAN SHAPOURIAN holds an MS in Electrical Engineering from Princeton University and a PhD in Theoretical Physics from the University of Chicago. He is currently a Senior Quantum Researcher at Cisco, where he leads a wide range of projects on photonic quantum information processing and hardware physics. He has received several awards including Microsoft Research Postdoctoral Fellowship, Simons Postdoctoral Fellowship (Collaboration on Ultra-Quantum Matter), and Kavli Institute for Theoretical Physics (KITP) Graduate Fellowship.



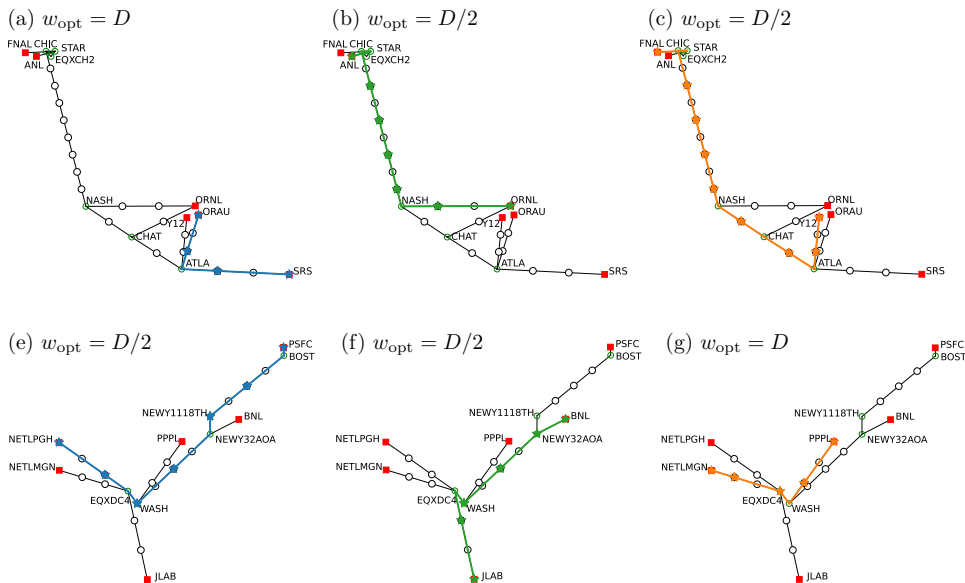
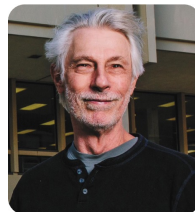


FIGURE 8: Optimal paths for various user pairs on the augmented ESnet with  $L_{\max} = 70\text{km}$ .  $w_{\text{opt}}$  denotes the number memories obtained for each user pair as a fraction of the repeater capacity  $D$ .



ALIREZA SHABANI is a scientist and entrepreneur who established a quantum lab for Cisco Systems. Prior to that, he founded Qulab, a pharmaceutical startup leveraging AI to automate drug design. He was also a senior scientist at Google Quantum AI Lab. His research has been at the intersection of quantum physics, engineering, and biology. He holds a Ph.D. in electrical engineering from the University of Southern California and was a postdoctoral scholar at Princeton

University and UC-Berkeley.



DON TOWSLEY (Fellow, IEEE) received the PhD degree in computer science from the University of Texas. He is currently a distinguished professor with the Manning College of Information & Computer Sciences. His research interests include performance modeling and analysis, and quantum networking. He has received several achievement awards including the 2007 IEEE Koji Kobayashi Award and the 2011 INFOCOM Achievement Award.

...



RAMANA KOMPPELLA is currently the Head of Cisco Research. His background is a perfect confluence of research and entrepreneurship, having spent a significant part of his career in academia (MS from Stanford, PhD from UCSD, tenured faculty at Purdue) as well as in startups (Co-Founder/CTO at AppFormix, Co-Founder/Head of Engineering/CTO at Candid alpha project inside Cisco). He was also part of the Google's network architecture team where he focused on large-scale

data center network operations.