2022 IEEE International Conference on Quantum Computing and Engineering (QCE) | 978-1-6654-9113-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/QCE53715.2022.00065

Scheduling Quantum Teleportation with Noisy Memories

Aparimit Chandra[1], Wenhan Dai[1,2], and Don Towsley[1]

[1]College of Information and Computer Science, University of Massachusetts Amherst
[2] Quantum Photonics Laboratory, Massachusetts Institute of Technology
Email: aparimitchan@umass.edu, whdai@cs.umass.edu, and towsley@cs.umass.edu

***Abstract*—Quantum teleportation channels can overcome the effects of photonic loss, a major challenge in the implementation of a quantum network over fiber. Teleportation channels are created by distributing an entangled state between two nodes, which is a probabilistic process requiring classical communication. This causes critical delays that can cause information loss as quantum data suffers from decoherence when stored in memory. In this work, we quantify the effect of decoherence on fidelity at a node in a quantum network due to the storage of qubits in noisy memory platforms. We model a memory platform as a buffer that stores incoming qubits waiting for the creation of a teleportation channel. Memory platforms are parameterized with decoherence rate and buffer size. We show that fidelity at a node is a linear sum of terms, exponentially decaying with time, where the decay rate depends on the decoherence rate of the memory platform. This allows us to utilize Laplace transforms to derive computable functions of average fidelity with respect to the load, buffer size, and decoherence rate of the memory platform. We prove that serving qubits in a Last In First Out order with pushout for buffer overflow management maximizes average fidelity. Last, we apply this framework to model a single repeater node to calculate the average fidelity of the end-to-end entanglement created by this repeater assuming perfect gate operations.**

***Index Terms*—Quantum Networks, Quantum Teleportation, Decoherence, Fidelity, Queuing Theory, Quantum Memory, Quantum Repeaters,**

I. Introduction

Quantum networks face many problems inherently different from those in classical networks, as qubits differ from bits [1]. One of these problems arises from the (in)famous no-cloning theorem [2]. Quantum networks implemented over fiber suffer from exponential photonic loss with respect to fiber length [3]. Classical networks overcome similar loss by using signal amplification. Unfortunately, the no-cloning theorem bars the use of signal amplification, which means quantum networks need to find another solution to the problem of loss. Quantum teleportation allows us to transfer quantum information between two spatially separated parties using a distributed entangled state and classical communication without transferring the physical entity carrying that information across the network [4], making it invulnerable to loss. Another essential property of quantum teleportation is that it allows for secure communication and is a central part of quantum key distribution [1].

Quantum teleportation is enabled through entangled quantum states. The most common examples of entangled states are Einstein–Podolsky–Rosen (EPR) states or Bell Pairs [4]. Therefore, a critical job for quantum networking devices like repeaters, switches, etc., is to distribute these EPR pairs between two nodes so that quantum information can be shared between them, creating a teleportation channel. Many protocols exist for generating and distributing EPR pairs [5], but all are probabilistic processes that can fail because of imperfections in physical operations such as gate errors, signal loss in fiber, etc. These probabilistic failures naturally give rise to many optimization, control, and design problems in quantum networking devices. There have been many recent results regarding the modeling and analysis of entanglement distribution rates for nodes in quantum networks [6], but a significant assumption in most of these is the presence of noiseless memories. This is an issue as most noisy intermediate-scale quantum (NISQ) era quantum memory platforms cause fidelity loss on any qubit stored in them due to decoherence [7].

When a qubit arrives at a node requesting teleportation, the node must await the generation of an EPR pair between itself and the destination, which causes delays. While the EPR pair is being generated, The request qubit has to be stored in a noisy memory platform. This means that even if the EPR pair is perfect, the request will suffer some decoherence. The sensitivity of a memory platform to noise is parameterized by decoherence time, or decoherence rate, and fidelity decays exponentially with time. Quantifying the effects of this decoherence on the fidelity of the teleportation allows us to come up with specifications for memory platforms for different applications. Another vital question is memory management. Given the fact that one can only store a finite amount of requests and EPR pairs, how should one schedule and service requests to minimize decoherence and deal with the arrival of new requests when memory is full.

In this paper, we quantify the fidelity loss for a node in a quantum network due to decoherence from memory and provide a way to derive the fidelity distribution or the average fidelity. We use dephasing noise characterized by a dephasing

This research was supported in part by the NSF grant CNS-1955744, NSF-ERC Center for Quantum Networks grant EEC-1941583, by the National Science Foundation to the Computing Research Association for the CIFellows 2020 Program, and the MURI ARO Grant W911NF2110325.

978-1-6654-9113-6/22/$31.00 ©2022 IEEE
DOI 10.1109/QCE53715.2022.00065

rate Γ to model noise in the memory platform and derive an expression for fidelity with respect to time spent in memory by a request. Furthermore, we model teleportation as a queuing process where requests are generated according to a Poisson process with rate λ and EPR pairs are also generated according to a Poisson process with rate μ. This allows us to calculate the wait times in memory using simple continuous-time Markov models.

The expected fidelity of a qubit teleported by the node depends on load, dephasing rate, memory size, and service discipline. Since the relation between fidelity of a qubit and its age is not linear, the order in which requests are served can affect the expected fidelity of the qubit. This order of service is referred to in this paper as the service discipline. Even when the average wait times of two service disciplines are the same, the average fidelities can differ. We consider both first in first out (FIFO) and last in first out (LIFO) disciplines with both finite and infinite buffers. When buffers are finite, we introduce a pushout buffer management policy. Coupled with FIFO and LIFO, we refer to the combined policies as FIFOPO and LIFOPO respectively. Here pushout operates as follows, if a qubit arrives to a full buffer, the oldest request in the buffer is kicked out to make space for the incoming qubit. We consider pushout because intuitively, it maximizes fidelity as older requests, i.e., qubits that have suffered the most decoherence, are kicked out. We prove that LIFOPO is the optimal discipline for optimizing fidelity.

We consider a scenario where we have two memory platforms, One for storing teleportation requests and one for caching EPR pairs. We model this as two competing queues where at least one is always empty. Lastly, we extend this model to show how this can be applied to calculate the average fidelity of the teleportation channel created by a single repeater chain. The novelty in our construction stems from its simplicity and flexibility for calculating fidelity distributions. It also considers the effects of scheduling disciplines which, to the authors' knowledge, have not been considered in quantum networks at the time of writing. The flexibility of this model also allows for easy extensions to different noise models, probability distributions, etc. This leads to future work in integrating elements from different works. We will go further into this in section VII-A.

A. Related work

As stated, the analysis and modeling of quantum network devices is an active field of study. There have been many studies on modeling switches and repeaters to analyze and design protocols [6], [8], but these studies focus on entanglement generation capacity, not on fidelity. This work can be thought of as adding noisy memory to those studies. [9] focuses on the fidelity of EPR pairs generated by repeater chains of different lengths, but does not account for a continuous stream of requests, so it can be seen as deriving a more accurate distribution for the EPR generation distribution for a node at the beginning of a repeater chain. [10] analyses fidelity loss from wait times in memory using queues and is a very flexible model as it also abstracts hardware implementations and protocols into a set of tunable parameters. However, it focuses primarily on the local network of a quantum processor whereas the model presented in this paper can be extended to model repeater nodes as well as other quantum nodes. They only consider a FIFO queue.

II. System model

In this section, we formally define the process we are modeling. We define the parameters that govern our physical process and how a memory platform in a quantum network node behaves.

Consider two nodes in a quantum network. one node constantly receives quantum information that it must teleport to the other node. We assume this node receives information as pure state qubits arriving according to a Poisson process with rate parameter λ. Any time this node receives a qubit, it teleports it to the other node. This requires the generation and consumption of an EPR pair between itself and the destination. Since it is rarely the case that the initial fidelity of a distributed Bell pair is one, we assume that the generated EPR pairs have initial fidelities of 0.9.

The distribution of EPR pairs between two nodes is a stochastic process [11] where the probability of successful EPR pair generation depends on the distribution protocol and physical implementation of the EPR pair generating platform. If we consider a discrete-time model, the number of time steps required to generate an EPR pair is characterized by a geometric distribution. If we consider the time for one trial to be very small and the probability of successfully generating an EPR pair also to be small, then we can approximate the EPR generation process by a Poisson process [10] where the time taken to generate an EPR pair is sampled from an exponential distribution with mean μ. We define the "load" on a node as λ/μ. We can then model the occupancy (number of stored qubits to be teleported and EPR pairs) of the memory platform as a continuous time Markov process (CTMC) and derive steady-state distributions, or Laplace transforms for the time a qubit spends in memory (wait time). We can then utilize memory error models to obtain statistical descriptions of the final unconditional fidelity of the teleported qubits.

A. Memory Model

A request qubit is stored in a noisy memory when waiting for an EPR pair. A queue forms when more requests arrive while one is already in memory. As stated previously, this allows us to model the memory platform as a CTMC, allowing for the calculation of wait time distributions. To quantify the effects of decoherence, we need a continuous time noise model that captures information loss. We choose the dephasing or the phase damping model [2] represented by the operator $\varepsilon(\rho)$, where ρ is the density matrix of a one qubit system. It is mathematically defined as

$$\varepsilon\left(\begin{bmatrix}\rho_{00} & \rho_{01}\\ \rho_{10} & \rho_{11}\end{bmatrix}\right) = \begin{bmatrix}\rho_{00} & e^{-\Gamma t}\rho_{01}\\ e^{-\Gamma t}\rho_{10} & \rho_{11}\end{bmatrix}, \tag{1}$$

where ρ_{ij} is the ijth entry of the density matrix ρ, Γ is a constant dephasing rate in a given environment, and t is the time elapsed. Dephasing noise is the most common noise associated with memories, and the results of this paper can be extended to account for any error model as long as it can be expressed as a linear sum of exponentially decaying terms.

One important caveat here is that we only account for dephasing errors, not erasure errors. Therefore there is no separate notion of efficiency as is common in the experimental literature [12]. The fidelity measured here is after the qubit has been served, i.e., retrieved from memory. The fidelity distributions derived later represent the probability of the fidelity given that the qubit has been teleported. This assumption can be relaxed, and we discuss this in VII-A.

B. Fidelity loss of a single qubit

Fidelity of a density matrix ρ to some pure state $|\psi\rangle$ is given by the formula

$$F(|\psi\rangle\langle\psi|, \rho) = \text{tr}(|\psi\rangle\langle\psi|\rho). \tag{2}$$

We can use this with (1) to calculate the fidelity of a single qubit $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ after spending time t in memory. We get the formula

$$F(t) = |\alpha|^4 + 2e^{-\Gamma t}|\alpha|^2|\beta|^2 + |\beta|^4, \quad t \geq 0. \tag{3}$$

Note that the fidelity depends on the initial state of the pure qubit, i.e., α and β influence the fidelity loss experienced by that qubit. We need the inverse of $F(t)$ in order to obtain the fidelity distribution from the wait time distribution ($|\alpha|^4 + |\beta|^4 \leq f \leq 1$),

$$F^{-1}(f) = \Gamma^{-1}(\ln(2|\alpha|^2|\beta|^2 - \ln(f - |\alpha|^4 - |\beta|^4)). \tag{4}$$

1) Fidelity loss in a Bell pair: The effect of dephasing on the fidelity of a Bell pair is well studied and is given by

$$F(t) = \frac{1 + e^{-2\Gamma t}}{2}, \quad t \geq 0$$

where t is time spent in the system and Γ is the dephasing rate of the memory [13]. If we consider an initial fidelity of 0.9 at time $t = 0$, it is modified to

$$F(t) = \frac{0.8 + e^{-2\Gamma t}}{2}, \quad t \geq 0. \tag{5}$$

Dephasing causes the Bell state to turn into a mixture of a Bell state and maximally mixed 2-qubit state $I/4$, which can be written in terms of its fidelity with respect to the Bell state as a non-maximally entangled Bell state:

$$\rho_w = \frac{1 - F}{3} I + \frac{4F - 1}{3} |\Phi^+\rangle\langle\Phi^+|.$$

Here F is the fidelity of ρ_w with respect to the Bell pair $|\Phi^+\rangle$. This is precisely the F that decays in (5).

2) Fidelity loss experienced by a qubit due to teleportation by a non maximally entangled state: Teleportation using a maximally entangled Bell pair results in perfect teleportation, and no information is lost. However, this is rarely the case in practice, so we look at how teleportation using a non-maximally entangled Bell state acts as a linear map on the input state. We consider a Werner state ρ_w as the teleportation resource and use it to teleport $\rho(t)$, which is the density matrix for some request qubit that has spent time t in memory. This allows us to represent the effect of teleportation on $\rho(t)$ as a linear map [14]:

$$\Lambda_T(\rho(t)) = \sum_{i,j=0}^{1} \langle\phi_{ij}|\rho_w|\phi_{ij}\rangle \cdot U_{ij}\rho(t)U_{ij}^\dagger, \tag{6}$$

where $|\phi_{ij}\rangle$ are Bell states, $\Lambda_T(\cdot)$ is the standard teleportation algorithm represented as a linear transformation and

$$U_{00} = I, U_{01} = \sigma_x, U_{10} = \sigma_z, U_{11} = i\sigma_y.$$

If F is the fidelity of the Werner state ρ_w with respect to the target Bell state Φ^+, then

$$\Lambda_T(\rho(t)) = F\rho(t) + \frac{1 - F}{3}(\sigma_x\rho(t)\sigma_x^\dagger + \sigma_z\rho(t)\sigma_z^\dagger + i\sigma_y\rho(t)(i\sigma_y)^\dagger).$$

This equation can be further simplified to get an equation for the fidelity of a qubit being teleported by a non-maximally entangled Bell state, both suffering dephasing errors for times t_1 and t_2 respectively. Therefore the final fidelity of the teleported qubit is

$$\begin{aligned}
&\text{tr}\big(\rho(0)\Lambda_T\rho(t_1)\big) \\
&= \frac{0.8 + e^{-2\Gamma t_2}}{2}(|\alpha|^4 + |\beta|^4 + 2e^{-\Gamma t_1}|\alpha|^2|\beta|^2) \\
&+ \frac{1.2 - e^{-2\Gamma t_2}}{6}(4e^{-\Gamma t_1}|\alpha|^2|\beta|^2) \\
&+ \frac{1.2 - e^{-2\Gamma t_2}}{6}(|\alpha|^4 + |\beta|^4 - e^{-\Gamma t_1}((\alpha^*\beta)^2 + (\beta^*\alpha)^2)).
\end{aligned} \tag{7}$$

In the considered model, either the EPR pairs or the request qubits have to be stored in memory, so if $t_1 > 0$, then $t_2 = 0$ and vice versa. Therefore, we further simplify (7) in these cases. If $t_1 = 0$, the error in teleportation is only due to dephasing suffered by the EPR pair, the formula simplifies to

$$F_2(t) = \frac{2.4 + 1.2c_1}{6} + \frac{3 - c_1}{6}e^{-2\Gamma t_2}, \quad t \geq 0$$

where $c_1 = 1 + 2|\alpha|^2|\beta|^2 - (\alpha^*\beta)^2 - (\beta^*\alpha)^2$. When $t_2 = 0$, we obtain the expression in (3):

$$F_1(t) = \frac{5.6}{6}c_2 + \big(\frac{5.8}{3}c_3 - \frac{0.2}{6}c_4\big) + e^{-\Gamma t}, \quad t \geq 0$$

where $c_2 = (|\alpha|^4 + |\beta|^4)$, $c_3 = (|\alpha|^2|\beta|^2)$, and $c_4 = ((\alpha^*\beta)^2 + (\beta^*\alpha)^2))$. The critical observation is that fidelity is a linear sum of exponentially decaying terms. If time is a

random variable and its Laplace transform with parameter s, denoted by $T^*(s)$ is known, we obtain the equation

$$\mathbb{E}[F_i] = c_i + c_j \mathbb{E}[e^{-\Gamma_i T}] = c_i + c_j T^*(\Gamma_i). \tag{8}$$

This approach of using the Laplace transform to get the moments for the fidelity is helpful as in the processes we consider, it is easier to obtain closed form solutions of the Laplace transforms for the wait time distributions than the distributions themselves. This will be especially useful when we consider models with finite memory. In these equations, c_i and c_j are determined by the input qubit being teleported. In this paper, we use $|+\rangle$ as the example input qubit and

$$F_1(t) = \frac{0.9\overline{3}}{2} + \frac{0.9\overline{3}}{2} e^{-\Gamma_1 t}, \quad t \geq 0 \tag{9}$$

$$F_2(t) = \frac{1.8}{3} + \frac{1}{3} e^{-2\Gamma_2 t}, \quad t \geq 0. \tag{10}$$

These are the simplified error models considered in this paper. All of the aforementioned functions are monotonic scalar functions of fidelity in terms of the wait time of a request. Therefore, we can transform wait time into fidelity. Thus it suffices to derive wait time distributions under different service and buffer management policies and then transform to fidelity distributions as we will see in the upcoming sections.

III. Double Queue Model

In this section, we consider a node with a memory platform available for storing multiple EPR pairs, which are generated according to a Poisson process with rate λ_e. Teleportation requests arrive according to a Poisson process with rate λ_r. We assume gate operations are instantaneous as times taken to perform gate operations are orders of magnitude smaller than the time taken to generate an EPR pair. This process can be modeled as two competing queues where the service rate for one queue is the request rate for another. The memory platform for the request qubits can store B_r qubits, and the platform for EPR pairs can store B_e qubits. We model this as a CTMC with the state being the number of request qubits in the system denoted as N. We represent a surplus of EPR pairs as a negative number of requests, making $-B_e \leq N \leq B_r$. This gives us the process presented in Fig. 1 From its Markov chain formulation,

$$\pi_n = \mathbb{P}[N = n] = \pi_{-B_e} \rho^{n+B_e}, \quad -B_e \leq n \leq B_r.$$

Since $\pi_{-B_e} \sum_{i=0}^{B_e+B_r} \rho^i = 1$,

$$\pi_n = \frac{1-\rho}{1-\rho^{B_e+B_r+1}} \rho^{n+B_e}, \quad -B_e \leq n \leq B_r. \tag{11}$$

Let p_e and p_r denote the probabilities that an arriving EPR pair and request are placed in a buffer, respectively. Then

$$p_e = \sum_{n=-B_e}^{0} \pi_n \text{ and } p_r = \sum_{n=0}^{B_r} \pi_n.$$

Let $P_{s,r}$ and $P_{s,e}$ be the probabilities that a request qubit is teleported and an EPR pair is used, respectively. Define: $P_{s,i} = \mathbb{P}[\text{an arrival of type } i \text{ gets served}]$, $i \in \{e, r\}$. Then

$$P_{s,i} = \frac{\sum_{j=0}^{B_i-1} \rho^j}{\sum_{j=0}^{B_i} \rho^j} = \frac{1-\rho^{B_i}}{1-\rho^{B_i+1}}. \tag{12}$$

The Markov chain formulation shows that this system alternates between two phases, as shown in Figure 2. In phase 1, request qubits are stored in memory waiting for EPR pairs and suffer decoherence during the wait. In phase 2, EPR pairs queue up in memory and wait for requests to arrive that they can teleport. The system alternates between these two phases, so either only the request qubit spends time in the memory or the EPR pair but never both. Due to this, We can individually analyze the fidelity loss for each phase and then derive a joint distribution by conditioning on the phase.

Let us take a closer look at the process during a phase. If we restrict ourselves to one phase, the memory platform behaves like a standard finite buffer M/M/1. We know from queuing theory that different orders of service for buffered requests and EPR pairs lead to different wait time distributions. If we have the wait time distributions, we can easily derive fidelity distributions using a Jacobian transformation. When it is too complex to explicitly derive the wait time distribution, it is usually straightforward to calculate the Laplace transform for the wait time and use it to calculate average fidelity.

Let $f_{W_i}(t)$ be the probability density function (pdf) for the wait time incurred by a random request during phase $i = 1, 2$. The qubit fidelity distribution will depend on whether the request qubit or the EPR pair incurred the wait (phases 1 and 2). In our case, we will use (9) or (10) depending on what type of qubit we are considering,

$$f_{F_i}(x) = f_W\big(F_i^{-1}(x)\big) \left| \frac{d}{dy}\big(F_i^{-1}(x)\big) \right|, \quad x \geq 0.$$

We also define $W_i^*(s) = \mathbb{E}[e^{-st}]$, i.e., the Laplace transform of the wait time during phase $i = 1, 2$. We use (8) to calculate $\mathbb{E}[F_i]$ given W_i^*.

In the next section, we give explicit expressions for $f_{F_i}(t)$ or $W_i^*(s)$ for four memory platforms differentiated by the type of memory management or service discipline used. Meanwhile, assuming we have descriptions for the wait times of the two individual queues, we can get the fidelity probability distribution of served requests accounting for both phases by adding the conditional distributions of fidelity of a served request which waited in a particular queue and normalizing it. This yields

$$f_F(x) = \frac{\lambda_e p_e P_{s,e} f_{F_e}(x) + \lambda_r p_r P_{s,r} f_{F_r}(x)}{\lambda_e p_e P_{s,e} + \lambda_r p_r P_{s,r}}, \quad x \geq 0. \tag{13}$$

Therefore,

$$\mathbb{E}[F] = \frac{\lambda_e p_e P_{s,e} \mathbb{E}[F_e] + \lambda_r p_r P_{s,r} \mathbb{E}[F_r]}{\lambda_e p_e P_{s,e} + \lambda_r p_r P_{s,r}}. \tag{14}$$

This expression is very flexible as it allows us to calculate the average fidelity for double queue models even when the

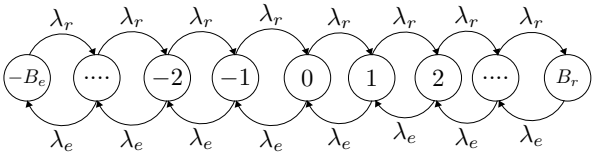

Fig. 1: Markov Chain formulation of Double Queue model with buffer size B_e for EPR pairs and B_r for request qubits

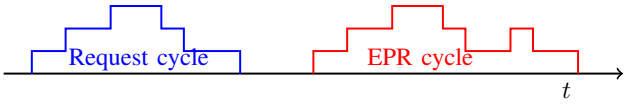

Fig. 2: Typical teleportation behavior.

memory platforms have different decoherence rates, buffer sizes, control, etc.

IV. Single Queue Models

In this section, we take a closer look at the phases mentioned in the previous section by focusing on a single memory platform for the incoming "request" with no memory for the "service". Applied to the double queue model, we must be careful whether the "request" is an actual request qubit or an EPR pair, as they will flip depending on which phase we are in. Once the request arrives in memory, it waits for service. As stated before, inter-arrival and service times are sampled from exponential distributions reducing the process to an M/M/1 queue. We will now take a look at five different kinds of queues, each modeling a different kind of memory control for the platform

A. *Infinite buffer* FIFO

The buffer size is infinite, and no incoming request is blocked. The buffer serves the qubits using FIFO. We know from literature [18] that

$$f_W(t) = (\lambda_e - \lambda_r)e^{-(\lambda_e-\lambda_r)t}, \quad t \geq 0$$

where λ_r is the arrival rate, and λ_e is the service rate. We can transform this into the probability density function for the fidelity using the formula derived in (3) as the function is scalar and monotonic to get

$$f_F(x) = \frac{(\lambda_e - \lambda_r)}{\exp(\ln|\alpha|^4 + |\beta|^4 - \ln x - 2|\alpha|^2|\beta|^2)^{(\lambda_e-\lambda_r)/\Gamma}} \cdot |\frac{\Gamma^{-1}}{(|\alpha|^4 + |\beta|^4 - x)}|, \quad x \geq 0. \tag{15}$$

The Laplace transform for $f_W(t)$ is

$$W^*(s) = \frac{\lambda_e - \lambda_r}{\lambda_e - \lambda_r - s}, \quad \text{Re}(s) \geq 0. \tag{16}$$

One key thing about this model is that it is only valid if $\lambda_r < \lambda_e$; otherwise the queue keeps growing, wait times increase, and fidelities decrease to a minimum. To combat this, we consider LIFO, which prioritizes younger requests.

B. *Infinite Buffer* LIFO

This is very similar to the previous model, except that the derivation of the wait time distribution is different. We consider a system where requests are served in a LIFO order, i.e., the buffer is a stack with infinite capacity. The busy period of a queue is defined as the time measured between the instant a request arrives to an empty buffer and the next time the buffer is empty. For an M/M/1 queue, the distribution of the busy period is given by

$$f_B(t) = \frac{1}{t\sqrt{\rho}}e^{-(\lambda_r+\lambda_e)t}I_1(2t\sqrt{\lambda_r\lambda_e}), \quad t \geq 0 \tag{17}$$

where $\rho = \lambda_r/\lambda_e$ is the load. Since LIFO always places a new request at the front of the buffer the wait time distribution is the same as the busy period distribution [18], i.e.,

$$f_W(t) = f_B(t), \quad t \geq 0. \tag{18}$$

With the inverse of the function of fidelity with respect to time

$$g^{-1}(f) = \Gamma^{-1}(\ln(2|\alpha|^2|\beta|^2 - \ln(f - |\alpha|^4 - |\beta|^4)) \tag{19}$$

we can now transform the wait time pdf into the the fidelity pdf

$$f_F(x) = f_B(g^{-1}(x))|\frac{\Gamma^{-1}}{(|\alpha|^4 + |\beta|^4 - x)}|, \quad x \geq 0, \tag{20}$$

We also know the Laplace transform of the busy period and by extension of the wait time is:

$$W^*(s) = \frac{1}{2\lambda_r}\left(\lambda_r + \lambda_e + s - \sqrt{(\lambda_r + \lambda_e + s)^2 - 4\lambda_r\lambda_e}\right). \tag{21}$$

C. FIFOPO

In this section, we consider a system in which incoming requests are stored in a queue with a maximum buffer capacity $B \leq \infty$. When a request arrives to find the buffer full, the oldest request is discarded, and the incoming request is stored in the queue. This makes the probability of service for a requesting FIFOPO dependent on its position in the queue (k) and the number of qubits behind it (j) as it might get pushed out. Therefore, we need to define a new probability, $W_r(j,k,t)$, which is the probability that a request in position k with j requests behind it is served and its remaining wait time is t.

We know from [17] that the Laplace transform $W^*(j,k,s) = \int_0^\infty e^{-st}W(j,k,t)dt$ can be described by

the following set of recursive equations where $j, k \in [0, B]$ and $\text{Re}(s) \geq 0$:

$$W^*(j,0,s) = 1 \quad (\forall j, s),$$
$$W^*(B-1,1,s) = \frac{\lambda_e}{\lambda_e + \lambda_r + s},$$
$$W^*(j,k,s) = \frac{\lambda_r}{\lambda_r + \lambda_e + s} W^*(j+1,k,s)$$
$$+ \frac{\lambda_e}{\lambda_r + \lambda_e + s} W_r^*(j,k-1,s), \quad k > 0, j < N-k,$$
$$W^*(B-k,k,s) = \frac{\lambda_r}{\lambda_r + \lambda_e + s} W^*(B-k+1,k-1,s)$$
$$+ \frac{\lambda_e}{\lambda_r + \lambda_e + s} W^*(B-k,k-1,s) \quad k > 1.$$

Since this is in terms of the joint probability of service and wait time. We need to turn this into a conditional probability of waiting time given the request will be served, which we can get by normalizing with the probability of a random request getting service P_s. Therefore, the Laplace transform is

$$W^*(s) = \mathbb{E}[e^{-sW}] = \frac{\left[\sum_{j=1}^{B_r} W^*(j,k,s)\right] + W^*(N,0,s)}{P_s}.$$

This can be used to calculate $\mathbb{E}[F]$. Next, we consider LIFOPO.

D. LIFOPO

The main difference between this section and the previous one is that incoming request qubits are stored in a stack instead of a queue. We still discard the oldest qubit when a request arrives and the stack is full. Unfortunately, it is difficult to work directly with the wait time pdf. Instead, we work with the Laplace Transform. Unlike the previous model, the wait time only depends upon its position in the queue k. Let $W(k,t)$ denote the probability density that a request in buffer position k gets served eventually, and its wait time will be t. Assuming $k = 1$ corresponds to the head of the stack and $k \in [0, B+1]$ and $\text{Re}(s) > 0$, from classical results [17],

$$W^*(0,s) = 1, W^*(B_i+1,s) = 0, \quad \text{Re}(s) \geq 0$$
$$W^*(k,s) = \frac{\lambda_i}{\lambda_e + \lambda_r + s} W^*(k+1,s)$$
$$+ \frac{\lambda_{i'}}{\lambda_e + \lambda_r + s} W^*(k-1,s) \quad 1 < k < B,$$
$$W^*(B,s) = \frac{\lambda_{i'}}{\lambda_e + \lambda_r + s} W(B-1,s),$$
$$W^*(1,s) = \frac{\lambda_i}{\lambda_e + \lambda_r + s} W^*(2,s)$$
$$+ \frac{\lambda_{i'}}{\lambda_e + \lambda_r + s} W^*(B-1,s).$$

They can be solved to produce

$$W^*(k,s) = \frac{r_1(s)^k r_2(s)^B - r_2(s)^k r_1(s)^B}{r_2(s)^B - r_2(s)^k} \tag{22}$$

where

$$r_{1,2}(s) = \frac{(\lambda_e + \lambda_r + s) \pm \sqrt{(\lambda_e + \lambda_r + s)^2 - 4\lambda_e \lambda_r}}{2\lambda_i}.$$

We need to normalize this Laplace transform as in the previous section with P_s, also, since a new request is always placed in the first position, we have

$$W^*(s) = \mathbb{E}[e^{-sW}] = \frac{W^*(1,s)}{P_s}, \quad \text{Re}(s) \geq 0.$$

V. Optimality of LIFOPO

We have analyzed several memory management and service disciplines. Naturally, this raises the question as to which performs best. In this section, we answer this question by establishing that, out of a large class of work conserving disciplines, LIFOPO is optimal in that it maximizes the final average fidelity of a teleported qubit. This result should not come as a surprise as it is well known that, out of the class of work conserving non-preemptive policies Π', LIFO maximizes $\mathbb{E}[f(W^\pi)]$ for any convex function f where W^π is the sojourn time under policy $\pi \in \Pi'$ for an infinite buffer G/G/1 queue [15], [16].

Let π denote a policy that assigns requests to EPR pairs and determines what teleportation qubits and EPR pairs to discard from the respective buffers to avoid overflows. We first observe that there is no benefit to removing a qubit from a buffer before it is full; hence we only consider polices that remove qubits at the time overflow occurs. Second, we restrict ourselves to work conserving policies; those that always teleport qubits whenever possible. Let Π denote the set of such double buffer policies. We introduce LIFOPO, which always assigns the youngest qubit to be teleported to a newly created EPR pair or the youngest EPR pair to a newly made teleportation request, and always discards the oldest qubit from the buffer when it is about to overflow. A formal definition of this policy is given in the Appendix. Henceforth we refer to LIFO-PO as γ.

***Theorem** 1:* Out of the class of policies Π LIFOPO maximizes average fidelity,

$$\mathbb{E}[F^\pi] \leq \mathbb{E}[F^{\text{LIFOPO}}].$$

where F_π is the teleportation fidelity under π.

Proof sketch. A complete proof is found in the appendix. Here we provide a sketch of the proof. The system can be decomposed into two single buffer subsystems, one for teleportation requests and the other for EPR pairs. Let F_{e}^π and F_{r}^π denote the fidelity for EPR pairs and teleportation requests respectively. We show that $\mathbb{E}[F_{\text{e}}^\pi]$ and $\mathbb{E}[F_{\text{r}}^\pi]$ are maximized when $\pi =$ LIFO-PO. As $\mathbb{E}[F^\pi]$ is a weighted average of $\mathbb{E}[F_{\text{e}}^\pi]$ and $\mathbb{E}[F_{\text{r}}^\pi]$, this establishes the theorem.

Focusing on the request buffer, we condition on the first n departures of qubits from the request buffer, either due to successful teleportation or removal due to overflow. Let $w^\pi = (w_1^\pi, \ldots, w_n^\pi)$ denote the wait times of these requests. Because request qubits can be removed from the buffer without service, we will assign wait times of infinity to those requests. Let m denote the number of these removed qubits. Our proof that γ is optimal is based on establishing the following majorization result between w^π an d w^γ, $\pi \in \Pi$, $\pi \neq \gamma$, $w^\pi \prec^w w^\gamma$. Here $\prec^w$ is defined as follows.

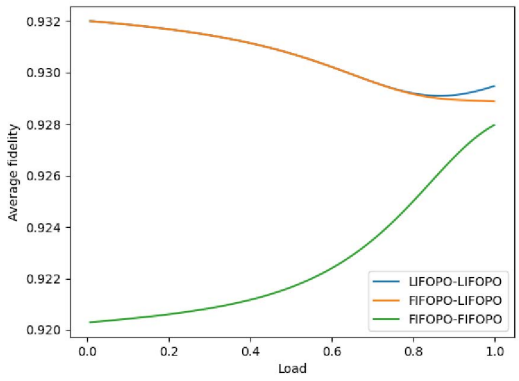

Fig. 3: Plot for average fidelity vs. load. $B = 10$, $\Gamma = 0.01$, $\lambda_e = 5$ and $\lambda_r \in (0, 5)$, therefore, load $\in [0, 1]$.

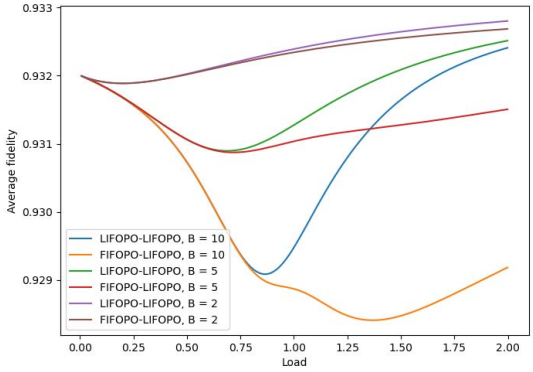

Fig. 4: Plot for average fidelity vs. load comparing different buffer sizes. $\Gamma = 0.01$, $\lambda_e = 5$ and $\lambda_r \in (0, 10)$, therefore, load $\in [0, 2]$.

Definition 1: Let $x, y \in \mathbb{R}_+^{n-m} \times \{\infty\}^m$; y weakly supermajorizes x written $x \prec^w y$ iff

$$\sum_{i=1}^{k} x_{(i)} \geq \sum_{i=1}^{k} y_{(i)}, \quad k = 1, \ldots, n-m.$$

where $x_{(i)}$ (resp. $y_{(i)}$) correspond to the components of x (y) in increasing order.

This is useful in our context because of the following property of $\prec^w$,

$$\sum_{i=1}^{n-m} \phi(x_{(i)}) \leq \sum_{i=1}^{n-m} \phi(y_{(i)}) \tag{23}$$

for any continuous decreasing convex function ϕ.

The proof that $w^\pi \prec^w w^\gamma$ is straightforward and consists of transforming π into γ by taking each non-LIFOPO decision and replacing it with an LIFOPO decision such that the weak majorization is propagated until the resulting policy is LIFOPO. Property (23) can now be applied with $\phi() = F()$ where $F()$ is given in (9), (10), n allowed to go to infinity, and the conditioning on arrival and departure times removed yielding $\mathbb{E}[F(W_r^\pi)] \leq \mathbb{E}[F(W_r^{\text{LIFOPO}})]$. The EPR buffer is handled in a similar manner.

VI. Results

It was proven in Section V that LIFOPO maximizes average fidelity. We can visualize this in Figure 3. We plot the average fidelity of a teleported request with respect to load for different service disciplines. The models are named in the format: X-Y, where X represents the buffer system used by memory for the qubit being teleported and Y represents the buffer system used by the EPR qubits, for example, FIFOPO-LIFOPO denotes a system where the incoming qubits are stored in a FIFOPO buffer, and the EPR pairs are stored in a LIFOPO buffer. We consider FIFOPO-LIFOPO, LIFOPO-LIFOPO and FIFOPO-FIFOPO. Another thing to note is that we only consider models where the buffer sizes for both the

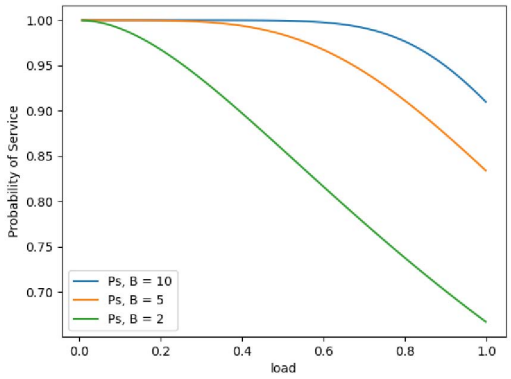

Fig. 5: Plot for Probability a random request reciever service vs. load for a LIFOPO-LIFOPO queue.

EPR buffer and the request buffer are equal, i.e. $B_e = B_r = B$. The dephasing rate $\Gamma = 0.01$, the EPR generation rate is $\lambda_e = 5$, the teleportation request rate λ_r between zero and ten. To reiterate, We observe that LIFOPO-LIFOPO outperforms FIFOPO-LIFOPO and FIFOPO-FIFOPO. The reason for the lower performance of FIFOPO-FIFOPO at low loads is explained by the fact that the EPR pairs are being queued up waiting for requests, but since they have to be used in order of creation, the requests are served by stale EPR pairs rather than fresh ones as is the case of systems that use LIFOPO for the EPR pair. Of course, LIFOPO-LIFOPO performing the best is consistent with Theorem 1. Another thing to note is the increasing nature of FIFOPO-FIFOPO. Since we have a pushout mechanism for the oldest qubit in the queue, increasing the load means a greater chance of older requests being discarded. On the other hand, in the case of LIFOPO-LIFOPO, we observe that as the load approaches one, the fidelity stops decreasing and starts increasing. This is because

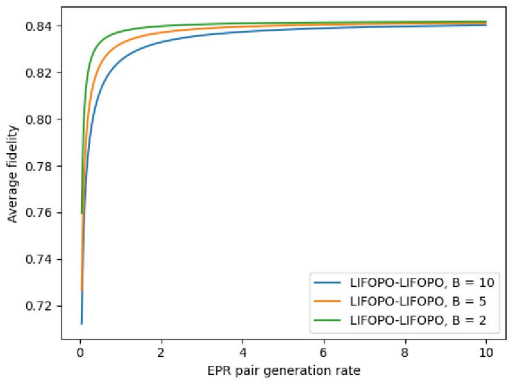

Fig. 6: Plot for average fidelity loss vs. load for a single repeater chain for different buffer sizes following LIFOPO-LIFOPO. $\Gamma = 0.01$.

EPR pairs are served more quickly and have incurred less decoherence. In Figure 4 we explore the performance of LIFOPO-LIFOPO and LIFOPO-FIFOPO as a function of load for three different buffer sizes of 2, 5, and 10. Here we allow the load to vary from zero to two (λ_r varies from zero to 10). In all cases, average fidelity first decreases and then increases. In the case of buffer size of 10, the minimum occurs close to a load of one. This behavior should not come as a surprise as the time qubits spend in either buffer is the same at load one. The minimum average fidelity does not occur at load one because request qubits and EPR pairs decohere at different rates, and the asymmetry becomes more pronounced as buffer size decreases. Last average fidelity decreases with buffer size because increasing the size allows qubits more time to decohere before use. Figure 5 examines the behavior of probability of service as a function of load for the three different LIFOPO-LIFOPO buffer sizes.

A. Application to a repeater node

In this section, we apply the double queue model to a quantum repeater between two nodes, A and B, as shown in Figure 7. This repeater constantly generates EPR pairs between A and B by generating EPR pairs between itself and A, and itself and B, and then performing entanglement swaps. Entanglement swapping is a form of teleportation of one qubit of an EPR pair 'a' using EPR pair 'b'. This "swaps" the entanglement as now one of the qubits of the resource EPR pair 'b' has been entangled with the non-teleported qubit of EPR pair 'a', and the original entanglements have been destroyed. This repeater contains two buffers, one to store EPR pairs between the repeater and node A and the other between the repeater and node B.

Our teleportation model requires the following modification: request qubits suffer the same type of decoherence as EPR pairs. Since we consider the initial fidelity of the Werner state created to be 0.9, the fidelity of the final EPR pair generated between A and B is equivalent to the fidelity loss suffered by the EPR pair that has waited in memory multiplied by 0.9 as the fidelity of the final Werner state created by the entanglement swapping of two Werner states is the product of the fidelities of the two initial Werner states [20]. The fidelity of a Bell pair dephasing with time is given by (5) which is the fidelity function we use for both queues.

As modeled before, EPR pairs are generated according to a Poisson process, and the time between consecutive EPR generations between the repeater and node x is sampled from an exponential distribution with mean μ_x, $x = A, B$. Assuming the router is equidistant between node A and node B and uses similar technologies for generating the EPR pair, we set $\mu_A = \mu_B$. To keep it consistent, we keep the decoherence rate, $\Gamma = 0.01$. We plot the average infidelity defined as $\mathbb{E}[F]$ with respect to μ in Figure 6. Average fidelity increases with the increasing rate because when the rate is low, one queue receives a pair, but since the other queue has a low rate, the arrived pair has to wait a long time before it has a counterpart for service. We also see larger infidelity in larger buffers, but this comes at the cost of a greater chance of rejection, as observed in Figure 5.

VII. Summary

In this paper, we have modeled and quantified the effects of decoherence in a teleportation process. We model memory platforms in networks as queues and utilize queuing theory to calculate how much time a request has to wait for teleportation. We then map these waiting times to fidelity loss due to dephasing. This allows us to derive efficiently computable functions for the average fidelity of the qubits teleported by a node. We consider a case where there are two queues to model caching of EPR pairs and provide a framework to extend results from classical queuing theory on single buffer queues to the double buffer systems. We quantify how service disciplines affect teleportation fidelities in NISQ era devices and calculate average fidelities for different disciplines. We prove the optimality of LIFOPO-LIFOPO for serving teleportation requests and compare it to other disciplines. We analyze the effects of buffer sizes and compare their Service probabilities. Lastly, we apply this framework to analyze the average transportation fidelity of a quantum repeater between two nodes and see how different buffer sizes compare in terms of fidelity and service completion probability.

A. Future Work

There are many open questions and directions this work can take. A most natural extension is to account for mixed states as requests. One can achieve this by modifying (7) and using the fidelity formula for comparing two mixed states instead of assuming a state is pure. Another direction would be to use more accurate distribution models for the EPR pair generation as in [9] and apply this model to longer repeater chains. Another natural extension would be to model a constant timeout policy so that if a request has been in the queue for longer than some time C, we can guarantee a minimum fidelity for the teleported information. As stated in II-A, we do not account for erasure or loss errors in the

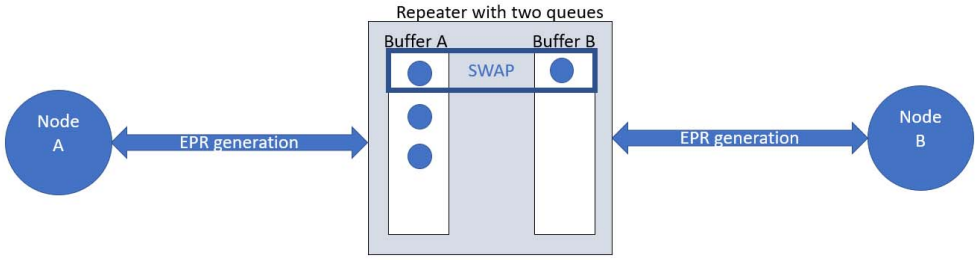

Fig. 7: A single repeater between two nodes, It has two buffers to store EPR pairs and only one can non empty at the same time. If the repeater has a qubit from an EPR pair in buffer A, and an EPR pair is generated between it and B, it performs a swap and discards the qubits.

memory. This can be rectified by considering queues with impatient customers. As long as the errors can be modeled as exponential equations, the CTMC formulation can be applied to account for them.

REFERENCES

[1] S. Wehner, D. Elkouss, and Ronald Hanson, "Quantum internet: A vision for the road ahead," Science, vol. 362, no. 6412, p. eaam9288, 2018, doi: 10.1126/science.aam9288.

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

[3] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, "Fundamental limits of repeaterless quantum communications," Nature Communications, vol. 8, no. 1, p. 15043, Apr. 2017, doi: 10.1038/ncomms15043.

[4] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels," Phys. Rev. Lett., vol. 70, no. 13, pp. 1895–1899, Mar. 1993, doi: 10.1103/PhysRevLett.70.1895.

[5] J. B. Brask, I. Rigas, E. S. Polzik, U. L. Andersen, and A. S. Sørensen, "Hybrid Long-Distance Entanglement Distribution Protocol," Phys. Rev. Lett., vol. 105, no. 16, p. 160501, Oct. 2010, doi: 10.1103/PhysRevLett.105.160501.

[6] W. Dai, A. Rinaldi, and D. Towsley, Entanglement Swapping in Quantum Switches: Protocol Design and Stability Analysis. arXiv, 2021. doi: 10.48550/ARXIV.2110.04116.

[7] J. Preskill, "Quantum Computing in the NISQ era and beyond," Quantum, vol. 2, p. 79, Aug. 2018, doi: 10.22331/q-2018-08-06-79.

[8] G. Vardoyan, S. Guha, P. Nain and D. Towsley, "On the Stochastic Analysis of a Quantum Entanglement Distribution Switch," in IEEE Transactions on Quantum Engineering, vol. 2, pp. 1-16, 2021, Art no. 4101016, doi: 10.1109/TQE.2021.3058058.

[9] S. Brand, T. Coopmans, and D. Elkouss, "Efficient computation of the waiting time and fidelity in quantum repeater chains," IEEE j. sel. areas commun., vol. 38, no. 3, pp. 619–639, 2020.

[10] G. Vardoyan, M. Skrzypczyk, and S. Wehner, "On the quantum performance evaluation of two distributed quantum architectures," Performance Evaluation, vol. 153, p. 102242, Feb. 2022, doi: 10.1016/j.peva.2021.102242.

[11] W. Dai, T. Peng and M. Z. Win, "Quantum Queuing Delay," in IEEE Journal on Selected Areas in Communications, vol. 38, no. 3, pp. 605-618, March 2020, doi: 10.1109/JSAC.2020.2969000.

[12] Y. Wang et al., "Efficient quantum memory for single-photon polarization qubits," Nature Photonics, vol. 13, no. 5, pp. 346–351, May 2019, doi: 10.1038/s41566-019-0368-8.

[13] W. J. Munro, K. Azuma, K. Tamaki and K. Nemoto, "Inside Quantum Repeaters," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 21, no. 3, pp. 78-90, May-June 2015, Art no. 6400813, doi: 10.1109/JSTQE.2015.2392076.

[14] S. Albeverio, S.-M. Fei, and W.-L. Yang, "Optimal teleportation based on bell measurements," Physical Review A, vol. 66, no. 1, Jul. 2002, doi: 10.1103/physreva.66.012301

[15] J.G. Shanthikumar, U. Sumita. "Convex ordering of sojourn times in single-server queues: Extremal properties of FIFO and LIFO service disciplines," *J. Appl. Prob.*,**24**, 737-748, 1987.

[16] Z. Liu, P. Nain, D. Towsley, "Sample path methods in the control of queues", *Queueing Systems – Theory and Applications*, **21**(3-4), 293-335, September 1995.

[17] B. Doshi and H. Heffes, "Overload Performance of Several Processor Queueing Disciplines for the M/M/1 Queue," in IEEE Transactions on Communications, vol. 34, no. 6, pp. 538-546, June 1986, doi: 10.1109/TCOM.1986.1096578.

[18] M. Zukerman, Introduction to Queueing Theory and Stochastic Teletraffic Models. arXiv, 2013. doi: 10.48550/ARXIV.1307.2968.

[19] A.W. Marshall, I. Olkin, B.C. Arnold. *Inequalities: Theory of Majorization and its Applications*, Springer, 2011.

[20] A. Sen(De), U. Sen, Č aslav Brukner, V. Bužek, and M. Żukowski, "Entanglement swapping of noisy states: A kind of superadditivity in nonclassicality," Physical Review A, vol. 72, no. 4, Oct. 2005, doi: 10.1103/physreva.72.042310.

APPENDIX

Proof of Theorem 1. We focus separately on the two buffers and focus on the amount of time qubits to be teleported and EPR pairs are allowed to decohere waiting to be matched up with each other. Henceforth we focus on the request buffer and only on requests that arrive when no EPR pair is stored in the EPR buffer. We focus on the arrivals and departures of n requests under policy $\pi \in \Pi$. Let $a_1, \ldots, a_n$ and $d_1, \ldots, d_n$ denote the arrival and departure times for these requests.

Here a departure corresponds either to a pairing with a newly creation of an EPR pair followed by a successful teleportation or removal from the buffer. Let $m \leq n$ denote the number of qubits removed from the buffer. A policy $\pi \in \Pi_{\mathsf{LIFO-O}}$ satisfies the following properties:

- There exists no pair of requests j, k that are served such that $a_k < a_j < d_k < d_j$,
- there exists no pair of requests j, k where k is served and j is discarded such that $a_k < a_j < d_k$,
- there exists no pair of requests j, K that are discarded such that $a_k < a_j < d_j < d_k$

Let $w^\pi = (w_1^\pi, \ldots, w_n^\pi)$ denote the wait times of these teleportation requests. Because requests can be removed from the buffer without service, we will assign wait times of infinity to those requests. Our proof that γ is optimal is based on showing $w^\pi \prec^w w^\gamma$, $\pi \in \Pi$ where $\prec^w$ is defined in Section V. Note that the standard definition [19] corresponds to the

case $m = 0$. We introduce an operator T_{ij}, called the "T-transform", as follows. Let $x \in \mathbb{R}^n_+$;

$$T_{ij} = \lambda I + (1-\lambda)Q_{ij}$$

where I is the identity operator, Q_{ij} is an operator that permutes the i-th and j-th components of x and $0 \le \lambda \le 1$. In other words,

$$T_{ij}x = (x_1, \ldots, x_{i-1}, \lambda x_i + (1-\lambda)x_j, x_{i+1}, \ldots, x_{j-1}, (1-\lambda)x_i + \lambda x_j, x_{j+1}, \ldots, x_n)$$

It is easily shown that $T_{ij}x \prec^w x$ provided $x_i, x_j < \infty$. Note that $x \prec^w Q_{ij}x$ ($\lambda = 0$). Last, define the function $S_j(x)$ as

$$S_j(x) = (x_1, \ldots, x_{j-1}, \alpha x_j, x_{j+1}, \ldots, x_n)$$

with $0 \le \alpha \le 1$. Then $x \prec^w S_j(x)$.

Consider the system with n requests arriving at times $a_1, \ldots, a_n$ and depart at times $d_1, \ldots, d_n$.

We transform π to γ through a sequence of steps that creates a sequence of policies $\pi_0 = \pi, \pi_1, \pi_2, \ldots, \pi_h = \gamma \in \Pi_{\mathsf{LIFO-O}}$ such that $w^{\pi_l} \prec^w w^{\pi_{l+1}}$, $l = 0, \ldots h-1$.

Assume π_l violates the LIFO-O property. There are three cases depending on whether the two requests are served, one is served and the other removed or both removed.

1) **Both are served.** Request k is served before a younger request j, $a_k < a_j < d_k < d_j$ (we omit dependence on π_l). We construct π_{l+1} from π_l by switching the order in which j and k are served. The wait times for requests j and k under π_l are $w_j^{\pi_l} = d_j - a_j$ and $w_k^{\pi_l} = d_k - a_k$ and under π_{l+1} are $w_j^{\pi_{l+1}} = d_k - a_j$ and $w_k^{\pi_{l+1}} = d_j - a_k$. Here w^{π_l} and $w^{\pi_{l+1}}$ satisfy

$$w^{\pi_l} = T_{jk} w^{\pi_{l+1}}$$

with

$$\lambda = \frac{a_j - a_k}{(a_j - a_k) + (d_j - d_k)}.$$

Hence $w^{\pi_l} \prec^w w^{\pi_{l+1}}$. See Figure 8.

2) **One request is served.** Request k is served while a younger request is discarded, $a_k < a_j < d_k$. We switch the order in which these two requests are handled resulting in the servicing of j at time d_k and removal of k at time d_j. Then w^{π_l} and $w^{\pi_{l+1}}$ satisfy

$$w^{\pi_{l+1}} = S(Q_{jk} w^{\pi_l})$$

with $\alpha = (d_k - a_j)/(d_k - a_k)$. Hence $w^{\pi_l} \prec^w w^{\pi_{l+1}}$. See Figure 9.

3) **Both are removed.** A younger request j is removed from the buffer before an older job k under π_l, $a_k < a_j < d_j < d_k$. We switch the order of the removals under π_{l+1}. This does not affect wait times and $w^{\pi_l} \prec^w w^{\pi_{l+1}}$. See Figure 10.

This procedure is repeated until the LIFO-O properties are satisfied and, consequently $w^\pi \prec^w w^\gamma$.

We fixed the arrival and service times. Remove the conditioning on them and let $W^\pi(n)$ denote the wait time of a randomly chosen request from the first n requests that are served. From the above majorization result and the equivalence (23), we conclude that $\mathbb{E}[\phi(W^{\mathsf{LIFOPO}}(n)] \ge \mathbb{E}[\phi(W^\pi(n))]$ for every convex decreasing function ϕ. Moreover if the limits $W^{\mathsf{LIFOPO}} = \lim_{n\to\infty} W^{\mathsf{LIFOPO}}(n)$ and $W^\pi = \lim_{n\to\infty} W^\pi(n)$ exist, then $\mathbb{E}[\phi(W^{\mathsf{LIFOPO}}] \ge \mathbb{E}[W^\pi]$.

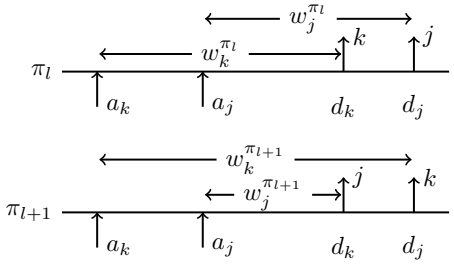

Fig. 8: Case 1.

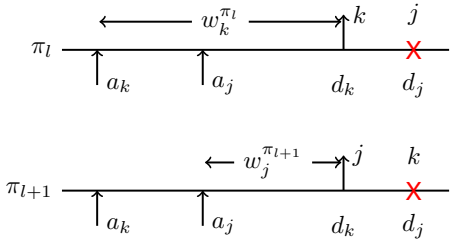

Fig. 9: Case 2.

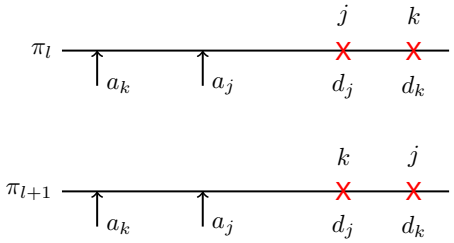

Fig. 10: Case 3.

Returning to our teleportation system, under the assumption that requests and EPR pairs are generated according to Poisson processes, when placed in their respective buffers, they will exhibit stationary wait time W_r^π and W_e^π respectively. The respective qubits decohere at different rates $F_r(t)$ and $F_e(t)$ in the two memories according to (9), (10), As these decoherence functions are decreasing and convex, we conclude that there exists a LIFOPO $\in \Pi_{\mathsf{LIFO-O}}$ such that $\mathbb{E}[F_r(W^{\mathsf{LIFOPO}}(n)] \ge \mathbb{E}[F_r(W^\pi(n))]$ and $\mathbb{E}[F_e(W^{\mathsf{LIFOPO}}(n)] \ge \mathbb{E}[F_e(W^\pi(n))]$. The expected teleportation fidelity for the entire system, $\mathbb{E}[F^\pi]$ is $\mathbb{E}[F^\pi] = q\mathbb{E}[F^\pi] + (1-q)\mathbb{E}[F^\pi]$ where q is the probability that a request qubit arrives to a system where no EPR qubits are available. Finally, we conclude $\mathbb{E}[F^{\mathsf{LIFOPO}}] \ge \mathbb{E}[F^\pi]$ for all $\pi \in \Pi$.