E2E Fidelity Aware Routing and Purification for Throughput Maximization in Quantum Networks

Yangming Zhao*, Gongming Zhao* and Chunming Qiao†
*School of Computer Science and Technology, University of Science and Technology of China
†Department of Computer Science and Engineering, University at Buffalo

***Abstract*—This paper studies reliable teleportation of quantum bits (called qubits) in a quantum data network with multiple sources (S) and destinations (D) as well as repeaters. To teleport qubits for a SD pair reliably, not only an entanglement path for the SD pair, but also appropriate purification of the links along the path is required to ensure that the end-to-end (E2E) fidelity of the established entanglement connections is high enough.**

This is the first work on quantifying the E2E fidelity, and also using this E2E fidelity to determine critical links to achieve the most resource efficient purification. A novel approach called E2E Fidelity aware Routing and Purification (EFiRAP) is proposed to maximize network throughput, *i.e.*, the number of entanglement connections among multiple SD pairs, with each connection having an E2E fidelity above a given required threshold. EFiRAP accomplishes this goal by first preparing multiple candidate entanglement paths and determining optimal purification schemes, and then selecting the final set of entanglement paths that can maximize network throughput under the given quantum resource constraints. Existing works only ensured the fidelity of individual links, rather than the E2E fidelity is above a given threshold. Extensive simulations show that the proposed EFiRAP can enhance network throughput by about 50% when compared with the state-of-the-art approach.

I. Introduction

Quantum computing holds the potential to solve certain types of problems more efficiently than classic computers [1]. For example, it can solve the integer factorization problem, which is NP-hard with classic computers, in polynomial time [2]. However, in a foreseeable future, it is expected that each quantum computer can deal with only a few quantum bits (called data *qubits*). To overcome such a limitation, we can network many small quantum computers using a quantum network to form a distributed processing system [3, 4], akin to a cloud computing system with classic computers [5]–[7].

A quantum network consists of *quantum nodes*, each serving as a source (Alice), destination (Bob) or quantum repeater. These nodes are interconnected with *quantum links*, which can be fibers or free space optical links. Each quantum node has some *quantum memory* to store qubits, and each quantum link carries *quantum channels* (*e.g.*, wavelengths) that can be used to deliver qubits from one end to the other. However, since a *data qubit* from Alice is likely to be lost if it were to be transmitted over one or more quantum channels, and moreover, the qubit cannot be simply copied by Alice for retransmission once it's lost due to the no-cloning theory [8], the prevailing approach used in a quantum network is to entangle Alice and Bob with the help of *non-data* qubits, and then use an approach unique to quantum communication known as *teleportation* to transfer the quantum state information carried by the data qubit from Alice to Bob.

In its simplest form, to entangle Alice and Bob when they are physically interconnected via one or more quantum links and repeaters, one first identifies a path from Alice to Bob. Then, over each link between every two physically adjacent quantum nodes along this path, a Bell pair of non-data qubits (which are ideally entangled) are generated and distributed to the two end nodes to create an *entanglement link*. As a result, Alice holds one qubit of a Bell pair and Bob holds a qubit of another Bell pair, while each repeater along the path holds two qubits, belonging to two different Bell pairs. An *E2E entanglement connection* over the path from Alice to Bob can then be established by "stitching" multiple entanglement links together in a pair-wise fashion. This can be accomplished by having each repeater perform the so-called *internal swapping* [9, 10], *i.e.* measure its two qubits, and send the measurement results to Bob (through classical channels). Bob then performs a unitary operation on its own qubit based on the measurement results received. Afterwards, the two qubits stored at Alice and Bob will be entangled, thus establishing an entanglement connection between Alice and Bob (see more details in Section II).

Note that due to the so-called entanglement decoherence, signal decay, or environment interference *etc.*, a pair of qubits may not be perfectly entangled, and even if they were at the time of creation, they may not remain in the desired entanglement state, rendering the corresponding entanglement link or connection non-usable for a reliable teleportation operation. In this paper, we use *fidelity* [9] (a value between 0 and 1) to quantify the probability that a pair of entangled qubits (*i.e.*, the corresponding entanglement link/connection) are in the desired state.

The fidelity of an entanglement link can be improved via purification [10, 11] by using so-called *sacrificial Bell Pairs* (which consumes precious quantum resources). Previous works focused on methods to improve the fidelity of each and every entanglement link over a given threshold if possible, with the expectation that doing so will increase the E2E fidelity of each established entanglement connection accordingly [12, 13]. However, none of the existing works quantified how the purification can help improve the E2E fidelity, and thus none can offer any guarantee on the E2E fidelity.

Quantifying the E2E fidelity in a quantum network is challenging since it is affected by many factors (and can't

be simply measured). This work represents the first attempt to formulate the E2E fidelity under bit flip errors [14, 15]. Based on our analysis, we find that i) as expected, purifying different entanglement links have different effects on the E2E fidelity; ii) somewhat counter-intuitively, purifying an entanglement link may even *degrade* the E2E fidelity of a certain entanglement connection. Accordingly, in order to maximize the network throughput in a resource limited quantum network, it is important to not only identify critical links to purify so as to establish entanglement connections with a desired E2E fidelity in the most resource-efficient way, but also establish as many entanglement connections as possible via a joint optimization of entanglement path routing and purification.

In this paper, we propose E2E Fidelity aware Routing And Purification (EFiRAP) to identify the appropriate entanglement paths to use for each SD pair, and the corresponding purification scheme that determines which links to purify, and how many sacrificial Bell pairs will be used to purify each entanglement link, in order to establish as many entanglement connections with a desired E2E fidelity as possible for multiple SD pairs, given limited quantum resources in the network. Though there are many previous works on the entanglement routing to maximize the network throughput [9, 12, 13, 16, 17], they either ignore the fidelity issue [9, 16, 17] or only focus on the fidelity of created entanglement links [12, 13], rather than the E2E entanglement connections. To the best of our knowledge, we are the first to maximize the network throughput with E2E fidelity guarantee.

We summarize the major technical contributions of this paper as follows:

- The first-of-its-kind method to calculate the E2E fidelity of an entanglement connection established by connecting multiple entanglement links with a given fidelity;
- Solid analysis to determine the most critical link to purify in order to resource-effectively improve the E2E fidelity of an entanglement connection;
- A novel algorithm to determine purification schemes for a set of candidate entanglement paths so that each corresponding established entanglement connection has a E2E fidelity larger than a predefined threshold;
- An efficient algorithm to select among the set of candidate entanglement paths in order to maximize the network throughput.

II. Background

In this section, we first present some preliminary background information, including quantum states, our assumptions, purification technique, and how to establish an entanglement connection by connecting multiple entanglement links. After that, we briefly review some recent works on the entanglement routing followed by a toy example to motivate the proposed EFiRAP approach. At the end of this section, we present a high-level overview of EFiRAP. To simplify the presentation, we will refer to the non-data qubits used to establish entanglement links simply as qubits hereafter.

A. Quantum States and Bit Flip Errors

Just as a classic bit has a state of either 0 or 1, the two basic states for a qubit are $|0\rangle$ and $|1\rangle$. Different from a classic bit, a qubit can be in a superposition state $|\zeta\rangle = a_0|0\rangle + a_1|1\rangle$. When we measure a qubit in the state $|\zeta\rangle$, we will get either 0, with probability $|a_0|^2$, or 1, with probability $|a_1|^2$, where $|a_0|^2 + |a_1|^2 = 1$. However, due to the fact that a qubit may experience a bit flip error [14, 15] after going through a quantum logic gate operation such as *Control NOT* (CNOT) as a part of the measurement process, we may get 0 with probability $|a_1|^2$, and 1 with probability $|a_0|^2$ instead.

A two-qubit system can be in a superposition of states $|00\rangle$, $|01\rangle$, $|10\rangle$ and $|11\rangle$, which can be described as $|\xi\rangle = a_{00}|00\rangle + a_{01}|01\rangle + a_{10}|10\rangle + a_{11}|11\rangle$ ($\sum_{x=0}^{1}\sum_{y=0}^{1}|a_{xy}|^2 = 1$). When we measure such a two-qubit system, we will get xy (*i.e.*, one qubit is x and the other is y) with probability $|\alpha_{xy}|^2$. A Bell pair is a special two-qubit system in any of the following four *Bell states*: $|\beta_{xy}\rangle = \frac{|0y\rangle + (-1)^x|1\bar{y}\rangle}{\sqrt{2}}$, where $\bar{y} = 1 - y$. A representative Bell pair is $|\beta_{00}\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}$. If there is a bit flip error associated with the first qubit of a Bell pair $|\beta_{00}\rangle$ during measurement, the result will neither be 00 nor 11. Instead, it will be either 10 or 01 with an equal probability (*i.e.*, 0.5).

B. Assumptions

Assumption 1. *In a quantum network, the state of any Bell pair is* $|\beta_{00}\rangle$, i.e.*, we do not consider three other possible Bell pairs.*

Since all Bell states can be easily converted to $|\beta_{00}\rangle$, this assumption does not impact the generality of our work.

Assumption 2. *The fidelity of all the Bell pairs generated over the same link (without any purification) is identical.*

This assumption is reasonable for the Bell pairs generated by the same devices in the same environment over the same quantum link. Of course, Bell pairs generated over different links will likely have different fidelity.

Assumption 3. *Each qubit of a Bell pair will flip with the same probability (although independently) during the measurement process.*

This assumption is reasonable since the two qubits have the same fidelity to start with but each will go through identical but imperfect measurement process.

For a Bell pair of qubits in state $|\beta_{00}\rangle$, it will keep its entangled Bell state if and only if i) both qubits stay in the original state or ii) both qubits suffer a bit flip. Accordingly, under Assumption 3, if the fidelity of Bell pair i (or entanglement link i) is F_i, then the probability that each of its qubits will *not* flip, denoted by p_i, will satisfy the equation $p_i^2 + (1-p_i)^2 = F_i$ [10]. In other words, we have

$$p_i = \frac{1}{2} + \frac{1}{2}\sqrt{2F_i - 1} \tag{1}$$

The above equation reveals an intuitive relationship between bit-flip error probability and fidelity: a lower fidelity may result in a small p or a higher bit-flop error probability. Accordingly, to reduce bit-flip errors, we need to improve the fidelity through purification.

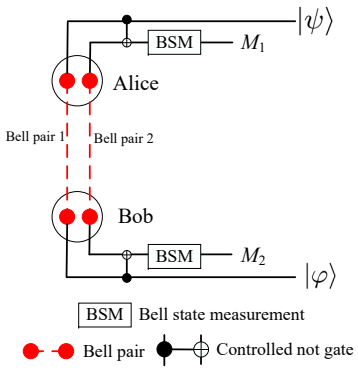


Fig. 1. Circuit for basic purification.

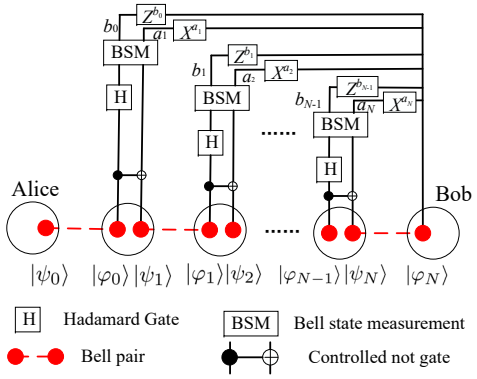


Fig. 2. Circuit to establish an entanglement connection.

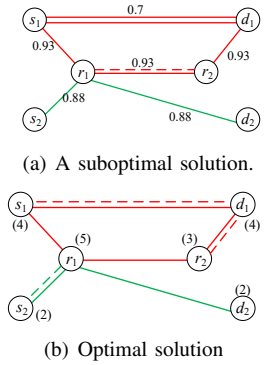


Fig. 3. A motivating example (Solid lines for entanglement links, while dotted lines for purification.).

C. Purification and Entanglement Connection

Purification: The main idea of purification of a Bell pair (or its corresponding entanglement link) is to use another Bell pair to test if the (first) entanglement link is in the expected state or not. If not, we will get rid of the entanglement link and generate two more Bell pairs, and repeat the purification process. Otherwise, the tested entanglement link is kept, with an improved fidelity, since we have ruled out a case in which this link is known to be "bad" based on the previous test.

The basic purification operation is shown in Fig. 1. We begin with two Bell pairs. Alice and Bob each hold one qubit of each Bell pair. Bell pair 1 will be used as an entanglement link, and Bell pair 2 will be used to test if *Bell pair 1 is in the state* $|\beta_{00}\rangle$. To this end, Alice and Bob each send their two qubits into a controlled-not (CNOT) gate, using the qubit of Bell pair 1 as the control bit while the qubit of Bell pair 2 as the target bit. Since CNOT is equivalent to XOR [18], after going through the CNOT gate, the target bit will hold the parity of the control and target bits. Then, Alice and Bob measure their qubits from Bell pair 2 (*i.e.*, the target bit). If their measurement results (*i.e.*, M_1 and M_2) agree, it means that *most likely* (but not surely) $|\psi\varphi\rangle = |\beta_{00}\rangle$ (*i.e.*, in the desired state), so Bell pair 1 (*i.e.*, corresponding entanglement link) will be kept. Otherwise, Bell pair 1 will be discarded. Note that Bell pair 2 will always be discarded since the two target qubits will no longer be entangled after being measured. This Bell pair 2 is referred to as a *sacrificial pair*.

There are two cases in which the two measurements (*i.e.*, M_1 and M_2) will agree: i) both Bell pairs are in the state $|\beta_{00}\rangle$ (*i.e.*, true positive); ii) both experienced a bit flip, or in other words, neither is in the state $|\beta_{00}\rangle$ (*i.e.*, false positive). Accordingly, given that the two Bell pairs have an initial fidelity of F, the probability of true positive is F^2, while the probability of getting a positive measurement result is $F^2 + (1-F)^2$. Accordingly, the fidelity of the entanglement link after purification will be $F' = \frac{F^2}{F^2+(1-F)^2} > F$. We can also leverage multiple sacrificial pairs to purify the same Bell pair for multiple times and further improve its fidelity.

If a Bell pair is purified by T sacrificial pairs, the fidelity $F^{(T)}$, can be iteratively calculated via

$$F^{(t)} = \frac{FF^{(t-1)}}{FF^{(t-1)} + (1-F)(1-F^{(t-1)})} \tag{2}$$

where $F^{(0)} = F$. Note that the sacrificial pairs may be generated on demand as needed, *i.e.*, only after the previous purification is performed, the purified Bell pair is kept but the entangled link needs further purification. Such on-demand generation of sacrificial pairs can minimize the quantum resources needed for generating entangled Bell pairs. However, the time it will take to generate the next sacrificial pair may be long, during which period, the entanglement link to be purified will become decoherent. Accordingly, in EFiRAP, all T sacrificial pairs are generated simultaneously at the cost of consuming more quantum resources.

Entanglement connection establishment: To establish an entanglement connection between Alice and Bob, we have to identify an entanglement path, and create an entanglement link between any two adjacent nodes along it. At last, all these entanglement links will be connected together via internal swapping to establish the entanglement connection.

As illustrated in Fig. 2, there are $N+1$ entanglement links created along a path and the state of the two qubits hosted by the n^{th} repeater (where $1 \leq n \leq N$) are $|\varphi_{n-1}\rangle$ and $|\psi_n\rangle$, respectively. Ideally, over any link, $|\psi_n\varphi_n\rangle = |\beta_{00}\rangle$ (where $0 \leq n \leq N$). To establish an entanglement connection via internal swapping, each repeater measures the two qubits it hosts, and sends the results, denoted by b_{n-1} and a_n, respectively, to Bob (through a classical channel). Then, Bob will perform a set of matrix operations, denoted by $M = Z^{b_0}X^{a_1}Z^{b_1}X^{a_2}\cdots Z^{b_{N-1}}X^{a_N}$ on his qubit in the state $|\varphi_N\rangle$, where X and Z are pauli-X matrix and pauli-Z matrix, respectively. If everything goes as expected, the two qubits hosted by Alice and Bob will be entangled in the state $|\beta_{00}\rangle$, thus establishing an E2E entanglement connection.

D. Previous Work

The problem studied in this paper is usually called entanglement routing problem [9, 16, 19]. In this area, most of the existing works only focused on a specific topology [20]–[23]. [17] and [24] focused on the entanglement routing problem on a general topology, but assumed that all the entanglement links will be successfully created and stay in the expected state. Neither studied the problem of having low-fidelity entanglement links and connections.

[9] and [16] took into consideration the entanglement link quality. However, they only considered the case where some entanglement links may fail to be created. While one can try multiple times until an entanglement link is successfully created, one cannot ensure that any entanglement link created has (and will keep) a high enough fidelity.

[19] is perhaps the first work that takes entanglement link fidelity into consideration. The main idea in [19] is to tackle the time-induced decoherence issue by reducing the duration that every entanglement link needs to be maintained before an internal swapping operation is performed, so that they are likely to stay in the expected state to enable entanglement connections to be established with high-fidelity entanglement links. That work didn't consider purification for an entanglement link may have a low fidelity when it is first created. [13] discussed several purification schemes but when it comes to entanglement routing, the main idea is to assign each link a cost which is inversely proportional to the total number of sacrificial pairs supported by the link, and then use the Dijkstra's algorithm to find a shortest path. The work didn't quantify, let alone guarantee, the E2E fidelity.

[12] is the closest related work, in which an elaborate entanglement routing solution based on the idea of first purifying the links was proposed, so that only the links whose fidelity can be purified above a given threshold will be used in the routing. However, it still didn't quantify the E2E fidelity. In addition, as to be shown, purifying all links so their fidelity is above a threshold will not only waste resources, but also fail to guarantee a high enough E2E fidelity.

As a result, the goodput (measured in terms of the number of entanglement connections with a high enough fidelity) achieved by all of the prior work could be low. In contrast, this work is the first that can maximize the number of entanglement connections, whose fidelity can be guaranteed to be above a given threshold, resulting in the highest throughput.

E. A Motivating Example

We use the example shown in Fig. 3 to motivate the design of EFiRAP. In this example, we are to establish entanglement connections with fidelity of at least 0.81 between two SD pairs (s_1, d_1) and (s_2, d_2). The fidelity of the Bell pairs generated over each quantum link, ranging from 0.7 to 0.93, is shown in Fig. 3(a), and the number of quantum memory units hosted by each node is shown in a bracket beside each node in Fig. 3(b). In addition, the capacity of every quantum link is 2 (*i.e.*, each link can have two Bell pairs at the same time). For the SD pair (s_1, d_1), we can find two entanglement paths $s_1 \rightarrow d_1$ and $s_1 \rightarrow r_1 \rightarrow r_2 \rightarrow d_1$, while for SD pair (s_2, d_2), there is only one entanglement path $s_2 \rightarrow r_1 \rightarrow d_2$. Without purification, the fidelity of the entanglement connections established over any of these entanglement paths cannot reach 0.81 (The way to calculate fidelity will be discussed in Section III-A).

Let us consider the entanglement connections between s_1 and d_1. There are two paths between this SD pair: $s_1 \rightarrow d_1$ directly, and $s_1 \rightarrow r_1 \rightarrow r_2 \rightarrow d_1$. Over the direct $s_1 \rightarrow d_1$ link, we can establish two entanglement connections, as shown in Fig. 3(a). However, the fidelity of each entanglement connection is only 0.7. If we use one Bell pair as a sacrificial pair as shown in Fig. 3(b), we can establish one entanglement connection over link (s_1, d_1) with fidelity 0.84. This example shows that *purification is necessary (and quite useful)*.

Now, consider the entanglement connections established over the path $s_1 \rightarrow r_1 \rightarrow r_2 \rightarrow d_1$. To derive an entanglement connection with fidelity larger than 0.81, we can perform one round of purification over any entanglement link along the path. If we would perform purification over link (r_1, r_2) as shown in Fig. 3(a), there would not be remaining quantum memory on repeater r_1 to generate sacrificial pairs to improve the fidelity of the entanglement connection established over $s_2 \rightarrow r_1 \rightarrow d_2$. However, if we purify the entanglement link (r_2, d_1) along the path for $s_1 \rightarrow r_1 \rightarrow r_2 \rightarrow d_1$ as shown in Fig. 3(b), then we can generate another sacrificial pair to purify link (s_2, r_1) along the path $s_2 \rightarrow r_1 \rightarrow d_2$. In this way, we can establish a total of 3 entanglement connections, each with fidelity larger than 0.81, compared with only 1 entanglement connection with a required minimum fidelity as in Fig. 3(a). This example shows that *it is important to pick the right purification scheme* since which link to purify for which connection can significantly affect the network throughput, due to the limited quantum resources such as quantum memory available at the quantum nodes.

F. EFiRAP in a Nutshell

As in [9, 16], we assume a time-slotted quantum network with many SD pairs requesting entanglement connections. The objective of EFiRAP is to maximize the number of entanglement connections whose E2E fidelity is above a given threshold. To do so, EFiRAP must not only figure out which entanglement paths to use, but also which entanglement links will be purified, and how many sacrificial pairs need to be used for purification. In other words, EFiRAP aims to address the joint optimization of entanglement routing and purification scheme.

Inspired by the huge success and benefits of Software Defined Networking (SDN) [25, 26], EFiRAP will run in a centralized manner. There are two key bedrocks to EFiRAP. One is a first-of-its-kind formulation to quantify the E2E fidelity, and the other is an analytic approach to identify the most critical link to achieve resource efficient purification. Built upon these two bedrocks, EFiRAP solves the E2E fidelity aware entanglement routing problem in two steps. In the first step, EFiRAP prepares a Candidate Entanglement Paths Set (CEPS). Each element in CEPS contains not only the information about an entanglement path, but also the corresponding purification scheme. Two elements in CEPS may have the same entanglement path but different purification schemes. Every element in CEPS can be used to derive an entanglement connection satisfying the minimum E2E fidelity requirement. This step is accomplished by an algorithm called Entanglement Path Preparation (EPP).

After preparing CEPS, EFiRAP will select elements from CEPS to establish entanglement connections and maximize the network throughput, taking into consideration the need to

allocate the limited resource constraint such as link capacity (number of quantum channels) and the quantum memory available at each node among all entanglement connections competing for such resources. This step is accomplished by an algorithm called Entanglement Path Selection (EPS).

III. EFiRAP Details

In this section, we describe EFiRAP in detail. We first quantify the fidelity of an entanglement connection established by connecting multiple entanglement links in Section III-A. We then identify critical links to purify to achieve the highest-resource-efficiency in Section III-B. Based on these analysis, we propose the EPP algorithm to prepare CEPS in Section III-C, and the EPS algorithm to establish entanglement connections by selecting entanglement paths from CEPS in Section III-D.

A. E2E Fidelity Quantification

In this subsection, we analyze the fidelity of an E2E entanglement connection established by connecting $N+1$ ($N \geq 1$) entanglement links. In the analysis, we assume that the fidelity of entanglement links is given and there are only bit flip errors resulted from quantum gate operations during measurement.

Refer to Fig. 2 and associated discussion on entanglement connection establishment, we first note that since $XZ = -ZX$, the sequence of matrix operations M performed by Bob to its qubit $|\psi_N\rangle$ can be reformulated as $\prod_{i=1}^{N} X^{a_i} \prod_{j=0}^{N-1} Z^{b_j}$ by ignoring any negative sign (as it won't affect the entanglement between Alice and Bob [10]). Secondly, since $X^0 = Z^0 = I$ where I is the identity matrix, and in addition, $XX = ZZ = I$, M can be reduced to either (i). applying no X (or Z) to $|\psi_N\rangle$ if M starts with an even number of X's (or Z's) as a result of having an even number of non-zero a_i (or b_j); or (ii). applying one X (or Z) if M starts with an odd number of X's (or Z's) as a result of having an even number of non-zero a_i (or b_j).

Since a bit flip error experienced when measuring $|\varphi_i\rangle$ and $|\psi_{i+1}\rangle$ at each and every repeater (*i.e.*, excluding the quits at Alice and Bob) will flip the value of b_i and a_{i+1}, respectively, and $XZ \neq I$, we have the following observation on the sequence of matrix operations M.

Observation 1. M will not be affected, if and only if, (i). we have zero or an even number of bit flip errors when measuring the set $\{|\psi_i\rangle\}_{i=1}^{N}$; and (ii). we also have zero or an even number of flip errors when measuring the set $\{|\varphi_j\rangle\}_{j=0}^{N-1}$.

Finally, we note that a bit flip error experienced when operating on the qubit at Bob ($|\psi_N\rangle$) to entangle Bob with Alice at last, or when measuring the qubit at Alice ($|\varphi_0\rangle$) for teleportation later on, can be treated as if we have a bit flip error in the "imagined" a_{N+1} or a_0, respectively, in the following sense: if only one of them gets flipped (while zero or an even number of other a_i does), then the entanglement connection will be affected.

We now turn to the calculation of the E2E fidelity F. Recall that p_i from (1) is the probability that $|\psi_i\rangle$ or $|\varphi_i\rangle$ will not flip (during measurement). Let P_k be the probability that there are k bit flips among $\{|\varphi_i\rangle\}_{i=0}^{N-1}$. In the special case where the bit flip probability on every link i is the same, *i.e.* $p_i = p$, for all i, we have $P_1 = C_N^1(1-p)p^{N-1}$ and $P_3 = C_N^3(1-p)^3p^{N-3}$, and thus $\frac{P_1}{P_3} = \frac{6}{(N-1)(N-2)}\frac{p^2}{(1-p)^2}$. When $N = 4$ and $p = 0.8$, we have $\frac{P_1}{P_3} = 16$. Similarly, in the general case (*i.e.* where p_i are different), we typically also have $P_1 >> P_3 >> P_5 >> \cdots$. Accordingly, let U be the probability for having an odd number of bit flips among $\{|\varphi_i\rangle\}_{i=0}^{N-1}$, we can approximate $U = \sum_{i \text{ is odd}} P_i$ with

$$U \approx P_1 = \sum_{i=0}^{N-1} (1-p_i)\Phi_i \tag{3}$$

where $\Phi_i = \prod_{k=0,k\neq i}^{N-1} p_k$. The i^{th} term of (3) is the probability that $|\varphi_i\rangle$ flip and others do not.

Similarly, let Q_k be the probability that there are k bit flips among $\{|\psi_i\rangle\}_{i=0}^{N} \bigcup \{|\varphi_N\rangle\}$, we also have $Q_1 >> Q_3 >> Q_5 >> \cdots$. If V is the probability there are odd number of bit flips among $\{|\psi_i\rangle\}_{i=0}^{N} \bigcup \{|\varphi_N\rangle\}$, then we can approximate $V = \sum_{i \text{ is odd}} Q_i$ with

$$V \approx Q_1 = 2p_N(1-p_N)\Phi + p_N^2 \sum_{i=0}^{N-1} (1-p_i)\Phi_i \tag{4}$$

where $\Phi = \prod_{k=0}^{N-1} p_k$. In (4), the first term is the probability that $|\psi_N\rangle$ or $|\varphi_N\rangle$ flips while $\{|\psi_i\rangle\}_{i=0}^{N-1}$ do not, and the second term is the probability that neither $|\psi_N\rangle$ nor $|\varphi_N\rangle$ flips while one of $\{|\psi_i\rangle\}_{i=0}^{N-1}$ flips (note that the last summation is equal to U due to approximation).

Accordingly, the probability that the entanglement connection will not be in the expected state can be calculated as $P_{fail} = 1-(1-U)(1-V) = U+V-UV$, and its fidelity F (*i.e.*, the probability that an entanglement connection is in the expected state) will be

$$F = 1 - P_{fail} = 1 + UV - U - V \tag{5}$$

B. Critical Link for Purification

When an entanglement connection has a low E2E fidelity F, it is possible to improve F by purifying the fidelity of some of the entanglement links along the path. However, doing so would consume precious quantum resources such as quantum channels over quantum links, and quantum memory at quantum nodes. More importantly, as to be shown, sometimes, purifying a link may intentionally lead to a decreased E2E fidelity. Therefore, it is important to identify critical links in order to achieve the most resource efficient purification. To this end, we first determine the partial derivative of the E2E fidelity with respect to the fidelity of each entanglement link, that is, $\frac{\partial F}{\partial F_i}$. Intuitive, it is the most resource efficient to purify link i with the highest positive value of $\frac{\partial F}{\partial F_i}$.

More specifically, we have

$$\frac{\partial F}{\partial F_i} = -\frac{\partial P_{fail}}{\partial F_i} = -(1-U)\frac{\partial V}{\partial F_i} - (1-V)\frac{\partial U}{\partial F_i} \tag{6}$$

Based on the chain rule of derivative, we know

$$\frac{\partial U}{\partial F_i} = \frac{\partial U}{\partial p_i}\frac{dp_i}{dF_i}, \quad \frac{\partial V}{\partial F_i} = \frac{\partial V}{\partial p_i}\frac{dp_i}{dF_i} \tag{7}$$

From (1), (3), and (4), we can derive

$$\frac{dp_i}{dF_i} = \frac{1}{4p_i - 2} \tag{8}$$

$$\frac{\partial U}{\partial p_i} = \begin{cases} 0, & \text{if } i = N \\ \sum_{k=0}^{N-1} \Phi_{ik} - N\Phi_i, & \text{otherwise} \end{cases} \tag{9}$$

and

$$\frac{\partial V}{\partial p_i} = \begin{cases} 2(1-2p_N)\Phi + 2p_N U, & \text{if } i = N \\ 2p_N(1-p_N)\Phi_i + p_N^2 \frac{\partial U}{\partial p_i}, & \text{otherwise} \end{cases} \tag{10}$$

where $\Phi_{ik} = \begin{cases} 0 & \text{if } i = k \\ \Phi_i / p_k & \text{otherwise} \end{cases}$. According to (6)–(10), we can deduce the derivative of the E2E fidelity of an entanglement connection respect to the fidelity of each entanglement link.

Note that $\frac{\partial F}{\partial F_i}$ in (6) will be negative if $\frac{\partial V}{\partial F_i} > 0$ and $\frac{\partial U}{\partial F_i} > 0$ for entanglement link i (since $1 - U > 0$ and $1 - V > 0$). When this happens, *purifying link i will lead to a smaller E2E fidelity of the established entanglement connection*. To prevent the above pitfall, it is sufficient to ensure that both $\frac{\partial V}{\partial F_i}$ and $\frac{\partial U}{\partial F_i}$ are negative. Note that from (7), (8) and (9), we have $\frac{\partial U}{\partial F_i} = (\sum_{k=0,k\neq i}^{N-1} \frac{1}{p_k} - N)\frac{\Phi_i}{4p_i - 2}$. Since $\Phi_i > 0$ and $\frac{1}{4p_i - 2} > 0$, we know that $\frac{\partial U}{\partial F_i} < 0$ if and only if $\sum_{k=0,k\neq i}^{N-1} \frac{1}{p_k} < N$, or in other words, p_k are sufficiently large. For example, when $p_k = 1$ for all k, we have $\sum_{k=0,k\neq i}^{N-1} \frac{1}{p_k} = N - 1$. According to (1), p_k increases with F_k. This implies that as long as we can purify each link to keep F_k to be above a certain threshold F_{min}, we can ensure $\frac{\partial U}{\partial F_i} < 0$.

To calculate F_{min}, let $p_0 = p_1 = \cdots = p_N = p$, and solve

$$\begin{cases} \frac{\partial U}{\partial F_i} = [(N-1)p^{N-2} - Np^{N-1}]\frac{1}{4p-2} < 0, & \forall \quad i \neq N \\ \frac{\partial V}{\partial F_i} = [(3N-1)p^N - 3Np^{N+1}]\frac{1}{4p-2} < 0, & \forall \quad i \neq N \\ \frac{\partial V}{\partial F_N} = 2Np^N(2-3p)\frac{1}{4p-2} < 0 \end{cases}$$

We derive $p > \frac{3N-1}{3N}$. Accordingly, from (1), we have

$$F_{min} > \frac{(3N-1)^2 + 1}{(3N)^2}, \quad N \geq 1 \tag{11}$$

C. Candidate Entanglement Path Preparation

Based on above discussions, we design the Entanglement Path Preparation (EPP) algorithm to determine a set of candidate entanglement paths for each SD pair and their corresponding purification schemes.

Given the SD pair i, EPP first prepares K promising entanglement paths (Lines 1–2). To this end, we set the weight of quantum link l as $-\ln F_l$, where F_l is the fidelity of a Bell pair generated over quantum link l (Line 1), and then find out K different paths with Yen's algorithm [27] (Line 2). The rationale behind this setting is that if all the Bell pairs along an entanglement path R stay in the expected state, the established entanglement connections would be in the expected state. This probability is $\prod_{l\in R} F_l$. To maximize $\prod_{l\in R} F_l$, it can be reformulated as minimizing $\sum_{l\in R}(-\ln F_l)$. Along each entanglement path, EPP calculates corresponding purification schemes (Lines 3–15). To this end, EPP first performs purification, *i.e.*, adds sacrificial pairs, to ensure that the derivative of the E2E fidelity of the entanglement connection to be established with respect to every entanglement link along the entanglement path is positive (Lines 4–6), and then performs more purification until the fidelity of established entanglement connection achieves the minimum fidelity requirement F^* (Lines 7–15). In each iteration, EPP only adds one more sacrificial pair to purify the entanglement link with the largest derivative (Line 13). When there are multiple entanglement links that derive the same largest derivative, EPP will record all the possibilities in $\mathcal{A}$. This step can derive different purification schemes along the same entanglement path. By preparing candidate paths for every SD pair with Algorithm EPP, we can derive the CEPS.

Algorithm 1: Entanglement Path Preparation (EPP)

Input: Network topology, SD pair i, number of entanglement paths K, minimum fidelity requirement F^*, Bell pair fidelity over each link F_l

Output: A set of entanglement paths and corresponding purification scheme $\mathcal{C} = \{\{r_{ijl}\})\}$

1 Initialize $j \leftarrow 1$, set weight of each link l to be $-\ln F_l$;
2 Find K paths with Yen's Algorithm [27];
3 **for** *All K paths found in Line 2* **do**
4 Calculate F_{min} according to (11);
5 Calculate required number of sacrificial pairs on each link l (to ensure $F_l \geq F_{min}$);
6 Update purificatioin scheme $\{r_l\}$, $\mathcal{A} \leftarrow \mathcal{A} \cup \{r_l\}$;
7 **while** $\mathcal{A} \neq \Phi$ **do**
8 Pick out the first item $\{r_l\}$ from $\mathcal{A}$;
9 Calculate the E2E fidelity of corresponding entanglement connection F according to (5);
10 **if** $F \geq F^*$ **then**
11 $r_{ijl} \leftarrow r_l$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{r_{ijl}\}$, $j \leftarrow j + 1$;
12 **else**
13 Find out the set of entanglement links $\mathcal{L}$ that can derive largest derivative according to (6);
14 **for** *All $l' \in \mathcal{L}$* **do**
15 Add one more sacrificial pair to purify link l', update $\{r_l\}$ and add $\{r_l\}$ into $\mathcal{A}$ if $\{r_l\} \notin \mathcal{A}$;
16 **Return** $\mathcal{C}$.

D. Entanglement Path Selection

The problem to maximize the network throughput can be formulated as follows:

$$\text{maximize} \quad \sum_{i,j} x_{ij} \tag{12}$$

$$\text{subject to:} \quad \sum_{i,j} r_{ijl} x_{ij} \leq C_l \quad \forall l \tag{12a}$$

$$\sum_{l\in A(u)} \sum_{i,j} r_{ijl} x_{ij} \leq M_u \quad \forall u \tag{12b}$$

$$x_{ij} \text{ is integer}, \quad \forall i, j \tag{12c}$$

Algorithm 2: Entanglement Path Selection (EPS)

Input: CPES $\{r_{ijl}\}$, algorithm approximation ratio requirement ϵ

Output: Entanglement path selection to maximize efficient network throughput $\{x_{ij}\}$

1 Based on $\{r_{ijl}\}$, formulate Problem (12)
2 Relax constraint (12c) to be $x_{ij} \geq 0$ and solve derived LP model, say the solution is $\hat{x}_{ij}$ and corresponding objective value is $\hat{z}$
3 Let $s \leftarrow \min\{\lfloor \hat{z} \rfloor, \frac{m(1-\epsilon)}{\epsilon}\}$, where m is the number of constraints in (12a) and (12b)
4 Initialize $z^{ALG} \leftarrow 0$, $x_{ij}^{ALG} \leftarrow 0$
5 **for** *t from $\sum_{i,j} \lfloor \hat{x}_{ij} \rfloor$ to s* **do**
6 **for** *each solution $T = \{\bar{x}_{ij}\}$ satisfying $\sum_{i,j} \bar{x}_{ij} = t$* **do**
7 Substitute constraint (12c) to be $x_{ij} \geq \bar{x}_{ij}$ and solve the derived LP, say the solution is x_{ij}^*
8 **if** *the derived LP is feasible and $z^{ALG} < \sum_{i,j} \lfloor x_{ij}^* \rfloor$* **then**
9 $z^{ALG} \leftarrow \sum_{i,j} \lfloor x_{ij}^* \rfloor$, $x_{ij}^{ALG} \leftarrow x_{ij}^*$
10 **Return** $\{x_{ij}^{ALG}\}$ and z^{ALG}

where x_{ij} is an integer variable which indicates how many entanglement connections we should establish by adopting the j^{th} entanglement paths (and corresponding purification scheme) of SD pair i; C_l is the capacity of link l; and M_u is the number of quantum memory units hosted by node u, whose adjacent link set is $A(u)$. The objective is to maximize the network throughput. The first constraint says the number of Bell pairs that will be generated over each quantum link should not exceed the link capacity. Similarly, the second constraint is used to state the limitation enforced by the quantum memory hosted by each node. This is a generalized multidimensional Knapsack problem, which cannot be solved with a fully polynomial time approximation scheme (FPTAS) [28]. Though there is an existing algorithm designed for a generalized multidimensional Knapsack problem [29], it only solves the 0-1 Knapsack problem, while variables in (12) are non-negative integers. Accordingly, we design a PTAS algorithm named Entanglement Path Selection (EPS) to solve (12), which is shown in Algorithm 2.

EPS algorithm first "guesses" the optimal value of problem (12). For each guessed objective value, EPS tries all the possible entanglement path selections to achieve it if possible (Lines 6–9). After that, EPS also tries to fully utilize the remaining resources and further improve the solution by solving a relaxed LP problem (Line 7). In the end, EPS returns the entanglement paths selected scheme that can maximize the network throughput.

Theorem 1. *EPS algorithm is ϵ-approximation.*

Proof. Let the optimal objective value of problem (12) be z^{OPT}, while the corresponding solution is $\{x_{ij}^{OPT}\}$. The objective value should be larger than or equal to $\sum_{i,j} \lfloor \hat{x}_{ij} \rfloor$, as $\{\lfloor \hat{x}_{ij} \rfloor\}$ is a feasible solution to problem (12). If $z^{OPT} \leq s$, since EPS will try all the entanglement path selection schemes, it will achieve the optimal solution, *i.e.*, $z^{ALG} = x^{OPT}$.

When $z^{OPT} > s$, we have $s = \lceil \frac{m(1-\epsilon)}{\epsilon} \rceil$ and $\sum_{ij} \bar{x}_{ij} = s$. Given $T = \{\bar{x}_{ij}\}$, we construct $\{x_{ij}(T)\}$ as follows: i) remove the integer constraint in problem (12); ii) if $x_{ij}^{OPT} \geq \bar{x}_{ij}$ for some i, j, add constraint $x_{ij} \geq \bar{x}_{ij}$; iii) solve the derived LP model and get the solution $\{x_{ij}(T)\}$.

Let $D = \{x_{ij} : x_{ij}(T) > \lfloor x_{ij}(T) \rfloor\}$, then $|D| \leq m$, where m is the number of constraints in (12a) and (12b). We know $z^{OPT} \leq \sum_{i,j} x_{ij}(T) \leq \sum_{i,j} \lfloor x_{ij}(T) \rfloor + |D| \leq z^{ALG} + m$. Since EPS will try all the solutions satisfying $\sum_{i,j} x_{ij} = s$ and $z^{OPT} > s$, we have $z^{ALG} \geq s$. Accordingly, $z^{OPT} \leq z^{ALG} + m \leq z^{ALG} + m\frac{z^{ALG}}{s} \leq z^{ALG}(1 + \frac{\epsilon}{1-\epsilon})$. In other words, $z^{OPT} - z^{ALG} \leq \frac{\epsilon}{1-\epsilon} z^{ALG} \leq \frac{\epsilon}{1-\epsilon} z^{OPT}$. □

E. Discussions

Though the proposed PTAS algorithm can achieve a near-optimal performance, its time complexity could be high in large scale networks. To reduce the time complexity, we introduce several improvements. First, in Line 6 of Algorithm EPS, we check the feasibility of $T = \{\bar{x}_{ij}\}$ (*i.e.*, whether or not (12a) and (12b) hold when $x_{ij} = \bar{x}_{ij}$). If we find T is not a feasible solution, all the solutions $\hat{T} = \{\hat{x}_{ij}\}$ satisfying $\hat{x}_{ij} \geq \bar{x}_{ij}$ for all i, j will no longer be checked. This will significantly reduce the time complexity of Algorithm EPS.

Second, in Line 7, for a given t, whenever we find a feasible solution, we will go to $t+1$ without checking the feasibility of all the remaining solutions of $T = \{\bar{x}_{ij}\}$ satisfying $\sum_{i,j} \bar{x}_{ij} = t$. This procedure will repeat until we cannot find a feasible solution. Then, we will go back to $t-1$ to see if we can further improve the solution. In addition, EPS may try only a part of the possible solutions in Line 6.

Third, we can slice a network to reduce the problem complexity. For example, we can divide the entire network into multiple parts by minimizing the number of SD pairs that will be allocated to different parts. This can be accomplished with the minimum k-cut algorithm [30]. Then, we optimize the network throughput within every part, and finally leverage the remaining resources to provide services to the SD pairs cross multiple network parts.

IV. Performance Evaluation

In this section, we evaluate the performance of EFiRAP through extensive simulations using a custom in-house simulator. Simulations involve randomly generated networks with randomly chosen SD pairs, and control parameters for quantum memory and link capacity, and initial link fidelity. For a given set of parameters, simulations are run 100 trials and the averaged results are shown.

A. Simulation Methodology

Test case generation. To generate a network with N quantum nodes, we randomly connect these nodes with $2N$ quantum links. By default, there are 100 quantum nodes (correspondingly 200 quantum links) and 20 SD pairs in the network.

The fidelity of a Bell pair generated over a quantum link F_i is evenly distributed between [0.70, 0.95]; each node hosts 100 units of quantum memory; and the quantum link capacity is evenly distributed between [26, 35]. The minimally required E2E fidelity F^* is 0.8.

Comparative schemes. We compare EFiRAP with two types of routing and purification schemes mainly in terms of network throughput. One is a variation of REPS [16]. By treating the given fidelity of a Bell pair as if it were the probability of successfully creating the associate entanglement link in REPS, we leverage the Provisioning for Fault Tolerance (FPT) algorithm in REPS to determine the number of Bell pairs that should be generated over each link, and then use the Entanglement Path Selection (EPS) algorithm in REPS to determine the entanglement paths. At last, we use any remaining Bell pairs on each link to perform the purification. When such purification can contribute to multiple entanglement paths, we first assign it to the one whose current E2E fidelity is closest to meet the minimally required F^*. Only the entanglement connections whose E2E fidelity meets the minimally required F^* will be counted as a part of the throughput.

The other type of comparative scheme is a variation of the PS/PF/PU algorithms in [12]. We refer to the variation as sequential local purification and routing (SPAR) scheme. In SPAR, each link is purified with sacrificial Bell pairs to create as many entanglement links with fidelity exceeding a threshold as possible. Entanglement link whose fidelity is not high enough will be discarded in the next phase where entanglement routing takes place. As in the case for the variation of REPS, once the entanglement connections are generated, only those whose E2E fidelity meets the minimally required F^* will be counted as a part of the throughput. Since [12] didn't specify any value for the fidelity threshold of each entanglement link, in our simulations, we have tried three values, namely 0.9, 0.94, and 0.97. The corresponding curves are labeled as SPAR-0.9, SPAR-0.94, and SPAR-0.97, respectively.

Performance Metrics. In addition to network throughput, we also compare the percentage of quantum resource (*i.e.*, quantum memory and quantum channel) used by EFiRAP with the resource utilization in REPS and SPAR. A high quantum memory/channel utilization means that more resources are consumed for the establishment of all the entanglement connections. It also implies that fewer quantum resources will be available in the network for other applications or purposes.

B. Evaluation Results

Effect of network scale. In this set of simulations, we vary the number of quantum nodes in the network, while keeping all other parameters, in order to evaluate how different algorithms perform with the network scale. Simulation results are shown in Fig. 4. When the network scale is relatively small, the network throughput achieved by all schemes except SPAR-0.9 increases with the network scale since there are more quantum resources that can be used to establish entanglement connections. However, when the network scale continues to increase, the throughput of all algorithms reduces since the number of hops between a source and a destination increases, making it difficult to establish an entanglement connection achieving the minimally required E2E fidelity.

We note that when there are fewer than 100 nodes in the network, SPAR-0.94 outperforms SPAR-0.97 since the former can create more entanglement links that can be used to establish entanglement connections. However, when the network scale increases, the sources and destinations are far from each other, and many of the entanglement connections established by SPAR-0.94 cannot achieve the minimal E2E fidelity requirement. Accordingly, SPAR-0.97 performs slightly better in this case, although neither is close to achieving the same performance in throughput as EFiRAP.

Interestingly, the throughput achieved by SPAR-0.9 does not increase with the network scale even when the network scale is relatively small. This is because when the fidelity threshold of an entanglement link is set to only 0.9, an entanglement connection established by SPAR along a path longer than 2 hops will mostly likely fail to achieve the required E2E fidelity of $F^* = 0.8$. Accordingly, SPAR-0.9 performs the worst even in a small scale network and its performance will degrade with the network scale increase.

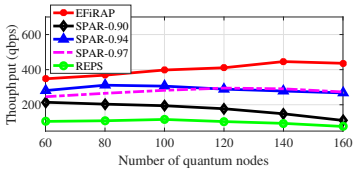
In our simulations, EFiRAP outperforms the best case of the three SPAR schemes and REPS by up to 50.94% and 471.86%, respectively. Since SPAR-0.94 and SPAR-0.97 achieve the similar performance and SPAR-0.94 performs better in our default setting. In the following, we only show the performance of SPAR-0.94, which performs the best among all three SPAR schemes in almost all other non-default settings as well.

From Figs. 4(b) & 4(c), we can observe that the quantum resource utilization will decrease with network scale. This is because that when the network scale increases, there are more network resources, much of which are not useful for establishing more entanglement connections.

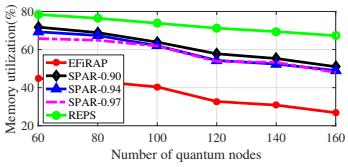
Effect of the amount of quantum resources. To study how the performance of EFiRAP will be impacted by the amount of quantum resources in the network, we first change the link capacity, *i.e.*, the number of quantum channels (also the number of Bell pairs that can be supported) over each quantum link, from 10 to 60, and then vary the quantum memory hosted by each quantum nodes from 70 to 120. In both cases, we record the network throughput and quantum resource utilization, and show them in Figs. 5 & 6.

In both figures, we can observe that more quantum resources, regardless of quantum memory or quantum channels, can lead to a larger network throughput. By increasing the link capacity, the throughput increase rate would decrease when there are more than 30 quantum channels over each quantum link, since the quantum memory becomes the bottleneck which limits the increase of network throughput. Similarly, when each quantum node hosts more than 90 units of quantum memory, the link capacity becomes the bottleneck and the throughput will increase in a slower pace.

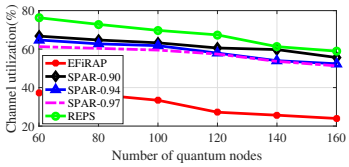
From Figs. 5(b) & 5(c), we can see that a higher link capacity will increase the quantum memory utilization, since more quantum channels will consume more quantum memory. However, the channel utilization itself will decrease, especially



(a) Network scale vs. throughput.

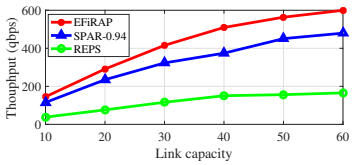


(b) Network scale vs. memory utilization.

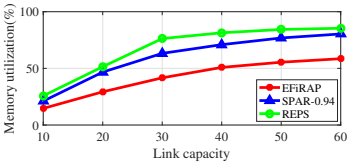


(c) Network scale vs. channel utilization.

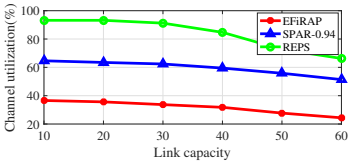
Fig. 4. Effect of network scale.



(a) Link capacity vs. throughput.

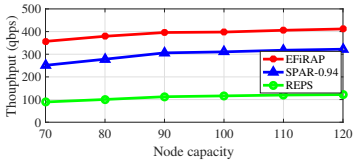


(b) Link capacity vs. memory utilization.

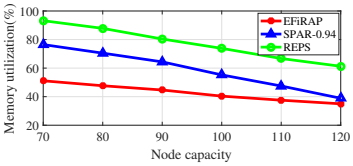


(c) Link capacity vs. channel utilization.

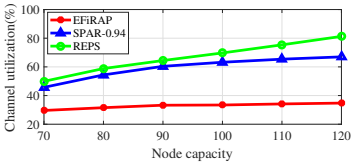
Fig. 5. Effect of link capacity.



(a) Node capacity vs. throughput.

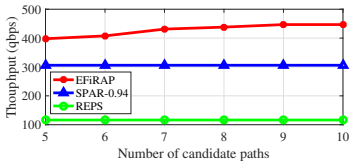


(b) Node capacity vs. memory utilization.

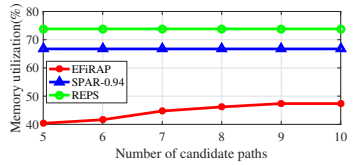


(c) Node capacity vs. channel utilization.

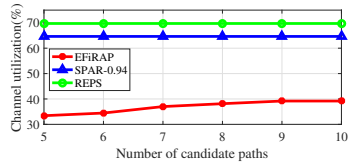
Fig. 6. Effect of node capacity.



(a) Number of candidate paths vs. throughput.



(b) Number of candidate paths vs. memory utilization.



(c) Number of candidate paths vs. channel utilization.

Fig. 7. Effect of number of candidate paths.

when the quantum memory becomes the bottleneck, resulting in a larger decrease in the quantum channel utilization. Similar observations can be made from Figs. 6(b) & 6(c). The increase in quantum memory hosted by each quantum node will result in a larger channel utilization, but the memory utilization will decrease faster and faster.

Effect of number of entanglement paths. In EFiRAP, we first identify K entanglement paths for each SD pair, and then determine the purification schemes. The number of entanglement paths prepared will significantly impact the performance of EFiRAP. We simulate the performance of EFiRAP when the number of entanglement paths prepared for each SD pairs varies from 5 to 10 and show the results in Fig. 7.

In general, preparing more entanglement paths for each SD pair improves the performance of EFiRAP, and results in larger quantum resource utilization, since it provides more options to establish entanglement connections and uses more quantum resources. However, when there are more than 9 entanglement paths for each SD pair, having more entanglement paths cannot further increase the throughput. Over a long entanglement path, more purification is needed to meet the required E2E fidelity F^*. Thus, the EPS algorithm will not select the long paths since it is not resource-efficient to do so. The performances of SPAR-0.94 and REPS don't change with the candidate path number, since they don't need to prepare CEPS.

V. Conclusions

In this paper, two basic ideas have been proposed for the first time, one is to use the E2E fidelity as a metric when establishing an entanglement connection, and the other is to identify critical links to achieve the most resource-effective purification. The paper describes a novel E2E fidelity-aware routing and purification (EFiRAP) approach based on these ideas with the objective of maximizing the throughput in quantum networks with limited quantum resources. Compared with previous works that either ignore fidelity altogether or use fidelity as a metric for entanglement links only, EFiRAP jointly optimizes entanglement routing and purification to guarantee that each established entanglement connection can meet the minimally required E2E fidelity. Extensive simulations show the superior performance of EFiRAP compared with existing solutions in terms of both network throughput and quantum resource utilization.

Acknowledgment

The work of Yangming Zhao has been done when he was a research scientist at UB. The work of Gongming Zhao is partially sponsored by the Anhui Initiative in Quantum Information Technologies under Grant AHY150300.

REFERENCES

[1] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, 1996, p. 212–219.

[2] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 1994, pp. 124–134.

[3] D. Cuomo, M. Caleffi, and A. S. Cacciapuoti, "Towards a distributed quantum computing ecosystem," *IET Quantum Communication*, vol. 1, no. 1, p. 3–8, Jul 2020.

[4] V. S. Denchev and G. Pandurangan, "Distributed quantum computing: A new frontier in distributed systems or science fiction?" *SIGACT News*, vol. 39, no. 3, p. 77–95, sep 2008.

[5] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: A scalable and flexible data center network," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 51–62, Aug. 2009.

[6] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "Bcube: A high performance, server-centric network architecture for modular data centers," in *Proceedings of the ACM SIGCOMM*, 2009, pp. 63–74.

[7] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "Dcell: A scalable and fault-tolerant network structure for data centers," in *Proceedings of the ACM SIGCOMM*, 2008, pp. 75–86.

[8] J. L. Park, "The concept of transition in quantum mechanics," *Foundations of Physics*, vol. 1, p. 23–33, 1970.

[9] S. Shi and C. Qian, "Concurrent entanglement routing for quantum networks: Model and designs," in *Proceedings of the ACM SIGCOMM*, 2020.

[10] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*, 10th ed. USA: Cambridge University Press, 2011.

[11] J.-W. Pan, C. Simon, Časlav Brukner, and A. Zeilinger, "Entanglement purification for quantum communication," p. 1067–1070, 2001.

[12] C. Li, T. Li, Y.-X. Liu, and P. Cappellaro, "Effective routing design for remote entanglement generation on quantum networks," *npj Quantum Information*, vol. 7, 01 2021.

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

[14] G. Q. AI", "Exponential suppression of bit or phase errors with cyclic error correction," p. 383–387, 2021.

[15] D. Ristè, S. Poletto, M.-Z. Huang, A. Bruno, V. Vesterinen, O. Saira, and L. DiCarlo, "Detecting bit-flip errors in a logical qubit using stabilizer measurements," *Nature communications*, vol. 6, 11 2014.

[16] Y. Zhao and C. Qiao, "Redundant entanglement provisioning and selection for throughput maximization in quantum networks," in *Proceedings of the IEEE INFOCOM*, 2021.

[17] 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, Dec. 2019.

[18] S. Rosenblum, Y. Y. Gao, P. Reinhold, C. Wang, C. J. Axline, L. Frunzio, S. M. Girvin, L. Jiang, M. Mirrahimi, M. H. Devoret, and R. J. Schoelkopf, "A cnot gate between multiphoton qubits encoded in two cavities," *Nature Communications*, vol. 9, no. 652, 2018.

[19] W. Kozlowski, A. Dahlberg, and S. Wehner, "Designing a quantum network protocol," in *ACM CoNEXT*, 2020, pp. 1 – 16.

[20] S. Pirandola, "End-to-end capacities of a quantum communication network," *Communications Physics*, vol. 3, pp. 1—-10, 2019.

[21] E. Schoute, L. Mancinska, T. Islam, I. Kerenidis, and S. Wehner, "Shortcuts to quantum network routing," *CoRR*, vol. abs/1610.05238, 2016. [Online]. Available: http://arxiv.org/abs/1610.05238

[22] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the stochastic analysis of a quantum entanglement switch," *SIGMETRICS Perform. Eval. Rev.*, vol. 47, no. 2, pp. 27—-29, dec 2019.

[23] M. Caleffi, "Optimal routing for quantum networks," *IEEE Access*, vol. 5, pp. 22 299–22 312, 2017.

[24] R. V. Meter, T. Satoh, T. D. Ladd, W. J. Munro, and K. Nemoto, "Path selection for quantum repeater networks," *Networking Science*, vol. 3, pp. 82–95, 2013.

[25] K. Benzekki, A. El Fergougui, and A. El Belrhiti El Alaoui, "Software-defined networking (sdn): A survey," *Security and Communication Networks*, vol. 9, no. 18, pp. 5803–5833, 2016.

[26] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese, and D. Walker, "P4: Programming protocol-independent packet processors," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 87—-95, 2014.

[27] J. Y. Yen, "An algorithm for finding shortest routes from all source nodes to a given destination in general networks," *Quarterly of Applied Mathematics*, vol. 27, pp. 526–530, 1970.

[28] B. Korte and R. Schrader, "On the existence of fast approximation schemes," *Nonlinear Programming 4 Academic Press*, pp. 415–437, 1981.

[29] A. Frieze and M. Clarke, "Approximation algorithms for the m-dimensional 0–1 knapsack problem: Worst-case and probabilistic analyses," *European Journal of Operational Research*, vol. 15, pp. 100–109, 1984.

[30] H. Saran and V. V. Vazirani, "Finding k cuts within twice the optimal," *SIAM J. Comput.*, vol. 24, no. 1, pp. 100–108, 1995.