Dynamic Queuing Analysis and Buffer Management for Entanglement Swapping Buffers with Noise

Zhouyu Li, Huayue Gu, Xiaojian Wang, Ruozhou Yu
North Carolina State University

ABSTRACT

Entanglement swapping is a core operation in a quantum network. It consumes a pair of entanglements to build a remote entanglement between two parties without direct interaction. In a buffered quantum network, unpaired entanglements can be stored in a quantum buffer for future uses. However, suffering from noises in the quantum buffer, fidelities of buffered entanglements degrade exponentially over time. Entanglements with low fidelity are no longer suitable for certain quantum applications and ought to be discarded. This paper analyzes the dynamic queuing process for a pair of link-level quantum buffers in entanglement swapping. By modeling the quantum buffer pair as a double-sided queue, we derive a closed-form buffering time distribution for every incoming entanglement with respect to the real-time buffer backlog. The distribution reveals the probability that entanglement will be discarded due to low fidelity and helps us design an active buffer management policy that controls the buffer backlog with negligible impact on the entanglement swapping throughput. A discrete-time simulator is developed to demonstrate the correctness of our analysis result and validate the effectiveness of our proposed policy.

1 INTRODUCTION

Entanglement swapping assists in distributing entanglement between two remote parties who do not share a direct quantum link. This process involves consuming a pair of entanglements between both of two remote parties and a common quantum repeater, to generate an entanglement between the two parties whose particles have never interacted directly. However, the stochastic entanglement generation process makes it hard to have simultaneously heralded entanglement pairs for entanglement swapping. When only one of the pair of entanglements is ready for swapping, a traditional bufferless quantum repeater would have to drop it since it will decohere almost instantly [16, 20]. This can significantly limit the throughput of entanglement generation for a quantum network, which leads to the recent proposal of buffered quantum repeaters, where these unpairs entanglements can be stored in quantum buffers before being swapped [5, 8, 9]. Meanwhile, entanglements in near-term quantum memory are subject to noise, leading to degradation of entanglement quality (*fidelity*) over time. If the fidelity drops below a certain threshold, the entanglement is no longer suitable for certain quantum applications and must be discarded [19].

To manage quantum buffers effectively and provide sufficient high-fidelity remote entanglements, it is essential to understand the queuing process behind stochastic entanglement buffering. Recent studies on queuing processes of quantum buffers or buffer-like devices [6, 10, 14, 21] have primarily focused on stationary statistical properties such as mean buffer occupancy or queueing delay in stationary distribution. A dynamic queueing analysis based on real-time buffer conditions is lacking yet essential for designing efficient active buffer management policies like the one in classical networks [1]. This paper aims to fill this research gap. Specifically, we study a pair of link-level quantum buffers for entanglement swapping, where entanglements are buffered with a first-in-first-out (FIFO) discipline and swapped immediately upon pairing. When stored in the buffer, entanglements suffer from dephasing noise over time, and will be discarded when fidelity is below a threshold. We derive the distribution of entanglement buffering time, that is, the time an entanglement stays in the quantum buffer, with respect to the buffer backlog when the entanglement arrives. Using the distribution, we design an active buffer management policy to control the quantum buffer backlog within an adequate value and maintain a high *entanglement swapping throughput* (number of swapped entanglements in a period). Evaluation results show the necessity of such an active management protocol designed upon our dynamic queuing analysis. Our main contributions are summarized below:

- We analyze the dynamic queuing process of a link-level quantum buffer pair with dephasing noise, and derive closed-form probability density function of the n^{th} buffered entanglement's buffering time.
- We design an active queuing control policy that upper bounds the quantum buffer to an adequate size under a tolerable risk of swappable entanglement loss, and maintains the entanglement swapping throughput.
- We implement a discrete-time simulator, and validate the correctness of our queuing analysis and effectiveness of our proposed policy.

The rest is organized as follows. Sec. 2 reviews related works. Sec. 3 introduces our quantum buffer pair model. Sec. 4 analyzes the dynamic queuing process of a quantum buffer pair. Sec. 5 defines the risk-aware adequate buffer size and proposes an active buffer management policy. Sec. 6 presents evaluation results. Sec. 7 discusses concerns and future directions. Sec. 8 concludes the paper.

2 RELATED WORKS

Existing literature on quantum queueing analysis mainly focuses on three scenarios: link-level quantum buffer pair, quantum teleportation queue pair, and the quantum switch.

A link-level quantum buffer pair models queueing in entanglement swapping of link-level entanglements. Razavi et al. [15] proposed a partial nesting protocol for quantum buffer management and throughput estimation for multi-hop remote entanglements with or without entanglement purification. Khatri et al. [10] considered one-slot quantum buffer. Three memory management policies were proposed, and the backward recursion policy, though exponentially slow, was proved fidelity-optimal.

A teleportation queue pair contains a data qubit queue and an entanglement queue, where teleportation is performed when both queues are not empty. Dai et al. [6] analyzed a fidelity-agnostic qubit-entanglement queue pair and gave closed-form average queuing delay for the qubit queue, as well as upper bound of the average queue length with cognitive-memory-based policy. Chandra et al. [3] also studied the teleportation queue pair by reducing the double-sided queue to a single queue system. Different queuing disciplines were analyzed with and without push-out.

A quantum switch generalizes a single pair of queues by considering multiple request-entanglement queues to be matched. Panigraphy et al. [14] studied the capacity region of a quantum switch. The conditional yield distribution of different purification protocols was computed, and a max-weight scheduling policy was proposed. Zubeldia et al. [21] extended request-entanglement queues into a Y-topology, where the stability and throughput of two-way and three-way matching were analyzed. Vasantam et al. [18] assumed entanglements would decohere after one time step and proposed a max-weight scheduling policy for the quantum switch.

All the existing work above analyzes the stationary behavior of the queue, which is not suitable for active buffer management. This paper considers paired link-level quantum buffers with enough capacities and dephasing noise, and focuses on the dynamic queuing process of the noisy system. Our dynamic analysis leads to the design of a real-time active buffer management policy that effectively reduces buffer backlog and maintains entanglement swapping throughput.

3 SYSTEM MODEL

3.1 Quantum Buffer Pair

In this paper, we consider a link-level quantum buffer pair in a quantum repeater chain. As shown in Fig. 1, the quantum repeater chain consists of three quantum repeaters S, R, and D connected by two quantum links. Entanglement sources ES_1 and ES_2 reside on the quantum links between

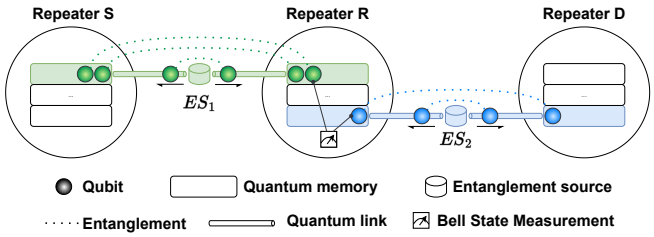


Figure 1: System model. The green and blue quantum buffers form a link-level quantum buffer pair.

repeater pair $S-R$ and $R-D$, respectively. ES_1 and ES_2 generate entangled qubit pairs following independent Poisson processes [6]. For each generated entanglement, the pair of entangled qubits are split and sent to both ends of the quantum link. Quantum repeaters are all equipped with quantum memories. A quantum buffer is a pair of quantum memories at the two repeaters to store their entangled qubits respectively. An entanglement generated between two repeaters, if not immediately consumed, will be stored in the quantum buffer for future usage, following a FIFO discipline. The quantum buffer $\mathcal{E}_1$ between repeaters S and R and the quantum buffer $\mathcal{E}_2$ between repeaters R and D form a **quantum buffer pair**, and their stored entanglements will be used to establish remote entanglements between S and D via entanglement swapping. This pair-and-serve operation consumes one entanglement from each buffer simultaneously and will be conducted whenever there are available entanglements in both quantum buffers of the quantum buffer pair.

3.2 Fidelity Model

Fidelity. Fidelity of an entanglement is the probability that it is measured in its desired pure state upon measurement. Assume an entanglement is desired in the $|\Phi^+\rangle$ state. An entanglement suffering from dephasing noise results in a mixed state $|\psi\rangle$ with fidelity $F \in [0,1]$, written as

$$|\psi\rangle = F \cdot |\Phi^+\rangle\langle\Phi^+| + (1-F)\cdot|\Phi^-\rangle\langle\Phi^-|. \tag{1}$$

Dephasing in quantum buffers. Following existing work [3], we assume entanglements suffer from dephasing noise in the quantum memory, where their fidelity will decrease exponentially over time due to the dephasing noise. According to [13], a single qubit suffering from dephasing noise is modeled with composed influence of the following two operations

$$E_0 = \sqrt{\alpha(t)}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad E_1 = \sqrt{1-\alpha(t)}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{2}$$

where $\alpha(t) = \frac{1+e^{-2\Gamma t}}{2}$ and Γ is the dephasing parameter of the quantum buffer. Assuming the original qubit's density matrix is ρ, state of the influenced qubit ρ' is

$$\rho' = \sum_{i=0,1} E_i \rho E_i^\dagger. \tag{3}$$

Hence phase remains intact with probability $\alpha(t)$ and flips with probability $1-\alpha(t)$ for the qubit. An entangled state is preserved when qubits' phases are either both unchanged or both flipped. For simplicity of presentation, we assume quantum memories have homogeneous dephasing parameters across repeaters, while noting that having heterogeneous noise across repeaters does not invalidate our analysis or design except complicating the mathematical presentation. For an entanglement with initial fidelity F_0, after being stored in a quantum buffer for time t, its fidelity $F(t)$ is

$$F(t) = \left[\alpha^2(t) + (1-\alpha(t))^2\right] F_0 + 2\alpha(t)\,(1-\alpha(t))\,(1-F_0)\,. \tag{4}$$

Replacing $\alpha(t)$ by its definition, we have

$$F(t) = \frac{1 + (2F_0 - 1)\, e^{-4\Gamma t}}{2}. \tag{5}$$

Application and buffer fidelity bounds. Consider now that a pair of entanglements with fidelity F_1 and F_2 are to be swapped. An entanglement swapping uses one-qubit (unitary) and two-qubit operations as well as a Bell state measurement, which may cause fidelity loss due to imperfect quantum operations and measurement. Following the model in [7], assume the fidelity loss factor of one-qubit operation, two-qubit operation, and BSM are α_1, α_2, and α_{BSM}, respectively. Based on the derivation in [8], the fidelity of entangled qubit pair after a successful swapping can be calculated as

$$F' = \frac{1}{2}\left(1 + \alpha_1\alpha_2\alpha_{BSM} \cdot (2F_1 - 1) \cdot (2F_2 - 1)\right). \tag{6}$$

For a quantum buffer, we assume that all the entanglements come with the same initial fidelity and suffer from the same dephasing noise. Denote the initial fidelity of entanglements arriving at buffers $\mathcal{E}_1$ and $\mathcal{E}_2$ as F_1 and F_2, respectively. Assume the target quantum application requires a fidelity lower bound Υ_a for the entanglement acquired via swapping. Υ_a is the *application fidelity bound*. Since entanglement swapping is performed immediately when a pair of entanglements from both buffers become available, at any time only one of the paired quantum buffers can be non-empty. In other words, at most one entanglement has suffered from storage-induced dephasing between a swapped entanglement pair. With Eq. (6), we can calculate the *buffer fidelity bounds* Υ_1 and Υ_2, for $\mathcal{E}_1$ and $\mathcal{E}_2$ respectively, as

$$\Upsilon_i = \frac{1}{2} + \frac{2\Upsilon_a - 1}{2\left(2F_j - 1\right)\alpha_1\alpha_2\alpha_{BSM}}, \tag{7}$$

where $i, j \in \{1, 2\}$ and $i \neq j$.

Maximum Entanglement lifetime. Due to the dephasing noise in the quantum buffer, the fidelity of entanglement will decrease over time following Eq. (5). When an entanglement's fidelity is too low to support a specific application, it will be discarded. Based on the buffer $\mathcal{E}_i$'s dephasing parameter Γ_i and its fidelity bound Υ_i, we can calculate the maximum time θ_i that an entanglement with initial fidelity F_i can stay in $\mathcal{E}_i$ before expired, named *maximum entanglement lifetime*

$$\theta_i = -\frac{1}{4\Gamma_i} \ln\left(\frac{2\left(F_i - \Upsilon_i\right) - 1}{2F_i - 1}\right), \tag{8}$$

where $i \in \{1, 2\}$. As we assume a constant F_i and Υ_i for $i \in \{1, 2\}$, θ_i is also constant in our system model.

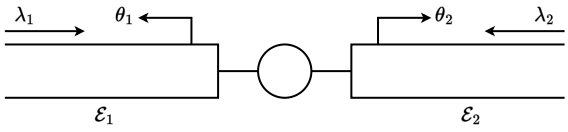


Figure 2: Double-sided queue for quantum buffer pair with noise.

4 QUANTUM QUEUING ANALYSIS

Double-sided queue. We abstract the quantum buffer pair as a double-sided queue with impatient customers, as shown in Fig. 2. Buffer $\mathcal{E}_1$ and $\mathcal{E}_2$ form the left and right sides of the queue, and entanglements are regarded as the queuing customers. The Poisson entanglement generation processes of ES_1 and ES_2 are the arrival process of $\mathcal{E}_1$ and $\mathcal{E}_2$, whose parameters are λ_1 and λ_2, respectively. Entanglement swapping is the service that is immediately performed once both buffers contain available entanglements. Entanglements are buffered and swapped in following a FIFO discipline. Every entanglement buffered to $\mathcal{E}_1$ or $\mathcal{E}_2$ has the patience of its maximum entanglement lifetime θ_1 or θ_2, after which it will be deemed *expired* and forced to leave the system. According to Eq. (8), θ_1 and θ_2 are constant for buffer $\mathcal{E}_1$ and $\mathcal{E}_2$. We also assume $\mathcal{E}_1$ and $\mathcal{E}_2$ have infinite capacity.

Single-sided queue and notations. The double-sided queue with impatient customers has been studied in taxi-customer matching [2, 4, 11]. Because paired entanglements are served instantly, at least one side of the queue is always empty. This allows us to only focus on the non-empty side and re-model the double-sided queue into a single-sided queue. Let $i, j \in \{1, 2\}$ and $i \neq j$. Without ambiguity, we assume hereafter that the *arrival process* always refers to the Poisson entanglement generation process of the entanglement source at the non-empty buffer $\mathcal{E}_i$ with parameter λ_i. The *service process* refers to the Poisson entanglement generation process of the entanglement source at the corresponding empty buffer $\mathcal{E}_j$ with parameter λ_j. Entanglements can stay in the buffer for at most *maximum entanglement lifetime* θ_i, after which they will *expire* and be discarded due to low fidelity. For the n^{th} customer arrives at the buffer, we define the duration between its arrival and its departure, either due to swapping or low fidelity, as its *buffering time*, denoted as $W_{i,n}$, $n \in \mathbb{N}^+$, $0 \leq W_{i,n} \leq \theta_i$. The probability that the n^{th}

customer will be expelled from the buffer due to low fidelity is defined as its *expiration probability*, denoted as γ_n.

Buffering time distribution. Consider the first arriving entanglement at a buffer (*i.e.*, it arrives when the buffer is empty). Because the Poisson service process leads to an exponentially distributed service time, and the buffering time is capped by θ_i when the service takes longer than it, we have the probability density function of $W_{i,1}$ as

$$\Pr\{W_{i,1}=t\}=\begin{cases}\Pr\{W_{i,1}=t\} & t<\theta_i\\ \Pr\{W_{i,1}\geq\theta\} & t=\theta_i\end{cases}=\begin{cases}\lambda_j e^{-\lambda_j t} & t<\theta_i\\ e^{-\lambda_j\theta_i} & t=\theta_i\end{cases}. \tag{9}$$

For the n^{th} customer, its buffering time is the buffering time of the $(n-1)^{\text{th}}$ customer plus the time it waits for being served. We have the following recurrence relation for $W_{i,n}$

$$W_{i,n}=\min\left(W_{i,1}+W_{i,n-1},\theta_i\right)\quad n\geq 2. \tag{10}$$

By solving the recurrence relation, we can get the probability density function of $W_{i,n}$ (with details referred to [12]):

$$\begin{aligned}\Pr\{W_{i,n}=t\}&=\begin{cases}\Pr\{W_{i,n-1}=\tau,W_{i,1}=t-\tau|\tau<t\} & t<\theta_i\\ \Pr\{W_{i,n-1}=\tau,W_{i,1}=\theta-\tau|\tau<\theta\}\\ +\Pr\{W_{i,n-1}=\theta_i,W_{i,1}=0\} & t=\theta_i\end{cases}\\ &=\begin{cases}\dfrac{\lambda_j^n e^{-\lambda_j t}t^{n-1}}{(n-1)!} & t<\theta_i\\ \displaystyle\sum_{k=0}^{n-1}\frac{(\lambda_j\theta_i)^k}{k!}e^{-\lambda_j\theta_i} & t=\theta_i\end{cases}.\end{aligned} \tag{11}$$

Expiration probability. The probability the n^{th} entanglement will expire is equivalent to the probability its buffering time $W_{i,n}=\theta_i$. So we can get the n^{th} customer's expiration probability from Eq. (11)

$$\gamma_n=\Pr\{W_{i,n}=\theta_i\}=\sum_{k=0}^{n-1}\frac{(\lambda_j\theta_i)^k}{k!}e^{-\lambda_j\theta_i}. \tag{12}$$

Expected queuing time. According to Eq. (11), we can also get the expected duration the n^{th} customer stays in the queue

$$\begin{aligned}\mathbb{E}(W_{i,n})&=\int_0^{\theta} t\Pr\{W_{i,1}=t\}\,dt\\ &=\frac{n}{\lambda_j}+(\theta_i-\frac{n}{\lambda_j})\sum_{k=0}^{n}\frac{\lambda_j\theta_i{}^k}{k!}e^{-\lambda_j\theta_i}-\theta\frac{(\lambda_j\theta_i)^n}{n!}e^{-\lambda_j\theta},\end{aligned} \tag{13}$$

where $\mathbb{E}_n(W_{i,n})<\theta_i,\forall n\in\mathbb{N}^+$.

5 ACTIVE BUFFER MANAGEMENT

In this section, we propose an active buffer management policy leveraging the queuing analysis result in Sec. 4, which controls the buffer backlog within an *adequate buffer size*. We first define the risk-aware adequate buffer size for a pair of quantum buffers. Then we introduce how we control the backlog under the adequate buffer size.

Risk-aware adequate buffer size. The core of the active buffer management policy lies in a risk-aware *adequate buffer size*. From Eq. (12), the expiration probability of a newly generated entanglement increases as the buffer backlog $(n-1)$ grows. So if a small probability (**risk**) is tolerable for losing an entanglement that will eventually be swappable before expiration, we can confine the buffer size to a limited value, which is adequate to store most of the swapped entanglements. We call such a buffer size the *adequate buffer size* under a tolerable entanglement loss risk. Given a tolerable entanglement loss risk $r\in[0,1]$, the corresponding expiration probability is $\gamma_{L_r}=1-r$. Though given γ_{L_r}, direct solving Eq. (12) for the adequate buffer size L_r is hard, the corresponding L_r is an integer upper bounded by $e^{\lambda_j\theta_i}$ and can be found efficiently by a binary search on $\left[0,\lceil e^{\lambda_j\theta_i}\rceil\right]$.

Active buffer management policy. Our policy aims to control the size of quantum buffers under an adequate value and maintain the entanglement swapping throughput. Suppose the buffer backlog has already reached the adequate size when a new entanglement arrives in the buffer. In that case, we discard the oldest entanglement (with the lowest fidelity) and store the new entanglement to make sure the backlog is within the adequate buffer size.

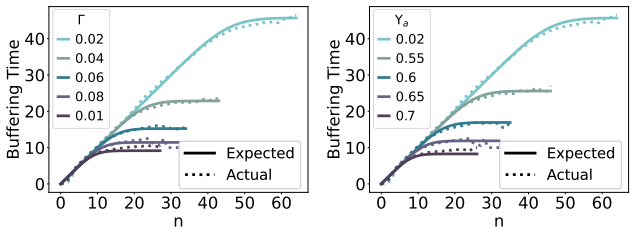
6 EVALUATION

6.1 Experiment Setup

To validate our queuing analysis result and the effectiveness of our proposed active buffer management policy, we implement a discrete time simulator to simulate the link-level quantum buffer pair. We set the entanglement generation for both repeaters to follow a Poisson process with parameter $\lambda_1=\lambda_2=1$. The initial fidelity is set as 0.94 for all the entanglement sources according to [17]. We run the simulator with different application fidelity bounds $\Upsilon_a\in\{0.51,0.55,0.6,0.65,0.7\}$ and dephasing parameter $\Gamma_1=\Gamma_2=\Gamma\in\{0.02,0.04,0.06,0.08,0.1\}$ to show the broad applicability of our derivation result and active buffer management policy. Factors of one-qubit, two-qubit, and Bell state measurement are all set as one. The range of the application fidelity bounds starts from 0.51 instead of 0.5 because it takes infinite time for an entanglement's fidelity to drop to 0.5 according to Eq. (8). For every setup, we run the simulator five times with different random seeds, each time for 10000 time steps. Results produced by different random seeds are averaged out to cancel the randomness.

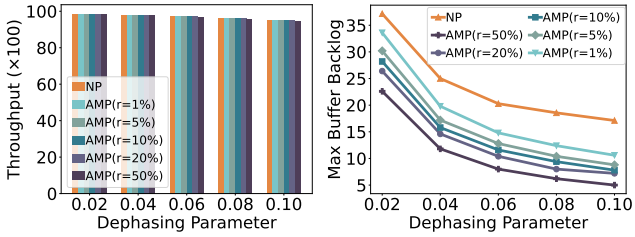
6.2 Expected Buffering Time

We first verify the correctness of the derived expected buffering time Eq. (13) without applying any buffer control policy.



(a) Different Γs with $\Upsilon_a = 0.51$. (b) Different Υ_as with $\Gamma = 0.02$.

Figure 3: Expected and actual buffering time of entanglements for various application fidelity bounds Υ_a and dephasing parameters Γ.



(a) Ent. swapping throughput. (b) Maximum buffer backlog.

Figure 4: Entanglement swapping throughput and maximum buffer backlog with and without policy applied when the application fidelity bound is 0.51 and the dephasing parameter varies.

Dotted lines in Fig. 3(a) and Fig. 3(b) show the average buffering time for entanglement concerning the backlog when it arrives at the buffer, under different application fidelity bounds and dephasing parameters, respectively. We also plot the curve of expected buffering time calculated according to Eq. (13) with solid lines. From both figures, we find negligible differences between the expected curve and the simulated result. The difference slightly increases when n is large, which is due to the lack of enough data points that result in those large buffer backlog values for accurate mean estimation. This validates the correctness of our derived expected buffering time in Eq. (13) computed with the dynamic buffering time distribution in Eq. (11).

6.3 Active Buffer Management Policy

We next evaluate our active buffer management policy under different application fidelity bounds and buffer dephasing parameters. For a fair comparison, in each round of the simulation, we let entanglement sources generate entanglements following the same process for different policies, including no policy applied (**NP**) and the active management policy (**AMP**) with different risks(r). We compare the entanglement swapping throughput (number of swapped entanglements) and the maximum buffer backlogs among the 10000 time steps. Fig. 4 and Fig. 5 show the comparison results for different dephasing parameters and application fidelity bounds.

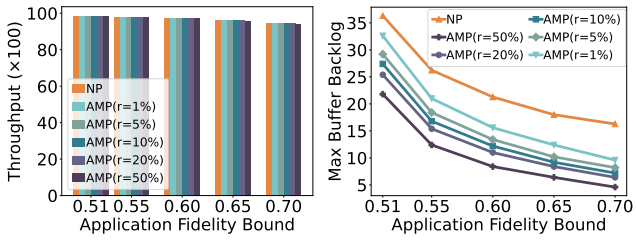
Specifically, Fig. 4 shows that the proposed management policy is effective for various buffer noise levels (dephasing parameters). In Fig. 4(a), the entanglement swapping throughput decreases as the dephasing parameter increases, and a higher risk factor for applied policy lead to lower throughput. However, even applying the policy with the highest tolerable risk 50%, the throughput only drops 0.7% compared with no policy applied. The active management policy with risk factor 1% even maintains the same throughput as no policy applied. On the other hand, Fig. 4(b) shows applying an active management policy with risk factor 1% can obviously decrease the buffer backlog. It saves 9.6% to 38.0% buffer spaces compared to not applying any policy. Increasing the tolerable risk factor can further decrease the maximum buffer backlog. Applying a management policy with 50% risk factor can save up to 70.8% buffer spaces, with only 0.7% throughput trade-off.

Fig. 5 validates the effectiveness of the proposed buffer management policy for different applications (application fidelity bounds). In Fig. 5(a), for the same fidelity bound, applying the buffer management policy does a minor impact on the entanglement swapping throughput. Though the negative impact grows as the risk and fidelity bound increase, applying the highest evaluated risk 50% at the highest fidelity bound 0.7 only causes the average entanglement swapping throughput drop from 9482.2 to 9404, which is a 0.8% decrement. Meanwhile, Fig. 5(b) shows that the buffer management policy with 50% risk can control the average maximum buffer backlog to 4.6, which is only 28.3% of the value (16.28) when no management policy is applied. Even though applying policies with smaller risks increase the maximum backlog, when the risk is 1%, it can still reduce the maximum backlog of non-policy applied from 10.2% to 41.0%, while causing almost zero degradation to the entanglement swapping throughput for different application fidelity bounds.

Hence, our buffer management policy can effectively reduce the maximum buffer backlog while maintaining the entanglement swapping throughput under various buffer conditions for different applications.

7 DISCUSSION

Stability analysis. Different from the analysis of stationary distribution, it is not necessary to analyze the stability condition in dynamic queuing process. This makes our dynamic queuing analysis result applicable to both stable and unstable quantum buffer pairs.



(a) Ent. swapping throughput. **(b) Maximum buffer backlogs.**

Figure 5: Entanglement swapping throughput and maximum buffer backlog with and without policy applied when the dephasing parameter is 0.02 **and the application fidelity bound varies.**

Entanglement purification. Entanglement purification is an operation to improve entanglement fidelities. However, performing entanglement purification is at the cost of potential entanglement loss and can not always improve the entanglement swapping throughput. We conduct preliminary experiment with entanglement purification but its influence on entanglement swapping throughput is not always positive. We leave the analysis of how purification affect the quantum buffer pair to our future works.

Buffer pair parameters. Due to the page limits in our evaluation we only present the result of paired quantum buffers and entanglement sources with the same parameter. Theoratically, our dynamic queuing analysis result still holds for heterogeneous quantum buffers and entanglement sources.

8 CONCLUSION

In this paper, we analyzed the queuing process in a link-level quantum buffer pair with FIFO discipline, and derived the distribution of each entanglement's buffering time with respect to the buffer backlog it encounters. We derived an adequate buffer size with a tolerable loss of swappable entanglement, and designed an active buffer management policy that controls the buffer backlog within the adequate size and maintains the entanglement swapping throughput. A discrete-time simulator was developed to validate our queuing analysis result and the effectiveness of our proposed policy.

REFERENCES

[1] Vamsi Addanki, Maria Apostolaki, Manya Ghobadi, Stefan Schmid, and Laurent Vanbever. 2022. ABM: Active Buffer Management in Datacenters. In *ACM SIGCOMM*. ACM, 36–52.

[2] Francisco Castro, Hamid Nazerzadeh, and Chiwei Yan. 2020. Matching Queues with Reneging: A Product Form Solution. *Queueing Syst* 96, 3-4 (Dec. 2020), 359–385.

[3] Aparimit Chandra, Wenhan Dai, and Don Towsley. 2022. Scheduling Quantum Teleportation with Noisy Memories. (May 2022). arXiv:quant-ph/2205.06300

[4] B.W. Conolly, P.R. Parthasarathy, and N. Selvaraju. 2002. Double-Ended Queues with Impatience. *Computers & Operations Research* 29, 14 (Dec. 2002), 2053–2072.

[5] Wenhan Dai, Tianyi Peng, and Moe Z. Win. 2020. Optimal Remote Entanglement Distribution. *IEEE J. Select. Areas Commun.* 38, 3 (March 2020), 540–556.

[6] Wenhan Dai, Tianyi Peng, and Moe Z. Win. 2020. Quantum Queuing Delay. *IEEE J. Select. Areas Commun.* 38, 3 (March 2020), 605–618.

[7] W. Dür, H.-J. Briegel, J. I. Cirac, and P. Zoller. 1999. Quantum Repeaters Based on Entanglement Purification. *Phys. Rev. A* 59, 1 (Jan. 1999), 169–181. arXiv:quant-ph/9808065

[8] Huayue Gu, Zhouyu Li, Ruozhou Yu, Xiaojian Wang, Fangtong Zhou, and Jianqing Liu. 2023. FENDI: High-Fidelity Entanglement Distribution in the Quantum Internet. (March 2023). arXiv:quant-ph/2301.08269

[9] Huayue Gu, Ruozhou Yu, Zhouyu Li, Xiaojian Wang, and Fangtong Zhou. 2023. ESDI: Entanglement Scheduling and Distribution in the Quantum Internet. (March 2023). arXiv:quant-ph/2303.17540

[10] Sumeet Khatri. 2021. Policies for Elementary Links in a Quantum Network. *Quantum* 5 (Sept. 2021), 537. arXiv:quant-ph/2007.03193

[11] Xin Liu. 2019. Diffusion Approximations for Double-Ended Queues with Reneging in Heavy Traffic. *Queueing Syst* 91, 1-2 (Feb. 2019), 49–87.

[12] A. Movaghar. 1996. On Queueing with Customer Impatience until the Beginning of Service. In *IEEE IPDS*. IEEE Comput. Soc. Press, 150–157.

[13] Michael A. Nielsen and Isaac L. Chuang. 2010. *Quantum Computation and Quantum Information* (10th anniversary ed.). Cambridge University Press.

[14] Nitish K. Panigrahy, Thirupathaiah Vasantam, Don Towsley, and Leandros Tassiulas. 2022. On the Capacity Region of a Quantum Switch with Entanglement Purification. (Dec. 2022). arXiv:quant-ph/2212.01463

[15] M. Razavi, M. Piani, and N. Lütkenhaus. 2009. Quantum Repeaters with Imperfect Memories: Cost and Scalability. *Phys. Rev. A* 80, 3 (Sept. 2009), 032301.

[16] Shouqian Shi and Chen Qian. 2020. Concurrent Entanglement Routing for Quantum Networks: Model and Designs. In *ACM SIGCOMM*. ACM, 62–75.

[17] L. J. Stephenson, D. P. Nadlinger, B. C. Nichol, S. An, P. Drmota, T. G. Ballance, K. Thirumalai, J. F. Goodwin, D. M. Lucas, and C. J. Ballance. 2020. High-Rate, High-Fidelity Entanglement of Qubits Across an Elementary Quantum Network. *Phys. Rev. Lett.* 124, 11 (March 2020), 110501.

[18] Thirupathaiah Vasantam and Don Towsley. 2022. A Throughput Optimal Scheduling Policy for a Quantum Switch. In *Quantum Computing, Communication, and Simulation II*. 22. arXiv:quant-ph/2206.03205

[19] Sören Wengerowsky, Siddarth Koduru Joshi, Fabian Steinlechner, Hannes Hübel, and Rupert Ursin. 2018. An Entanglement-Based Wavelength-Multiplexed Quantum Communication Network. *Nature* 564, 7735 (Dec. 2018), 225–228.

[20] Yangming Zhao, Gongming Zhao, and Chunming Qiao. 2022. E2E Fidelity Aware Routing and Purification for Throughput Maximization in Quantum Networks. In *IEEE INFOCOM*. IEEE, 480–489.

[21] Martin Zubeldia, Prakirt R. Jhunjhunwala, and Siva Theja Maguluri. 2023. Matching Queues with Abandonments in Quantum Switches: Stability and Throughput Analysis. (Jan. 2023). arXiv:cs/2209.12324