On the Trade-off between Fidelity and Latency for the Quantum Link Layer with few Memories and Entanglement Purification

Karim S. Elsayed*, Wasiur R. KhudaBukhsh†, Amr Rizk*
*Faculty of Computer Science, University of Duisburg-Essen, Germany
†School of Mathematical Sciences, University of Nottingham, UK

***Abstract*—Generating high fidelity link-level entanglements is essential for the distribution of end-to-end connectivity across quantum communication networks. The efficiency of entanglement buffering is, however, limited by the short lifetime of the required qubits as noise causes a rapid decay in their fidelities. Entanglement purification, which is essentially a probabilistic operation, counteracts this decay. In its simplest form it uses two entanglements to obtain a higher fidelity one.**

In this paper, we combine entanglement buffering and purification by applying a purification protocol that acts on the stored entanglements beyond their generation time. Considering a link-level system with a few quantum memories to buffer entanglements in addition to application requests for entanglement, we use a Markov Chain model to derive the steady state distribution of the service performance metrics, specifically, (i) the fidelity of the entanglements that are consumed by application requests and (ii) the distribution of their waiting time in the request queue. Our evaluations show a trade-off between the fidelity and the request waiting time, specifically, given purification as well as for different memory sizes.

I. Introduction

Quantum Networks fundamentally require long-distance quantum communications enabled by the use of entanglement. Particularly, carrying out a quantum application between two nodes requires the consumption of an entangled qubit pair, shortly denoted as entanglement, that is distributed between them [1]. Distributing entanglements between distant nodes requires first creating entanglements on the link level between adjacent nodes [2], [3]. Taking advantage of the capability to buffer qubits using quantum memories, quantum networks use protocols that continuously create, buffer and distribute entanglements among quantum nodes [4].

The quality of the quantum applications depends on the quality of the consumed entanglement, which is measured using a metric denoted as Fidelity. As buffered qubits suffer from inevitable quantum noise, i.e., decoherence, their fidelity decays significantly with time [5], [6]. At the expense of the probabilistic loss of entanglements, *purification* can overcome the loss in fidelity, where it involves, in its simplest form, using two entanglements to generate a higher fidelity one [7]. Consequently, purification leads to a trade-off performance between the generation rate and the fidelity of entanglements.

Purification is mainly applied in quantum networks to generate entanglements with high initial fidelity using recurrence and pumping protocols [8], [9]. Specifically, this involves continually purifying a stored entanglement at the time of the generation of a second one. Considering the continuous generation of entanglements, the authors in [10] incorporate link-level pumping purification to obtain a maximum of one high-fidelity stored entanglement. To allow storing more than one entanglement, the authors in [11] introduce a purification protocol[1] applied on the buffered entanglements beyond their generation time, which was denoted as PBG protocol. Using the PBG protocol raises an interesting question, which we address in this paper, about the trade-off in performance achieved by *the later purification of the entanglements to allow buffering more entanglements.*

The analytical modeling of the link-level system will help in understanding the system performance and dependence on its variables, thus it facilitates its optimization as an integral part of the quantum internet. However, the analytical evaluation of the fidelity of buffered entanglements is challenging, especially considering purification, as the fidelity at a certain time depends on the history of the probabilistic events it has undergone.

To that extent, the authors in [10] derive bounds on the fidelity and the availability of entanglements in a link-level model with one long-term memory at each node. Using the PBG protocol in a model with two long-term quantum memories at each node, the work in [11] analytically derives the steady state distribution of the fidelity and the number of entanglements in the queue without considering any request process. We denote the one long-term memory and the two long-term memory link-level systems as 1M-Pur. and 2M-Pur., respectively.

Taking the request process into account, which influences the steady state of the 2M-Pur. system, we consider the PBG protocol and extend the work in [11] by incorporating a request queue and generalizing the analytical approach under any purification algorithm[2] and initial entanglement generation fidelity.

- We provide a general framework that analytically evaluates the 2M-Pur. model including a request queue for any

[1]The term "protocol" refers to the high level purification approach in terms of parameters like the time of applying purification, the number of rounds, the number of entanglements etc.

[2]The term "algorithm" refers to the type of the purification approach in terms of the hardware operations applied

purification algorithm and any initial fidelity.
- We analytically derive the distribution of the fidelity of the entanglements used for service and the request waiting time.
- We provide numerical evaluations that study the service performance trade-off between different models with different numbers of quantum memories and different considerations of purification.

II. Related work

The continuous generation and buffering of entanglements between nodes provide ready-to-use entanglements for quantum application as introduced in [4]. Other works study different parts of the quantum network under buffering, e.g., the work in [12] evaluates the quantum buffering delay, while the authors in [13] analyzes buffering in the quantum switch.

A detailed analysis of the decaying effect of decoherence on the fidelity of qubits can be found in [14]. Particularly, the phase damping decoherence is one of the dominant types of quantum noise where the fidelity decays exponentially with time [15]. To overcome the fidelity decay, the works in [7], [16], [17] propose different purification algorithms to improve the fidelity of entanglements. On the quantum network level, the authors in [18]–[20] study the effect of purification in the context of the system rates. Specifically, the work in [18] evaluates the quantum key distribution rate while the work in [19] numerically evaluates the hashing rate related to quantum routing. We mainly differ than [18]–[20] as we focus the analytical evaluation of the purification effect on the link-level considering a PBG protocol.

The most similar works addressing the analytical derivation of the performance metrics in a link-level model under purification are [10], [11]. We fundamentally differ from [10] since we use the PBG protocol enabling a larger system consisting of two long-term quantum memories at each node. Additionally, we buffer the requests in case of unavailable entanglements and evaluate their waiting time distribution. We generalize the 2M-Pur. system in [11] to model any purification algorithm under any initial fidelity values and incorporate a request process that consumes the entanglements. Along with the performance trade-off that originates by the application of purification, different systems in terms of the number of memories and the type of the purification protocol leads to a *service performance trade-off*. We evaluate the system in comparison to others and numerically show the service trade-off in Sect. V.

III. Model and Problem statement

We consider the 2M-Pur. link-level quantum system consisting of two quantum nodes, each containing two long-term quantum memories that allow storing a maximum of two entanglements between the two nodes. We denote the two quantum memories at each node as an *entanglement queue* of length two. This model is depicted in Fig. **??**, where we show the following system design aspects: (i) entanglements are generated between the two nodes, (ii) entanglements stored in the entanglement queue possess a time-decaying fidelity, (iii) entanglements can be purified and (iv) application requests for link-level entanglements consume the contents of the entanglement queue.

We consider a link-level entanglement generator with a constant initial generation fidelity $F_0 \leq 1$. We model each link-level entanglement generation as a Bernoulli trial with success probability p_g and generation time duration $\triangle t$. This is a common model in literature since an entanglement generation attempt duration is constant depending on the transmission time in the quantum link [11], [12], [21], [22]. The link-level entanglement generation model we adopt here creates two entangled photons at one node and sends one of them to the other node. In this model, the probability of qubit absorption in the optical fiber is 1pg1-p$_g$$andthedurationofthetransmissionist\triangle t.Eachnodecontainsanext$
$termmemorythatstoresthenewlygeneratedentanglementforminima$
$termmemory.Uponthegenerationofanentanglementgivenafullentan$
we apply a PBG protocol to improve the average fidelity of the entanglements.

Definition 1 (PBG Protocol)**.** *Given an entanglement generation upon a full entanglement queue containing entanglements with fidelities larger than the threshold F_l, the PBG protocol applies purification between the two smallest fidelity entanglements out of the entanglements in the system.*

Given the above definition for a two-memory system (2M-Pur.) the PBG protocol considers the three entanglements available in the system, i.e., the two stored ones in the long-term memories and a newly generated one in the short term memory, and applies purification between the two smallest fidelity entanglements of these three. Entanglements with fidelity levels smaller than F_l are not used for purification but get replaced by the newly generated ones. For modelling purposes we ignore the time taken by the PBG protocol to perform purification.

We introduce a Bernoulli request process that arrives at one of the nodes. During the time slot $\triangle t$, a request for using an entanglement arrives with probability p_r. The node contains a FIFO (First In, First Out) request queue of length m to store the unserved request, in addition to its two quantum memories. When the request queue is full, arriving requests will be dropped. We note that this model of the request process approximates the commonly used continuous time Poisson request process by a discrete geometric distribution. This approximation is reasonable as the time scale of an entanglement generation attempt (in the microseconds) is very small compared to the request time scale [22]. A request service protocol predefines the choice of the entanglement that serves the requests. We choose the service protocol according to the following

Definition 2 (Service Protocol)**.** *The application request consumes the entanglement with the maximum fidelity in the system.*

IV. A DISCRETE-TIME APPROACH

We aim to evaluate the request service performance in terms of the consumed entanglement fidelity and the request waiting time. Therefore, we use a discrete time Markov chain (DTMC) $(\mathbf{Q}, \boldsymbol{\pi}(n), S(n))$ that models the number of requests and the fidelity of the entanglements in the entanglement queue. We use $\mathbf{Q}$, $\boldsymbol{\pi}(n)$ and $S(n)$ to denote the transition matrix, the probability vector at time n and the system state at time n, respectively. Tracking the fidelity decay of the entanglements by the DTMC is attainable by discretizing the continuous fidelity function [11] into countable different levels proportional to the decay within each time slot.

A. Discretization of the Fidelity

We consider phase-damping decoherence, which is one of the most dominant type of quantum noise, as discussed in Sect. II, resulting in an exponential decay of the fidelity [15]. Let the function $A(t, F_u)$ represent the decaying fidelity of an entanglement with fidelity F_u at time 0. Similar to [11], we discretize $A(t, F_u)$ into fidelity levels according to $L_n := A(n\triangle t, F_u)$, where the n-th discrete fidelity level L_n represents the fidelity after n time slots passing each of duration $\triangle t$. We illustrate the discretization of the continuous fidelity for an example of $A(t, F_u)$ in Fig. 2.

In the following, the maximum fidelity value F_u which is solely used to define the fidelity levels is assumed to be constant. The discretization allows tracking the fidelity of an entanglement as it changes by one level each time slot $L_{n+1} = A(\triangle t, L_n)$.

We use the discrete levels to model the fidelity of the entanglements in the system by rounding the fidelity value to the nearest smaller fidelity level. We use this for the generation fidelity F_0 as well as the improved fidelity $F_p(i,j)$ after the purification of two entanglements with fidelity levels L_i and L_j. Specifically, we set F_0 and F_p to the n_0-th and the n_p-th fidelity levels, respectively, according to

$$n_0 = \left\lceil A^{-1}(F_0, F_u) \right\rceil ,$$
$$n_p(i,j) = \max(\left\lceil A^{-1}(F_p(i,j), F_u) \right\rceil, 0) , \quad (1)$$

where A^{-1} is the inverse of the function A and the maximum operation guarantees that $n_p \geq 0$ in case that $F_p > F_u$. In the following, we fix $F_u = 1$ to maximize the number of fidelity levels for accurate modeling, however, we note that for computational reasons F_u that is sufficiently larger than F_0 may suffice. Recall that the result of the purification attempt is probabilistic. Indeed, it is associated with a success probability depending on the fidelities L_i, L_j of the involved entanglements, which is denoted $P_s(i,j)$. Recall that we do not attempt to purify entanglements, but rather replace them, when the fidelity of one of the stored entanglements is L_N according to Def. 1. This is expressed by defining the purification result as $n_p(a, N) = a,\ P_s(a, N) = 1$.

Note that the fidelity of entanglements must exceed a certain quantum application-specific threshold to successfully serve requests [23]. Hence, as we set the number of fidelity levels to N, and fit any unusable entanglement with fidelity strictly less than the given threshold to the N-th level. We calculate N in terms of the fidelity threshold F_l as $N = \max(\lceil A^{-1}(F_l, F_u) \rceil, \lfloor A^{-1}(F_l, F_u) \rfloor + 1)$, where the maximum guarantees that $F_N < F_l$ in the case that F_l is exactly equal to one of the fidelity levels.

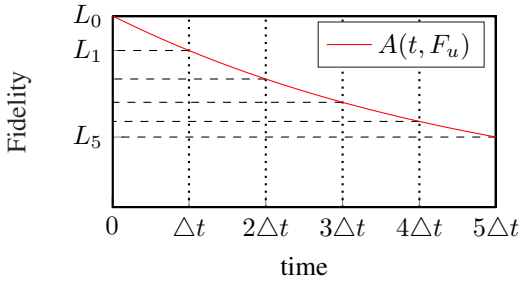

Fig. 2: Discretization of the continuous decaying fidelity $A(t, F_u)$ into the fidelity levels L_n equivalent to the fidelity at the n-th time slot. The fidelity levels are not equidistant since the fidelity decay is exponential.

B. System States

The system states are described by two concurrent Bernoulli trials representing the request arrival and entanglement generation in one time slot constituting the sample space Ω. Simply, the mutually exclusive events are $\mathcal{F} = \{\mathcal{F}_{E\bar{R}}, \mathcal{F}_{E\bar{R}}, \mathcal{F}_{ER}, \mathcal{F}_{\bar{E}\bar{R}}\}$ with probabilities $P(\mathcal{F}_{\bar{E}R}) = P_{\bar{E}R},\ P(\mathcal{F}_{E\bar{R}}) = P_{E\bar{R}},\ P(\mathcal{F}_{ER}) = P_{ER},\ P(\mathcal{F}_{\bar{E}\bar{R}}) = P_{\bar{E}\bar{R}}$. The subscripts $E, \bar{E}$ denote successful and failed entanglement generation, respectively. The subscripts $R, \bar{R}$ denote the arrival and no arrival of a request in that time slot, respectively. Hence, the probabilities of the events are $P_{\bar{E}R} = p_r(1 - p_g),\ P_{E\bar{R}} = (1-p_r)p_g, P_{ER} = p_r p_g,\ P_{\bar{E}\bar{R}} = (1-p_r)(1-p_g)$ given the model and the explanation of p_g, p_r in Sect. III.

Now, recall the system model from Fig. **??**. The state of the system includes the number of the requests in the request queue $k \in \mathcal{N}$ and the fidelity of the two entanglements in the entanglement queue $L_i, L_j \in \{-\infty, 0, 1, \ldots, N\}$. Here, $L_{-\infty}$ is abstract and is solely used to refer to an empty memory in the entanglement queue. We represent the system state as a 3-tuple in terms of the indices of the fidelity levels of the stored entanglements and the number of waiting requests as (i, j, k). Here, j is the fidelity level index of the smaller fidelity entanglement when the entanglement queue is full. It also represents the fidelity level index when only one entanglement is in memory, thus $j \geq i$. Since the request queue starts filling only if the entanglement queue is empty, a state with $k > 0$ implies that $i = j = -\infty$.

We group the states into $N+3$ different blocks depending on the value of i and j. Each row of states in the DTMC depicted in Fig. 3 represents one of the following state blocks. We denote the $m+1$ states describing the queue filling by the state block $S_\phi := \{(-\infty, -\infty, k \geq 0)\}$. Similarly, we denote the states when only one entanglement is in the queue by $S_{-\infty} := \{(-\infty, j \geq 0, 0)\}$. Finally, we group the rest of the states when $i \geq 0$ as $S_i := \{(i, j \geq i, 0)\}$. Additionally, we use the column vector $\boldsymbol{\pi}_\upsilon$ to denote the steady state probability of the states within the state block S_υ for $\upsilon \in \{\phi, -\infty, 0, ..., N\}$.

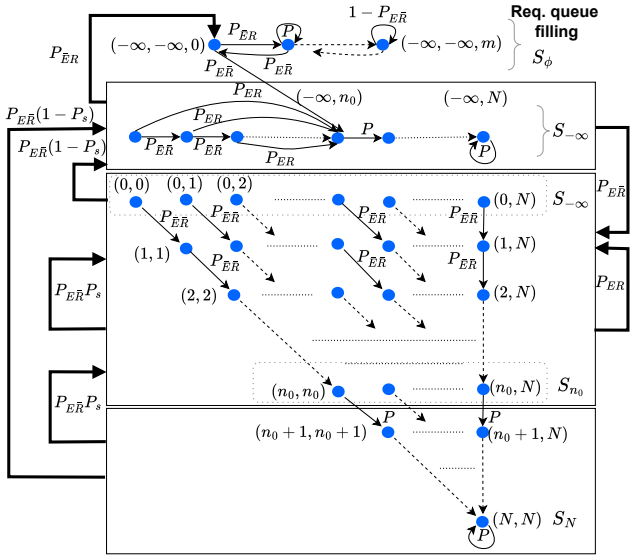

Fig. 3: The DTMC of the two-memory system with PBG purification. A state (i,j,k) refers to the fidelity levels L_i, L_j of the two stored entanglements, respectively, in the order $i \leq j$ and k is the number of the queued requests. For notation simplicity, except for the state $(-\infty,-\infty,0)$, we only write the two-tuple (i,j) when $k=0$. The thick arrows are abstract in the sense that they denote the existence of a transition from every state within a source block to a corresponding state to the destination block, whereas a block self transition does not imply a transition to the same state within the block. The transitions are precisely defined in (2)-(6). Note that $P_s(i,j)$ has a different value depending on the source state and the value of n_0.

C. System transitions and block transition matrix

We describe the system state transitions based on the occurrence of the events described above. First, the occurrence of $\mathcal{F}_{\bar{E}\bar{R}}$, i.e., no generation of an entanglement and no arrival of requests, does not impact the number of the entanglements or requests queued. As a result, the system state transitions only describe the change in the fidelity levels of the entanglements associated with the elapse of one time slot, i.e.,

$$(i,j,k) \to (\omega(i),\omega(j),k) \ , \tag{2}$$

where the function $\omega(x) := \min\{x+1,N\}$ for $x \in \{-\infty,0,\dots,N\}$ increments the x-th level and k being the number of queued requests. The minimum limits the change of the fidelity levels beyond N as we limit the number of levels to N such that L_N represents any fidelity less than the minimum threshold F_l.

Upon the occurrence of $\mathcal{F}_{ER}$, i.e., request arrival and entanglement generation, the number of requests and entanglements do not change as the request will in turn use one of the entanglements. According to the service protocol in Def. 2 the entanglement with the highest fidelity is used and the ones with the least fidelity levels will stay in the system. Thus the system state transits upon the occurrence of $\mathcal{F}_{ER}$ as

$$\begin{aligned}
(-\infty,-\infty,k) &\to (-\infty,-\infty,k) \ , \\
(-\infty,j\neq-\infty,k) &\to (-\infty,j',k) \ , \\
(i\neq-\infty,j,0) &\to \big(\min(i',\omega(j)),\max(i',\omega(j)),0\big) \ , \\
\text{with } i' = \max(\omega(i),n_0) \ &, \ j' = \max(\omega(j),n_0) \ .
\end{aligned} \tag{3}$$

The first line describes the case when the memory is empty and the generated entanglement is directly consumed by the request. The second line describes the case when only one entanglement is in the memory, thus the entanglement with the smaller fidelity j' remains. Similarly, the two entanglements with the smallest fidelities i' and $\omega(j)$ remain in the system in the case of a full entanglement queue as described in the third line of the equation. Note that the transitions in (3) coincide with those in (2) for $i \geq n_0$. This is reflected in the DTMC in Fig. 3 in the transition between state block S_i to the S_{i+1} with respect to $P = P_{\bar{E}\bar{R}} + P_{ER}$.

The occurrence of $\mathcal{F}_{\bar{E}R}$, i.e., the arrival of a request and no generation of an entanglement, results in either draining the entanglement queue by consuming an available entanglement or filling the request queue, otherwise. We summarize the state associated transitions as

$$\begin{aligned}
(-\infty,-\infty,k) &\to (-\infty,-\infty,\min\{k+1,N_R\}) \ , \\
(-\infty,j\neq-\infty,0) &\to (-\infty,-\infty,0) \ , \\
(i\neq-\infty,j,0) &\to (-\infty,\omega(j),0) \ .
\end{aligned} \tag{4}$$

The first line describes the request queue filling when the entanglement queue is empty, while the other lines describe the consumption of the highest fidelity entanglement to serve the request.

Finally, we consider the system transitions for $\mathcal{F}_{E\bar{R}}$, i.e., an entanglement generation and no request arrival. In case of a full entanglement queue, $\mathcal{F}_{E\bar{R}}$ triggers the PBG purification protocol. The system state is first described when a partially filled entanglement queue transits as

$$\begin{aligned}
(-\infty,-\infty,0) &\to (-\infty,n_0,0) \ , \\
(-\infty,j\neq0,0) &\to (\min(\omega(j),n_0),j',0) \ , \\
(-\infty,-\infty,k>0) &\to (-\infty,-\infty,k-1) \ ,
\end{aligned} \tag{5}$$

where the first two lines describe the entanglement queue filling while the third line denotes serving the request queue. Second, we describe the system states when a full entanglement queue changes upon an entanglement generation and no request arrival according to the PBG purification protocol in Def. 1 as

$$\begin{aligned}
&(i\neq-\infty,j,0) \to (-\infty,\hat{i},0), \text{w.p. } P_{E\bar{R}}(1-P_s(i',\omega(j))) \ , \\
&(i\neq-\infty,j,0) \to \Big(\min(\hat{i},\hat{j}),\max(\hat{i},\hat{j}),0\Big) \ , \\
&\qquad\qquad \text{w.p. } P_{E\bar{R}}P_s\,(i',\omega(j)) \ ,
\end{aligned} \tag{6}$$

where $\hat{i} = \min(\omega(i),n_0)$ is the fidelity level index of the entanglement that is not used for purification while $\hat{j} = n_p(i',\omega(j))$ is the fidelity level index resulting from the purification of the two entanglements with the two smallest fidelities. Recall that $P_s(a,b)$ is the success probability of purifying two entanglements with fidelity levels L_a and L_b.

We form the transition matrix $\mathbf{Q}$ of the DTMC depicted in Fig. 3 as a block transition matrix of sub-matrices $\mathbf{Q}_{x,y}$ as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{\phi,\phi} & \mathbf{Q}_{\phi,-\infty} & \mathbf{0} & \dots & \dots & \dots & \mathbf{0} \\ \mathbf{Q}_{-\infty,\phi} & \mathbf{Q}_{-\infty,-\infty} & \dots & \mathbf{Q}_{-\infty,n_0} & \mathbf{0} & \dots & \vdots \\ \mathbf{0} & \vdots & \vdots & \vdots & \mathbf{0} & \ddots & \vdots \\ \vdots & \mathbf{Q}_{n_0,-\infty} & \vdots & \mathbf{Q}_{n_0,n_0} & \mathbf{Q}_{n_0,n_0+1} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \mathbf{0} & \mathbf{Q}_{N-1,N} \\ \mathbf{0} & \mathbf{Q}_{N,-\infty} & \dots & \mathbf{Q}_{N,-n_0} & \mathbf{0} & \mathbf{0} & \mathbf{Q}_{N,N} \end{bmatrix},$$

where $\mathbf{Q}_{x,y}$ contains the transitions defined in (2-6) from the states within S_x to that of S_y. For example, for $P = P_{ER} + P_{\bar{E}\bar{R}}$, the matrices $\mathbf{Q}_{-\infty,-\infty}$ is given by

$$\mathbf{Q}_{\phi,\phi} = \begin{bmatrix} P & P_{\bar{E}R} & 0 & \dots & 0 \\ P_{E\bar{R}} & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & P & P_{\bar{E}R} \\ 0 & \dots & 0 & P_{E\bar{R}} & 1 - P_{E\bar{R}} \end{bmatrix}, \tag{7}$$

and the matrix $\mathbf{Q}_{-\infty,\phi} = P_{\bar{E}R}\,[\mathbf{e}_{N+1}|\mathbf{0}]$, where the vector $\mathbf{e}_{N+1}$ represents an all-one column vector of length $N+1$ and $[.|.]$ depicts the column-wise concatenation operation.

The transition matrix $\mathbf{Q}$ contains a sparse part depicted by the transitions to the state blocks $S_i : i > n_0$, which originates only as a result of the concurrent success ($\mathcal{F}_{ER}$) and the concurrent failure ($\mathcal{F}_{\bar{E}\bar{R}}$) of the generation of an entanglement and a request arrival as given in (2) and (3). Recall that the transitions in (2) and (3) coincide for $S_i : i \geq n_0$ and result in the transitions from each state block S_i to the next S_{i+1}. Now, the transitions to the other state blocks are possible from any state block mainly as a result of the PBG purification protocol attempt described in (6) resulting in the non-sparse part in $\mathbf{Q}$. Specifically, a successful PBG purification attempt results in the transition to a state within the block $S_i : 0 \leq i \leq n_0$, while a failed attempt results in a transition to a state within the block $S_{-\infty}$ describing that only one entanglement is in the queue.

D. Steady state distribution of the DTMC

Motivated by the sparsity of the transition matrix, we derive a reduced linear system of equations (LSE) to obtain the steady state of the DTMC with a constant generation fidelity $F_0 \leq 1$, i.e., we obtain the steady state distribution of the fidelity levels of the link-level quantum system. Recall from Sect. IV-A that n_0 is the fidelity level corresponding to the generation fidelity F_0. Starting with the classical DTMC solution, we write the balance equations for each state block as

$$\boldsymbol{\pi}_i^T = \sum_j \boldsymbol{\pi}_j^T \mathbf{Q}_{j,i} \,. \tag{8}$$

First, we recursively relate the steady state probabilities $\boldsymbol{\pi}_i$ for all fidelity level indices $i > n_0$ to $\boldsymbol{\pi}_{n_0}$ making use of the sparsity in $\mathbf{Q}$ as

$$\begin{aligned} \boldsymbol{\pi}_i^T &= \boldsymbol{\pi}_{n_0}^T \prod_{l=n_0+1}^{i} \mathbf{Q}_{l-1,l} := \boldsymbol{\pi}_{n_0}^T \boldsymbol{\Psi}_{n_0,i} \,,\ n_0 < i < N \,, \\ \boldsymbol{\pi}_N^T &= \boldsymbol{\pi}_{n_0}^T (1 - Q_{N,N})^{-1} \prod_{l=n_0+1}^{N} \mathbf{Q}_{l-1,l} := \boldsymbol{\pi}_{n_0}^T \psi_{n_0,N} \,. \end{aligned} \tag{9}$$

Recall that the state block S_ϕ contains all the states when the entanglement queue is empty. Let $S_{\phi,k}$ denote a state within the state block S_ϕ, which describes an empty entanglement queue and k requests waiting in the request queue and let $\pi_{\phi,k}$ denote the steady state probability of $S_{\phi,k}$. From the definition of the block matrix $\mathbf{Q}_{\phi,\phi}$ containing the transitions between the states within the state block S_ϕ in (7), we recursively derive the steady state probability of the states within the block S_ϕ in relation to the steady state probability of $S_{\phi,0}$ describing empty entanglement and request queues. Then, we relate the steady state probability of $S_{\phi,0}$ to the steady state probability vector $\boldsymbol{\pi}_{-\infty}$ of the state block $S_{-\infty}$, i.e., the block containing the states when only one entanglement is in the queue. In turn, this allows to derive the steady state probability vector $\boldsymbol{\pi}_\phi$ of the states in the block S_ϕ in terms of $\boldsymbol{\pi}_{-\infty}$. Using (7), we recursively derive $\pi_{\phi,i}$ as

$$\begin{aligned} \pi_{\phi,i} &= P_{\bar{E}R}\pi_{\phi,i-1} + P\pi_{\phi,i} + P_{E\bar{R}}\pi_{\phi,i+1} \,,\ \ 0 < i < m \,, \\ \pi_{\phi,m} &= P_{\bar{E}R}\pi_{\phi,m-1} + (1 - P_{E\bar{R}})\pi_{\phi,m} = \gamma\pi_{\phi,m-1} \,, \end{aligned}$$

with $\gamma := \frac{P_{\bar{E}R}}{P_{E\bar{R}}}$. By direct induction we can relate any state $\pi_{\phi,i}$ to $\pi_{\phi,0}$ through

$$\pi_{\phi,i} = \gamma^i \pi_{\phi,0}, \quad 0 < i \leq m \,. \tag{10}$$

Using the balance equation of the state $S_{\phi,0}$ along with (7), we derive $\pi_{\phi,0}$ in terms of $\pi_{-\infty}$ as

$$\begin{aligned} \pi_{\phi,0} &= P\pi_{\phi,0} + P_{E\bar{R}}\pi_{\phi,1} + P_{\bar{E}R}\boldsymbol{\pi}_{-\infty}^T \boldsymbol{e}_{N+1} \\ &= \gamma \boldsymbol{\pi}_{-\infty} \boldsymbol{e}_{N+1} \,. \end{aligned}$$

Let $\boldsymbol{\rho} := [1, \gamma, \gamma^2, \dots, \gamma^N]^T$, using (10), we rewrite $\boldsymbol{\pi}_\phi$ in a vector form in terms of $\boldsymbol{\pi}_{-\infty}$ as

$$\boldsymbol{\pi}_\phi = \gamma \boldsymbol{\pi}_{-\infty}^T \boldsymbol{e}_{N+1} \boldsymbol{\rho}^T := \boldsymbol{\pi}_{-\infty}^T \boldsymbol{\Psi}_{-\infty,\phi} \,. \tag{11}$$

Finally, Let the states $S_{n_0^-} = \{(i,j,0) : -\infty < i \leq n_0\}$ denote the state blocks corresponding to the non-sparse part in $\mathbf{Q}$. We derive the reduced LSE in terms of its steady state probability $\boldsymbol{\pi}_{n_0^-}$ by first expressing the balance equation for $S_{n_0^-}$ as

$$\boldsymbol{\pi}_{n_0^-}^T = \boldsymbol{\pi}_{n_0^-}^T \mathbf{Q}_{n_0^-,n_0^-} + \boldsymbol{\pi}_\phi^T \mathbf{Q}_{\phi,n_0^-} + \sum_{i=n_0+1}^{N} \boldsymbol{\pi}_i^T \mathbf{Q}_{i,n_0^-} \,. \tag{12}$$

Using (9) and (11) we obtain

$$\begin{aligned} \boldsymbol{\pi}_{n_0^-}^T = \boldsymbol{\pi}_{n_0^-}^T \mathbf{Q}_{n_0^-,n_0^-} &+ \boldsymbol{\pi}_{-\infty}^T \boldsymbol{\Psi}_{-\infty,\phi} \mathbf{Q}_{\phi,n_0^-} \\ &+ \boldsymbol{\pi}_{n_0}^T \sum_{i=n_0+1}^{N-1} \boldsymbol{\Psi}_{n_0,i} \mathbf{Q}_{i,n_0^-} + \boldsymbol{\pi}_{n_0}^T \psi_{n_0,N} \mathbf{Q}_{N,n_0^-} \,. \end{aligned} \tag{13}$$

Since the above expression is only in terms of $\boldsymbol{\pi}_{n_0^-}$, we rewrite it in a matrix form as

$$\boldsymbol{\pi}_{n_0^-}^T \left[\mathbf{I} - \mathbf{H}\right] = \mathbf{0} \,, \tag{14}$$

where $\boldsymbol{\pi}_{n_0^-}^T \mathbf{H}$ equals the right hand side of (13).

Next, we use the expressions from (9) and (11) in the normalization equation as

$$\begin{aligned}1 =&\boldsymbol{\pi}^T_{n_0^-}\mathbf{e}_u + \boldsymbol{\pi}^T_{\phi}\mathbf{e}_{m+1} + \sum_{i=n_0+1}^{N} \boldsymbol{\pi}^T_i \mathbf{e}_{N-i+1} \\ =&\boldsymbol{\pi}^T_{n_0^-}\mathbf{e}_u + \boldsymbol{\pi}^T_{-\infty}\boldsymbol{\Psi}_{-\infty,\phi}\mathbf{e}_{m+1} + \boldsymbol{\pi}^T_{n_0}\sum_{i=n_0+1}^{N-1} \boldsymbol{\Psi}_{n_0,\phi}\mathbf{e}_{N-i+1} \\ &+ \boldsymbol{\pi}^T_{n_0}\psi_{n_0,N}\,, \end{aligned} \tag{15}$$

where u is the number of the states within $S_{n_0^-}$. Similar to (14), we rewrite the above expressions in a shorter vector form as $\boldsymbol{\pi}^T_{n_0^-}\boldsymbol{\beta} = 1$. Using this short form of normalization equation along with that in (14), we obtain the short form expression of the reduced LSE as

$$\boldsymbol{\pi}^T_{n_0^-}\left[\mathbf{I} - \mathbf{H}|\boldsymbol{\beta}\right] = [\mathbf{0}|1]\,. \tag{16}$$

In summary, in this section we showed how to leverage the sparsity and structure of the DTMC describing the states of the two-memory link level quantum system with purification to reduce the linear system of equations that is required to compute the distribution of the fidelity levels in the system.

V. Numerical Evaluations

A. Performance metrics

In this subsection, we introduce the performance metrics used for evaluating the system and show their analytical derivation from the steady state solution of the DTMC. We evaluate the service provided to the requests based on the following metrics: (i) The fidelity level of the entanglements chosen to serve the request, denoted as the service fidelity level (SFL) index (n_s) (the lower the larger the fidelity of the consumed entanglements), (ii) the probability of service failure (ε), and (iii) the request waiting time (W) until it consumes an entanglement.

To derive the distribution of the service fidelity level n_s, we define the function $\hat{f}(S)$ that determines the service fidelity according to the service protocol (see Def. 2) given a request to the system in state $S = (i, j, k)$ with two stored entanglements of fidelity levels with indices i and j and the number of queued requests k. We define $\hat{f}(S)$ as follows: In the case of concurrent success of the entanglement generation and request arrival, i.e., $\mathcal{F}_{ER}$, as

$$\begin{aligned}\hat{f}(S|\mathcal{F}_{ER}) = n_0 1_{j<0} + \min(\omega(i), n_0)1_{i\geq 0} \\ + \min(\omega(j), n_0)1_{i<0}1_{j\geq 0}\,,\end{aligned}$$

where $1_{\{.\}}$ is the indicator function and ω as defined in Sect. IV-C. Further, we define $\hat{f}(S)$ in the case of failed entanglement generation associated with successful request arrival, i.e., $\mathcal{F}_{\bar{E}R}$, as

$$\hat{f}(S|\mathcal{F}_{\bar{E}R}) := n_0 1_{j<0} + \omega(i)1_{i\geq 0} + \omega(j)1_{i<0}1_{j\geq 0}\,.$$

Using $\hat{f}(S)$ we calculate the distribution $f_{n_s}(a)$ of the service fidelity level n_s as

Phase damping [20]	$A(t, F_u) = \frac{1}{2}\left(1 + (2F_u - 1)\,\mathbf{e}^{-\frac{t}{1\text{ms}}}\right)$
Polarization mode [16] Dispersion purification	$P_s(i,j) = L_iL_j + (1-L_i)(1-L_j)$ $F_p(i,j) = L_iL_j/P_s(i,j)$
Entanglement success [11]	$p_g = \mathbf{e}^{-(0.15\text{dB/km})l}$
Discretization	$F_u = 1, F_l = 0.55$
Link length	$l = 15\text{km}$
Initial fidelity index	$n_0 = 7$
Request queue length	$m = 4$

TABLE I: Numerical evaluation parameters

$$\begin{aligned}f_{n_s}(a) &= (1-\varepsilon)^{-1}\sum_S P[n_s = a|S]P[S] \\ &= (1-\varepsilon)^{-1}\sum_S \pi_S\big[1_{a=\hat{f}(S|\mathcal{F}_{ER})}p_g \\ &\qquad + 1_{a=\hat{f}(S|\mathcal{F}_{\bar{E}R})}1_{(k<m)}(1-p_g)\big],\end{aligned}$$

with the probability of service failure ε. Note that we only calculate the $f_{n_s}(a)$ for the successfully served requests, i.e., $f_{n_s}(a)$ is the conditional distribution given successful service. The service failure probability ε is given as $(1-p_g)\pi_{\phi,m}$, where $\pi_{\phi,m}$ is the steady state probability of the full request queue state, i.e., $(-\infty, -\infty, m)$.

The requests experience non-zero waiting time when the entanglement queue is empty, where an arriving request waits a random time depending on the number of already waiting requests k. Let W be a random variable denoting the discrete waiting time of an arriving request that is not dropped due to full request queue. The density of the request waiting time $f_W(w)$ is given through marginalization as

$$\begin{aligned}f_W(w) &= (1-\varepsilon)^{-1}\sum_S \pi_S f_W(w|S) \\ &= (1-\varepsilon)^{-1}\sum_S \pi_S\big[1_{j<0}f_W(w|j<0) + 1_{j\geq 0}1_{w=0}\big]\,,\end{aligned} \tag{17}$$

where $f_W(w|j<0)$ is the waiting time distribution when the entanglement queue is empty, i.e., $S = (i,j,k) = (-\infty, -\infty, k)$, while the term $1_{j\geq 0}1_{w=0}$ implies the deterministic zero-waiting time associated with serving the arriving request directly when the entanglement queue is not empty. The conditional distribution $f_W(w|j<0)$ depends on the outcome of the entanglement generation attempt concurrent to the new request arrival, which we express as

$$\begin{aligned}f_W(w|j<0) &= p_g f_W(w|j<0, F_{ER}) \\ &\quad +(1-p_g)1_{k<m}f_W(w|j<0, F_{\bar{E}R}) \\ &= p_g^{k+1}(1-p_g)^{w-k}\left[\binom{w-1}{k-1} + 1_{k<m}\binom{w-1}{k}\right]\,.\end{aligned} \tag{18}$$

Here the conditional waiting time distributions $f_W(w|j<0, F_{ER})$ and $f_W(w|j<0, F_{\bar{E}R})$, i.e., given concurrent successful or failed entanglement generation, respectively follow a negative binomial distribution. The first represents the distribution of the waiting time of k entanglement generations and the latter $k+1$ entanglement generations.

B. Model evaluation and comparison

Next, we evaluate the effect of the PBG purification protocol on the request service in the two-memory system. Additionally, we show the trade-off in the service provided when

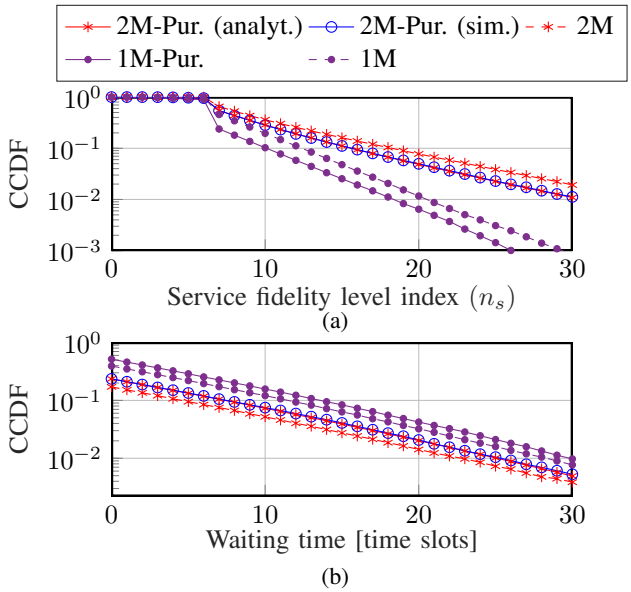

Fig. 4: Comparison between the two-memory (2M) entanglement queue system with and without PBG purification and the one-memory (1M) system in terms of (a) the service fidelity (SFL) index (n_s) (lower is better) and (b) the request waiting time performance metrics. The utilization $p_r/p_g = 0.5$. Note that the distributions are discrete and the lines are only for visual tractability.

the system contains one or two memories at each node. The purification associated with the one-memory system attempts to purify the stored entanglement upon the generation of a new entanglement (as proposed in [10]). Hence, we evaluate the following cases: The two-memory system with PBG (2M-Pur.), the two-memory system without purification (2M), the One-memory system with purification (1M-Pur.) and the One-memory system without purification (1M.). When not applying purification (for 1M. and 2M. systems), the oldest entanglement is replaced by the newly generated one when the memories are full, i.e., dropping entanglements from the entanglement queue is FIFO. If not mentioned otherwise, we use the model parameters for the evaluations from Tab. I.

In Fig. 4, we show the complementary cumulative distribution function (CCDF) of the SFL index and the waiting time for the one and two memory systems with and without PBG purification. Recall that for the fidelity indices, especially the SFL n_s, a lower index is better.

We use simulations to validate the analytical result from Sect. IV for the 2M-Pur. system. First we consider the two-memory system: The upper figure shows the improvement of the service fidelity (i.e. lower SFL index n_s) as a result of the purification protocol PBG. This happens at the expense of larger request waiting times that originate in part from the probabilistic loss of the entanglements as a result of the purification failure.

Now, the one-memory system shows a trade-off between service fidelity and the waiting time. Comparing the 1M-Pur. and 1M curves in Fig. 4 to the 2M-Pur. curves, we find that the one-memory cases show better service fidelity.

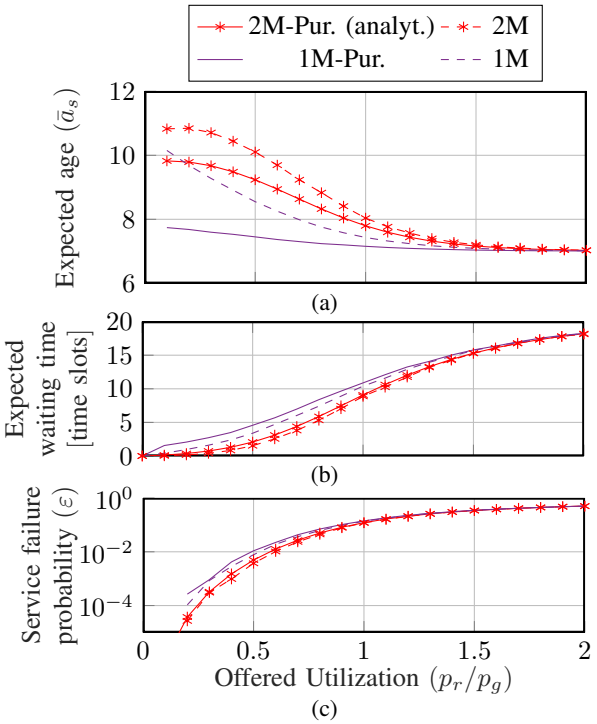

Fig. 5: Comparison between the two-memory (2M) entanglement queue system and the one-memory (1M) system in terms of (a) the mean service fidelity level (SFL) index ($\bar{n}_s$) (lower is better), (b) the mean waiting time and (c) the service failure probability (ε) for an increasing utilization. The choice of entanglement queue length as well as the application of purification provide a trade-off between the performance metrics, which diminishes for large utilization.

Here, the single available entanglement always gets refreshed either by replacement with the newly generated one in the 1M system or by improving its fidelity due to purification in the 1M-Pur. system. Note that since the improvement in service fidelity in the 1M system is limited to the initial fidelity, the 2M-Pur. can show better service fidelity depending on the system parameters as discussed in Fig. 5. This service fidelity improvement in comparison to the two memory cases comes at the expense of *much longer waiting times*. We conclude that the entanglement queue length and the application of purification offer a trade-off between the service fidelity and the waiting time of entanglement requests.

We evaluate in Fig. 5 the expected performance metrics for increasing utilization. For increasing offered utilization p_r/p_g, the expected SFL index $\bar{n}_s := \mathsf{E}[n_s]$ decreases and converges to the initial fidelity generation index n_0 as shown in Fig. 5a. As the request rate increases, the entanglements do not wait before getting consumed by requests, thus, suffer from less decoherence. However, for increasing offered utilization the probability of service failure due to full (finite) request queue increases as well as the expected waiting time as shown in Fig. 5b.

We evaluate the expected SFL index when $n_0 = 3$ and $n_0 = 7$, to show that the general behavior of the four settings remains unchanged independent of the value of the initial

fidelity index n_0. Additionally, we highlight in Fig. 5a that the 2M-Pur. system shows better service fidelity at very low utilization compared to the 1M system when the initial fidelity index $n_0 = 7$. This is because at very at low utilization, an entanglement in 2M-pur. can be purified enough times before service such that its fidelity exceeds that of the 1M system.

Note that the differences between the four settings as shown in Fig. 5 diminishes under *high* offered utilization. At high offered utilization, much more requests arrive than available entanglements, the entanglement queue is mostly empty, and the effect of its length as well as the purification diminishes. At low offered utilization, the entanglement queue is mostly non-empty leading to a high use of purification which creates the gap in the expected service fidelity index $\bar{n}_s$.

VI. Discussion & Open problems

In this paper, we provided an analytical framework that models the fidelity of stored entanglements together with requests for the link-level quantum communication system containing two long-term memories at each node. Using this model, we derive the steady state distribution of the service fidelity and request waiting time as well as the service failure probability. Our evaluations show a trade-off in the fidelity served to the requests as a result of applying purification in combination with the number of available memories.

Compared to the performance of exact continuous fidelity models, the discrete model bounds the fidelity after purification from below. Specifically, the derived metrics represent a lower bound on the average fidelity and an upper bound on the average waiting time and service failure.

One way to increase the accuracy of the discrete model at the expense of computational effort, is to use a predefined grid of discrete fidelity levels, e.g., equidistant values, instead of using the values proportional to the time slot decay as introduced here. Note that the transition matrix $\mathbf{Q}$ loses its recursive structure shown here and the LSE is not guaranteed to be computationally reducible as shown.

Our discrete model can accommodate different request queuing policies, which only affects the formulation of the performance metrics from the DTMC steady state solution shown before. In general, the model allows formulating joint compute and communication qubit memory management policies. Future work directions may include tracking the request fidelity for short request queues jointly with the entanglement fidelity.

Acknowledgement

This work has been funded by the German Research Foundation (DFG) as part of project B4 within the Collaborative Research Center (CRC) 1053 - MAKI.

References

[1] A. S. Cacciapuoti, M. Caleffi, R. Van Meter, and L. Hanzo, "When entanglement meets classical communications: Quantum teleportation for the quantum internet," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3808–3833, 2020.

[2] A. Fischer and D. Towsley, "Distributing graph states across quantum networks," in *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, 2021, pp. 324–333.

[3] W. Kozlowski, A. Dahlberg, and S. Wehner, "Designing a quantum network protocol," in *Proceedings of the international conference on emerging networking experiments and technologies*, 2020, pp. 1–16.

[4] Á. G. Iñesta and S. Wehner, "Performance metrics for the continuous distribution of entanglement in multiuser quantum networks," *Physical Review A*, vol. 108, no. 5, p. 052615, 2023.

[5] A. S. Cacciapuoti and M. Caleffi, "Toward the quantum internet: A directional-dependent noise model for quantum signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7978–7982.

[6] M. A. Schlosshauer, *Decoherence: and the quantum-to-classical transition*. Springer Science & Business Media, 2007.

[7] D. Deutsch, A. Ekert *et al.*, "Quantum privacy amplification and the security of quantum cryptography over noisy channels," *Physical review letters*, vol. 77, no. 13, p. 2818, 1996.

[8] C. H. Bennett *et al.*, "Purification of noisy entanglement and faithful teleportation via noisy channels," *Physical review letters*, vol. 76, no. 5, p. 722, 1996.

[9] W. Dür, H.-J. Briegel *et al.*, "Quantum repeaters based on entanglement purification," *Physical Review A*, vol. 59, no. 1, p. 169, 1999.

[10] B. Davies, Á. G. Iñesta, and S. Wehner, "Entanglement buffering with two quantum memories," *arXiv preprint arXiv:2311.10052*, 2023.

[11] K. Elsayed, W. R. KhudaBukhsh, and A. Rizk, "On the fidelity distribution of link-level entanglements under purification," *arXiv preprint arXiv:2310.18198*, 2023.

[12] W. Dai, T. Peng, and M. Z. Win, "Quantum queuing delay," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 3, pp. 605–618, 2020.

[13] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the stochastic analysis of a quantum entanglement switch," *ACM SIGMETRICS Performance Evaluation Review*, vol. 47, no. 2, pp. 27–29, 2019.

[14] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010.

[15] L. Hartmann, B. Kraus, H.-J. Briegel, and W. Dür, "Role of memory errors in quantum repeaters," *Physical Review A*, vol. 75, no. 3, p. 032310, 2007.

[16] L. Ruan, B. T. Kirby, M. Brodsky, and M. Z. Win, "Efficient entanglement distillation for quantum channels with polarization mode dispersion," *Physical Review A*, vol. 103, no. 3, p. 032425, 2021.

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

[18] S. Bratzik, S. Abruzzo *et al.*, "Quantum repeaters and quantum key distribution: The impact of entanglement distillation on the secret key rate," *Physical Review A*, vol. 87, no. 6, p. 062335, 2013.

[19] M. Victora, S. Tserkis *et al.*, "Entanglement purification on quantum networks," *Physical Review Research*, vol. 5, no. 3, p. 033171, 2023.

[20] W. J. Munro, K. Azuma, K. Tamaki, and K. Nemoto, "Inside quantum repeaters," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 21, no. 3, pp. 78–90, 2015.

[21] B. Davies, T. Beauchamp, G. Vardoyan, and S. Wehner, "Tools for the analysis of quantum protocols requiring state generation within a time window," *arXiv preprint arXiv:2304.12673*, 2023.

[22] M. Pompili, S. L. Hermans, S. Baier *et al.*, "Realization of a multinode quantum network of remote solid-state qubits," *Science*, vol. 372, no. 6539, pp. 259–264, 2021.

[23] A. Dahlberg, M. Skrzypczyk, T. Coopmans, *et al.*, "A link layer protocol for quantum networks," in *Proceedings of the ACM special interest group on data communication*, 2019, pp. 159–173.