On the Stochastic Analysis of a Quantum Entanglement Distribution Switch

Gayane Vardoyan, Saikat Guha, Philippe Nain, Don Towsley

On the Stochastic Analysis of a Quantum Entanglement Distribution Switch

Gayane Vardoyan[1], Saikat Guha[1], Philippe Nain[1], and Don Towsley[1]

[1]University of Massachusetts, Amherst, gvardoyan@cs.umass.edu
[2]The University of Arizona,saikat@optics.arizona.edu
[3]Inria, France, email: philippe.nain@inria.fr
[4]University of Massachusetts, Amherst, towsley@cs.umass.edu

March 20, 2021

Abstract

We study a quantum entanglement distribution switch that serves k users in a star topology. We model variants of the system as continuous-time Markov chains (CTMCs) and obtain expressions for switch capacity, expected number of qubits stored in memory at the switch, and the quantum memory occupancy distribution. We obtain a number of analytic results for systems in which measurements are imperfect, the links are homogeneous or heterogeneous and for switches that have an infinite or finite number of quantum memories, or buffers. In addition, we model the effect of decoherence of quantum states and associated cut-off times on their storage using a simple model. From numerical observations, we discover that decoherence-associated cut-off times have little effect on capacity and expected number of stored qubits for homogeneous systems. For heterogeneous systems, especially those operating near the boundaries of their stability regions (*i.e.*, systems that are nearly unstable), buffer size and decoherence can have significant effects on performance metrics. We also learn that in general, increasing the buffer size from one to two qubits per link is advantageous to most systems, while increasing the buffer size further yields diminishing returns. The analytical results obtained in this work can serve as a useful guide toward the future design of quantum switches – *e.g.*, by allowing the designer to determine how many quantum memories suffice for a given number of users – as well as provide valuable insight on the performance of these and similar devices.

1 Introduction

Entanglement is an essential component of quantum computation, information, and communication. Its applications include quantum cryptography (*e.g.*, [1, 2, 3, 4, 5]), distributed quantum computing (*e.g.*, [6, 7]), quantum sensing (*e.g.*, multipartite entanglement for quantum metrology [8, 9] and spectroscopy [10]; quantum machine learning [11]), and it offers advantages to quantum communication (see, *e.g.*, [12, 13]). These applications drive the increasing need for a quantum switching network that can supply end-to-end entanglement to groups of endpoints that request them [14, 15, 16, 17]. To realize such quantum systems, several architectures have been proposed to support high entanglement generation rates, high fidelity, and long coherence times [18, 19, 20, 21, 22, 23].

In this paper, we study in detail the most basic and fundamental component of a quantum network – a single quantum switch that serves k users in a star topology. Each user has a dedicated link connected to the switch. In the most general case, the switch serves n-partite entangled states to sets of users according to incoming requests, where $n \leq k$. To achieve this, link-level entangled states are generated at a constant rate across each link, resulting in two-qubit maximally-entangled states (*i.e.*, Bell pairs or EPR states). These qubits are stored at local quantum memories: one from each Bell pair at the user and the other at the switch. When enough link-level entanglement is accrued (at least n Bell pairs at n different links), the switch performs multi-qubit measurements to provide end-to-end entanglement to user groups of size n. When $n = 2$, the switch uses Bell-state measurements (BSMs) and when $n \geq 3$, it uses n-qubit Greenberger-Horne-Zeilinger (GHZ) basis measurements [24]. In this work, we focus on the case of $n = 2$ – *i.e.*, the case of bipartite-only switching, although some prior work on $n \geq 3$, as well as n being allowed to switch between 2 *or* 3 will be discussed in Section 2.

The objective of this work is to characterize the performance of such a device, for example by determining its capacity (defined as the maximum achievable rate of entanglement switching), and deriving expressions for the expected quantum memory occupancy under various assumptions – *e.g.*, while assuming a particular quantum memory coherence time or limitations on the available number of memories. We accomplish this objective by constructing a simple, yet descriptive model of a quantum switch: we determine a small number of important model parameters and abstract away the specifics of implementation and physical platform. For instance, we do not focus on a specific method of entanglement generation on a link, and we do not analyze a specific quantum memory implementation; rather, we include the rate of entanglement generation and memory coherence times as configurable system parameters. This way, our model is agnostic to hardware architecture and protocol specifics, and is kept general. Subsequently, when we analyze the model, we obtain results that are often interpretable and intuitive.

We consider systems in which links may generate entanglement at different rates and where the switch can store one or more qubits (each entangled with another qubit held by a user) per link. Throughout this paper, we refer to these pairs of stored qubits as *stored entanglements*. Another factor that impacts performance is decoherence of quantum states and subsequent qubit storage cut-off times that may be imposed by the switch or an application to prevent the consumption of low-fidelity states; we model this and study its effect. The main metric of interest for this network is its capacity C, *i.e.*, the maximum possible number of end-to-end entanglements served by the switch per time unit. Another metric of interest is the expected number of qubits Q in memory at the switch, $E[Q]$. Where possible, we also derive in closed-form the distribution of the number of stored qubits at the switch. Both C and $E[Q]$ depend on the values of k, n, entanglement generation and decoherence rates, number of quantum memories (often referred to as *buffer size* throughout this manuscript), and the switching mechanism, including the scheduling policy used by the switch.

Contributions of this work are as follows: by modeling the switch as a continuous-time Markov chain (CTMC), we derive C and $E[Q]$ for $n = 2$ for a first-in, first-out scheduling policy on successfully-generated entanglement, and study how they vary as functions of k, buffer size, and decoherence rate or qubit storage cut-off times. From our analysis, we gain valuable insight into which factors influence capacity the most, and which ones are of lesser consequence. For instance, we find that when $n = 2$ and links are identical, the number of links and their entanglement generation rate are the most impactful, while decoherence-associated cut-off times for qubit storage and link buffer size have little effect on capacity and $E[Q]$. However, the same is not true in the case of non-identical links, where the distribution of entanglement generation rates, combined with finite coherence time, can drastically affect both C and $E[Q]$. Last, we compare our results for the $n = 2$, identical-link, negligible decoherence, infinite buffer case against a logically more accurate discrete-time Markov chain (DTMC) model studied in [25] and find that the differences in predictions of the performance metrics are small.

The remainder of this paper is organized as follows: in Section 2, we discuss relevant background and related work. In Section 3, we cover modeling techniques, assumptions, and objectives. In Section 4, we introduce our CTMC models for $n = 2$ and present their analyses. Numerical observations are discussed in Section 5. In Section 6, we discuss ways in which some of our modeling assumptions may be relaxed. We conclude in Section 7.

2 Background

In [22], Herbauts *et al.* implement an entanglement distribution network intended for quantum communication applications. The fidelities of entanglement generated in this network were 93% post-distribution, and fidelities of 99% were shown to be achievable. The demonstration entails distributing bipartite entanglement to any pair of users wishing to share entanglement in a multi-user network (there were eight users in the experimental setup). Delivering multiple bipartite entangled states was shown to be possible virtually simultaneously. The authors specifically cite a possible application of the network in a scenario where a single central switch dynamically allocates two-party entanglement to any pair of users in a static network. In this paper, we study variants of this system, but here we additionally assume that the switch has the ability to store entangled qubits for future use.

In this work, we do not make an assumption about the fidelity of successfully-generated entanglement – neither at the link nor at the end-to-end levels – and focus only on a switching policy that maximizes the entanglement switching rate. While this is a good starting point for quantum switch analysis, being able to make quantitative statements about the fidelity of entanglement is another important question. The analysis of such a study, which will likely have to incorporate some form of entanglement purification, *e.g.*, [26], is left as an open question and a subject of future work, although we add some discussion on how this may be accomplished in Section 6. Since the original introduction of our quantum switch model in [27], Coopmans *et al.* studied the effect of memory coherence time on the average fidelity of the end-to-end entanglement served to the users by the switch, using NetSquid, a discrete-event simulation framework for quantum networks [28]. In their work, decoherence was modeled as exponential T_2 noise, and the simulated switch did not implement the qubit storage cut-off policy we consider here; but the authors were nevertheless able to use our decoherence-free models and validate our theoretical findings for the switch capacity as a function of buffer size, which are in close agreement with the simulation.

In recent years, there have been other promising experimental demonstrations for realizing the fundamental components of quantum repeater architectures. For instance, in [18], Bhaskar *et al.* implement quantum-memory-enhanced quantum communication to overcome the fundamental limit of repeaterless communication [29]. At the same time, new architectures and protocols, which promise to yield higher-fidelity states and quicker end-to-end entanglement generation rates, have been proposed – *e.g.*, the quantum router proposal in [19] achieves both of these objectives. Such advances further emphasize the importance of analysis and theoretical studies to help guide hardware specifications and protocol design for quantum communication architectures.

In [30], the authors use Markovian models to compute the expected waiting time in quantum repeaters with probabilistic entanglement swapping. Specifically, they consider entanglement distribution over a distance subdivided by repeater segments, and while they propose a method of computing the average waiting time for an arbitrary number of links, explicit expressions are provided for only up to four segments. In contrast, we consider a single quantum repeater-like device, but one that services an arbitrary number of links.

In [31], we analyze the capacity region of a quantum entanglement switch that serves users in a star topology and is constrained to store one or two qubits per link. The problem setup is similar to that of this work, with the exception that the switch has the ability to serve bipartite and tripartite end-to-end entanglement. There, we examine a set of randomized switching policies and find policies that perform better than time-division multiplexing between bipartite and tripartite entanglement switching. Note that while in [31], we allow the switch to choose between two types of entanglement to serve at every time step, in this work we fix $n = 2$ and analyze it in more detail: for instance, in [31] all links are assumed to be identical, while in this work links may be heterogeneous and buffer sizes can be larger than one or two per link.

In [32], we study a quantum switch serving n-partite end-to-end entangled states to $k \geq n$ users and for $n \geq 2$. The setup is identical to that of this work, but limited to the case of a homogeneous-link, infinite-buffer system with no quantum state decoherence. For the case of $n = 2$, the results are consistent with those of this paper, and we build on them to explore more complex bipartite switching systems. As new quantum architectures and technologies emerge, we expect quantum networks to become more prevalent and suitable for practical use. With link-level and especially end-to-end entanglement being a valuable commodity in these networks, proper resource management will be imperative for reliable and efficient operation, which further motivates our study.

3 Model and Objectives

Consider first a fairly general setting of the proposed problem: k users are attached to a quantum entanglement distribution switch via k dedicated links. At any given time step, any set of n users (with $n \leq k$) may wish to share an end-to-end entangled state. The creation of an end-to-end entanglement involves two steps. First, users generate pairwise entanglements with the switch, which we call link-level entanglements. Each of these results in a two-qubit entangled (Bell) state, with one qubit stored at the switch and the other stored at a user. Once there are n link-level Bell pairs available to fulfill a request between n users, the process enters step two: the creation of an end-to-end entanglement. The switch chooses the set of n locally-held qubits (that are entangled with n qubits held by the n distinct users) corresponding to the request and performs an entangling measurement. If such a measurement is successful, the result is an n-qubit maximally-entangled state between the corresponding n users. If after this step more link-level entanglements are available and can be used to fulfill another request, the switch repeats the second step until either there are fewer than n local qubits left or until no more requests can be fulfilled.

In this work, our objective is to derive a tight upper bound on the entanglement switching rate when $n = 2$, *i.e.*, the maximum possible rate at which the switch may serve bipartite end-to-end entangled states – we call this quantity the bipartite switching capacity of the system. Since this upper bound should hold for any workload, it is necessary for us to assume that *any* two users wish to share an entangled state; in fact, removing this assumption would necessarily decrease the rate at which the switch is allowed to serve end-to-end entanglement. With this request policy, the switch has no restrictions on which measurements to perform whenever two distinct link-level entanglements are available. Hence, in step two of entanglement distribution, the switch simply chooses a set of two qubits corresponding to Bell pairs on two distinct links, and uses them in the entangling measurement. Step two is repeated until at most one link has available Bell pairs. The results of our analysis on the capacity of the switch can be used as a comparison basis for other types of scenarios, in which, for example, each pair of users may specify a desired rate of communication with each other through the switch. Another utility of this analysis is that by examining a switch that operates at or near maximum capacity, one may gain insight on the practical memory requirements of a switch.

Both link-level entanglement generation and entangling measurements can be modeled as probabilistic phenomena [33]. In this work, we model the former as a Poisson process: each link attempts entanglement generation at rate λ, and for link $l \in \{1, \ldots, k\}$, each attempt succeeds with probability $p_l \approx e^{-\beta L_l}$, where L_l is the length of the lth link (*e.g.*, optical fiber) and β its attenuation coefficient. Hence, link l generates successful entanglements with rate $\mu_l{:=}\lambda p_l$. We refer to the special case of $\mu_l = \mu_m$, $\forall l, m \in \{1, \dots, k\}$ as a *homogeneous* system, and when they are not necessarily equal, as a *heterogeneous* system. We assume that measurements performed by the switch succeed with probability q[1].

In [25], we modeled a quantum switch as a DTMC. The basic setup there is the same as that of this work, but several more simplifying assumptions are made: the links are assumed to be identical, the buffer size infinite, and decoherence is assumed to be negligible. Relaxing any of these assumptions poses several difficulties and complicates the analysis, in some cases making it intractable. In fact, even with the simplifying assumptions of [25], we were only able to obtain a closed-form expression for the switch capacity, but not for $E[Q]$. To gain intuition on why the analyses of the two models are so different, consider a switch with k identical links and no decoherence, and consider a state where a link l has j stored Bell pairs, $j \geq 1$. In the CTMC, a "backward" transition may occur, when another link (other than l) successfully generates entanglement, and a "forward" transition occurs when link l generates another entanglement. This is illustrated in Figure 2. In the DTMC, there are several other transitions that must be considered, since within a given time slot, more than one link may generate a Bell pair successfully, including link l, so that transitions may occur between non-adjacent states. Further, all such combinatorial sets of links must be considered, sometimes yielding rather unwieldy expressions for the transition probabilities. Nevertheless, the DTMC is a logically more accurate way to model such a system; we later numerically compare the differences between the two models.

Next, we describe how the switch handles quantum state decoherence and how we model it. In quantum networking literature, there are several references to a *"cut-off* time" for quantum state storage; see, *e.g.*,[36, 37, 38, 39, 40, 41]. The cut-off time has slightly varying definitions in different contexts: in some cases, it is viewed as a quantum memory lifetime (or coherence time), and determines how long a qubit should be held in memory, as the effects of decoherence on the quantum state are considered too great beyond the cut-off time. In other contexts, it is instead viewed as a configurable parameter that may be determined by a routing or even an application-level protocol in such a way as to ensure that the final fidelity of the end-to-end states is above some required threshold (for instance, some QKD protocols can tolerate fidelities of no less than 0.81 [42]). In such scenarios, the cut-off strategy is used to reduce the effect of decoherence on stored quantum states or to increase the distance over which a secret key can be generated, at the cost of a lower end-to-end entanglement generation or secret-key rate. In summary, the cut-off time is a constant quantity that either corresponds to the platform-dependent quantum memory coherence time, or to some possibly optimized parameter specified by a user or an application, but regardless of the exact definition, it is closely tied to the quantum memory coherence time. In practical implementation proposals

[1] With a linear optical circuit, four unentangled ancilla single photons and photon number resolving detectors, with all the devices being lossless, $q = 25/32 = 0.78$ can be achieved for BSMs [34]. With other technologies q close to 1 can be achieved [35].

of this strategy, an entangled qubit is held in memory for some time $t^\star$, after which it is deterministically discarded.

To model this decoherence-associated cut-off time for qubit storage, we approximate this deterministic discarding procedure by a probabilistic one: the switch discards a qubit after an exponentially-distributed amount of time, with mean $1/\alpha$. In other words, our cut-off time (or, as we sometimes interchangeably use, the coherence time) is in effect an exponential random random variable (r.v.), instead of a constant quantity. We make this modeling choice because it seamlessly extends our decoherence-free model of a quantum switch and also because modeling deterministic components of system operation is a difficult task, and this system is no exception. In summary, while the exponential assumption on qubit discarding is not physically meaningful, it makes the analysis tractable. In Section 5.4 we simulate the probabilistic and deterministic qubit storage cut-off time policies and compare both simulations to the results of our analyses. While an exhaustive evaluation over all parameter values (buffer size, decoherence/cut-off rate, entanglement generation rate, etc.) is not possible, our limited results imply that, at least for realistic and representative use cases, our approximation of the deterministic cut-off policy using a probabilistic one is reasonable.

Next, we discuss the specifics of qubit prioritization for storage and measurements. If at any time there are fewer than $n = 2$ link-level entanglements, the switch may choose to store the available entangled qubits and wait until there are enough new ones generated to create an end-to-end entanglement. We assume that the switch can store $B \geq 1$ qubits in its buffer, per link. If on the other hand, there are more than $n = 2$ link-level entanglements, the switch must decide which set(s) of them to use in measurement(s). Such decisions can be made according to a pre-specified policy: for example, a user or a set of users may be given higher priority for being involved in an end-to-end entanglement. Other scheduling policies may be adaptive, random, or any number of hybrid policies. In this work, we assume that the switch uses the *Oldest Link Entanglement First (OLEF)* rule, wherein the oldest link-level entanglements have priority to be used in entangling measurements. A practical reason for this rule is that quantum states are subject to decoherence, which is a function of time; hence, our goal is to make use of link-level entanglements as soon as possible.[2] When we model systems with a finite number of quantum memories, then there may occur scenarios in which a link has used up all its available memories and must decide whether to discard an older Bell pair in order to store a newly-generated one. In such a case, the OLEF rule still applies, and we discard the qubit associated with the oldest stored entanglement to make space for the qubit from the newly-generated Bell pair. Note that the OLEF switching policy we consider in this work is one that optimizes the entanglement switching rate, but it may not be the optimal policy for other figures of merit, such as average end-to-end fidelity of entanglement. A fidelity-optimal switching policy, especially one that incorporates a purification protocol, is an open question

[2] If the system is operating in discrete time as in [25], there may arise instances in which two or more links are tied for having the oldest entanglements. In such cases, as long as the switch follows the OLEF rule, sets of link-level entanglements are chosen at random for measurements, provided that each set consists of n entanglements belonging to n distinct links.

and requires further analysis.

The state space of the system we have described can be represented by a vector $\boldsymbol{Q}(t) \in \{0, 1, \ldots, B\}^k$, where the lth element corresponds to the number of stored entanglements at link l at time t. One consequence of the assumption that any pair of users always wishes to share an entangled state is that at most one user will store entanglement at any time. Hence, throughout this work, up to one link may have a stored Bell pair after step two of entanglement distribution. Our goal is to derive expressions for system capacity C (*i.e.*, the number of end-to-end entanglements produced per time unit) and the expected number of stored qubits $E[Q]$. Throughout the paper, we use the result that if the balance equations of an irreducible CTMC have a unique and strictly positive solution, then this solution represents the stationary distribution of the chain.

4 Continuous Time Markov Chain for Bipartite Switching

In this section, we introduce and analyze a CTMC model of a bipartite entanglement distribution switch serving k users. We first assume that memories do not decohere and obtain expressions for capacity and the expected number of qubits stored at the switch. We then modify the model to incorporate decoherence and qubit storage cut-off times and analyze it. Last, we derive an upper bound for the capacity of the switch.

4.1 The Heterogeneous Case

Assume μ_l depends on l, *i.e.*, the links are heterogeneous. For subsequent analysis, it is useful to define

$$\gamma := \sum_{l=1}^{k} \mu_l,$$

the aggregate entanglement generation rate over all links. Also, let $\boldsymbol{e}_l$ be a size k vector with all zeros except for the lth component, which is 1, and let $\mathbf{0}$ be a vector of size k with all entries equal to 0.

We are interested in the stationary distribution and stability conditions for a heterogeneous system with infinite and finite buffers. As discussed in Section 3, in bipartite entanglement switching, only one link stores entanglements at a time, but since links generate entanglements at different rates, we must keep track of which link is associated with the stored entanglement(s). Let $\boldsymbol{Q}(t) = (Q_1(t), \ldots, Q_k(t)) \in \{0, 1, 2, \ldots\}^k$ represent the state of the system at time t, where $Q_l(t)$ is the number of entanglements stored at link l, $l \in \{1, \ldots, k\}$, at time t. As a consequence of the scheduling policy described in Section 3, if $Q_i(t) > 0$ for some i, then $Q_j(t) = 0$, $j \neq i$. In other words, $\boldsymbol{Q}(t)$ only takes on values $\mathbf{0}$ or $j\boldsymbol{e}_l$, $l \in \{1, \ldots, k\}$, $j \in \{1, 2, \ldots\}$. Here, $\mathbf{0}$ represents the state where no entanglements are stored, and $j\boldsymbol{e}_l$ represents the state where the lth link has j stored entanglements.

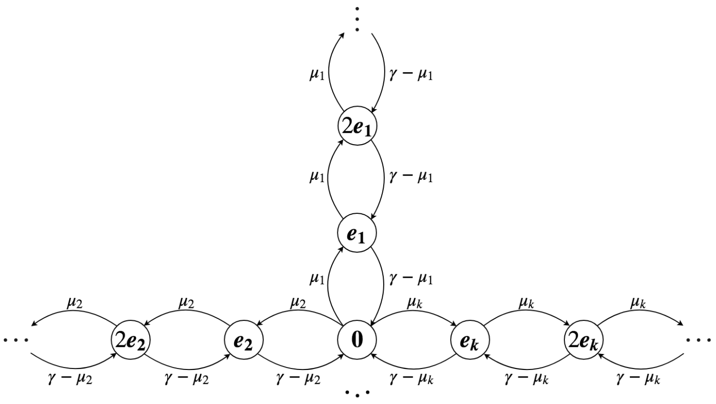


Figure 1: A CTMC for a k-user, infinite buffer, heterogeneous-link switch. μ_l is the entanglement generation rate of link l, while γ is the aggregate entanglement generation rate of all links. $\boldsymbol{e}_l$ is a vector of all zeros except for the lth position, which is equal to one.

Define the following limits when they exist:

$$\pi_0 = \lim_{t\to\infty} P(\boldsymbol{Q}(t) = \mathbf{0}),$$
$$\pi_l^{(j)} = \lim_{t\to\infty} P(\boldsymbol{Q}(t) = j\boldsymbol{e}_l).$$

Once we obtain expressions for π_0 and $\pi_l^{(j)}$, we can derive expressions for capacity and the expected number of stored qubits $E[Q]$.

4.1.1 Infinite Buffer

Figure 1 presents the CTMC for a switch with an infinite buffer. Consider state $\mathbf{0}$ (no stored entanglements). From there, a transition along one of the k "arms" of the CTMC occurs with rate μ_l, when the lth link successfully generates an entanglement. For a BSM to occur, any of the $k-1$ other links must successfully generate an entanglement: this occurs with rate $\gamma - \mu_l$. The balance equations are

$$\pi_0\mu_l = \pi_l^{(1)}(\gamma - \mu_l),\ l \in \{1, \dots, k\},$$
$$\pi_l^{(j-1)}\mu_l = \pi_l^{(j)}(\gamma - \mu_l),\ l \in \{1, \dots, k\},\ j \in \{2, 3, \dots\},$$
$$\pi_0 + \sum_{l=1}^{k}\sum_{j=1}^{\infty} \pi_l^{(j)} = 1.$$

From above, we see that for $j = 1, 2, \dots,$

$$\pi_l^{(j)} = \rho_l^j \pi_0, \quad \text{where} \quad \rho_l \equiv \frac{\mu_l}{\gamma - \mu_l}, \forall\ l.$$

It remains to obtain π_0; we can use the normalizing condition:

$$\pi_0 + \pi_0 \sum_{l=1}^{k} \sum_{j=1}^{\infty} \rho_l^j = \pi_0 \left(1 + \sum_{l=1}^{k} \left(\sum_{j=0}^{\infty} \rho_l^j - 1\right)\right) = 1.$$

Now, assume that for all $l \in \{1, \ldots, k\}$, $\rho_l < 1$. This implies that for all l, $\mu_l < \gamma/2$. This is the stability condition for this chain. Then,

$$\pi_0 = \left(1 + \sum_{l=1}^{k} \frac{\rho_l}{1 - \rho_l}\right)^{-1}$$

and the capacity is

$$C = q \sum_{l=1}^{k} \sum_{j=1}^{\infty} \pi_l^{(j)} (\gamma - \mu_l) = \frac{q \sum\limits_{l=1}^{k} \frac{\mu_l}{1-\rho_l}}{1 + \sum\limits_{l=1}^{k} \frac{\rho_l}{1-\rho_l}} = \frac{q\gamma}{2}. \tag{1}$$

See Appendix 8 for a proof of the last equality. The distribution of the number of stored entanglements is

$$P(Q = j) = \begin{cases} \pi_0, & \text{if } j = 0, \\ \sum\limits_{l=1}^{k} \pi_l^{(j)} = \pi_0 \sum\limits_{l=1}^{k} \rho_l^j, & \text{if } j > 0. \end{cases}$$

The expected number of stored entanglements is

$$E[Q] = \sum_{j=1}^{\infty} j P(Q = j) = \sum_{j=1}^{\infty} j \pi_0 \sum_{l=1}^{k} \rho_l^j = \frac{\sum\limits_{l=1}^{k} \frac{\rho_l}{(1-\rho_l)^2}}{1 + \sum\limits_{l=1}^{k} \frac{\rho_l}{1-\rho_l}}, \tag{2}$$

where, in the last equality, we apply Tonelli's Theorem.

4.1.2 Finite Buffer

In the case of heterogeneous links and a finite buffer of size B, the CTMC has the same structure as in Figure 1, except that each "arm" of the chain terminates at $B\boldsymbol{e}_l$, $\forall l \in \{1, \ldots, k\}$. The balance equations are

$$\begin{aligned}
&\pi_0 \mu_l = \pi_l^{(1)} (\gamma - \mu_l),\ l \in \{1, \ldots, k\}, \\
&\pi_l^{(j-1)} \mu_l = \pi_l^{(j)} (\gamma - \mu_l),\ l \in \{1, \ldots, k\},\ j \in \{2, \ldots, B\}, \\
&\pi_0 + \sum_{l=1}^{k} \sum_{j=1}^{B} \pi_l^{(j)} = 1
\end{aligned}$$

and have solution

$$\pi_l^{(j)} = \rho_l^j \pi_0,\ l \in \{1,\dots,k\},\ j \in \{1,\dots,B\},$$

where ρ_l is defined as in the infinite buffer case. Then,

$$\pi_0 \left(1 + \sum_{l=1}^{k}\sum_{j=1}^{B} \rho_l^j\right) = 1, \quad \text{hence}$$

$$\pi_0 = \left(1 + \sum_{l=1}^{k}\sum_{j=1}^{B} \rho_l^j\right)^{-1},$$

and the capacity is

$$C = q\sum_{l=1}^{k}\sum_{j=1}^{B}(\gamma - \mu_l)\pi_l^{(j)} = \frac{q\sum\limits_{l=1}^{k} \frac{\mu_l(1-\rho_l^B)}{1-\rho_l}}{1 + \sum\limits_{l=1}^{k} \frac{\rho_l(1-\rho_l^B)}{1-\rho_l}}. \tag{3}$$

The distribution of the number of stored qubits is given by

$$P(Q = j) = \begin{cases} \pi_0, & \text{if } j = 0, \\ \sum\limits_{l=1}^{k} \pi_l^{(j)} = \pi_0 \sum\limits_{l=1}^{k} \rho_l^j, & \text{if } 0 < j \leq B. \end{cases}$$

The expected number of stored qubits is

$$E[Q] = \sum_{j=1}^{B} jP(Q = j) = \frac{\sum\limits_{l=1}^{k} \frac{\rho_l\left(B\rho_l^{B+1} - (B+1)\rho_l^B + 1\right)}{(1-\rho_l)^2}}{1 + \sum\limits_{l=1}^{k} \frac{\rho_l(1-\rho_l^B)}{1-\rho_l}}.$$

The rate received by user l (connected to link l) is given by

$$C_l = q\left((\gamma - \mu_l)\sum_{j=1}^{B} \pi_l^{(j)} + \mu_l \sum_{\substack{m=1,\\ m\neq l}}^{k} \sum_{j=1}^{B} \pi_m^{(j)}\right), \tag{4}$$

where the first term represents the production of entanglements by link l (which get consumed by other links at rate $\gamma - \mu_l$) and the second term represents the consumption by link l of stored entanglements at other links. Note then, that if we were to sum all C_l, each end-to-end entanglement would be double-counted. Hence, $\sum C_l = 2C$. (Note: in the infinite-buffer case, $C_l = q\mu_l$, $l \in \{1, \ldots, k\}$; see Appendix 8 for a proof. Then, $\sum C_l = q\gamma = 2C$, another proof of the last equality in Eq. (1).) The expected number of stored qubits at link l, $E[Q_l]$ can be obtained by taking the lth component of the sum in the numerator of the expression for $E[Q]$. In other words, when $B = \infty$,

$$E[Q_l] = \frac{\frac{\rho_l}{(1-\rho_l)^2}}{1 + \sum\limits_{l=1}^{k} \frac{\rho_l}{1-\rho_l}}.$$

For a homogeneous system, $E[Q_l] = E[Q]/k$.

4.2 The Homogeneous Case

Suppose all links (or users) have the same entanglement generation rates, *i.e.* $\mu_l = \mu$, $\forall\ l \in \{1, \ldots, k\}$. We can take advantage of this homogeneity as follows: since only one link can be associated with stored qubits at the switch at any given time, and all links have equal rates, it is only necessary to keep track of the *number* of stored entanglements, and not the *identity* of the link (or user). Hence, the state space of the CTMC can be represented by a single variable taking values in $\{0, 1, \ldots, B\}$ where $B = \infty$ corresponds to the infinite buffer case, and $B < \infty$ the finite buffer case. We discuss each of these in detail next.

4.2.1 Infinite Buffer

Figure 2 depicts the CTMC for k homogeneous links and $B = \infty$. When no entangled qubits are stored (system is in state 0), any of the k links can generate a new entanglement, so the transition to state 1 occurs with rate $k\mu$. Let S represent the link associated with one or more stored entanglements. From states 1 and above, transitioning "forward" (or gaining another entanglement in storage) occurs whenever link S generates a new entanglement. This event occurs with rate μ. Finally, moving "backward" through the chain (corresponding to consuming a stored entanglement, when the switch performs a BSM) occurs whenever any of the $k - 1$ links other than S successfully generate an entanglement; this event occurs with rate $(k - 1)\mu$. It is easy to show that when there are two links, the system is not stable (and a stationary distribution does not exist). Take, for instance, the stability condition for a heterogeneous system with infinite buffer from Section 4.1.1:

$$\mu_l < \frac{\gamma}{2} = \frac{\sum\limits_{l=1}^{k} \mu_l}{2}.$$

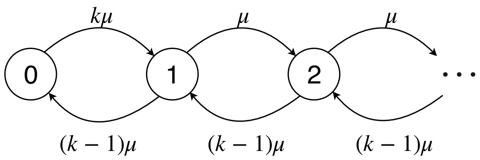


Figure 2: A CTMC model with k users, infinite buffer, and homogeneous links. μ is the entanglement generation rate.

Setting all μ_l's equal yields the stability condition $k > 2$ for the homogeneous system with infinite buffer. Henceforth, we only consider $k \geq 3$.

Note that the CTMC in Figure 2 is a birth-death process whose stationary distribution can be obtained using standard techniques found in literature (*e.g.*, [43]). The steady-state probability of being in state 0 is $\pi_0 = (k-2)/(2(k-1))$ and of being in state j is $\pi_j = k(k-2)/(2(k-1)^{j+1})$. The capacity is

$$C = q\sum_{i=1}^{\infty} \pi_i (k-1)\mu = q(k-1)\mu(1-\pi_0) = \frac{q\mu k}{2}.$$

Note that this result is also obtained by setting all μ_l equal to μ in Eq. (1). The expected number of stored entangled pairs is given by

$$E[Q] = \sum_{i=0}^{\infty} i\pi_i = k\pi_0 \sum_{i=1}^{\infty} i\left(\frac{1}{k-1}\right)^i = \frac{k}{2(k-2)}.$$

Note that this result can be obtained by setting all μ_l equal to μ in Eq. (2). An interesting outcome of setting all $\mu_l = \mu$ is that for $E[Q]$, there is no longer a dependence on the entanglement generation rate; this is in contrast to the heterogeneous system with infinite-size buffer. Further, when the links are homogeneous, as their number grows, $E[Q]$ approaches 1/2, implying that in such a scenario, as long as the switch operates at or near capacity (as it does under our switching policy), little quantum storage is required – one or two quantum memories per link would suffice, to be precise. An interesting question left for future study is to investigate how these storage requirements would change under a different entanglement switching policy.

The more general case of multipartite entanglement switching (*i.e.*, $n \geq 2$) for homogeneous-link systems with infinite buffer and no quantum state decoherence is covered in [32].

4.2.2 Finite Buffer

Figure 3 illustrates the CTMC for a system with k homogeneous links being served by a switch with finite buffer space B. When there are B stored entanglements and a new one is generated on link S, we assume that the switch drops the oldest stored entanglement, adhering to the

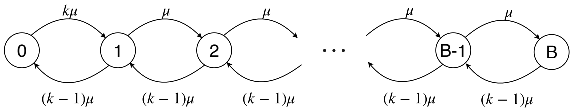


Figure 3: A CTMC model with k users, finite buffer of size B, and homogeneous links. μ is the entanglement generation rate.

OLEF policy. This CTMC is also a standard birth-death process whose solution can be found in literature (*e.g.*, [43]) and has

$$\pi_0 = \frac{(k-2)(k-1)^B}{2(k-1)^{B+1} - k}.$$

Using the fact that $\sum_{i=1}^{B} \pi_i = 1 - \pi_0$, the capacity is

$$C = q \sum_{i=1}^{B} \mu(k-1)\pi_i = \frac{q\mu k \left(1 - \left(\frac{1}{k-1}\right)^B\right)}{2 - k\left(\frac{1}{k-1}\right)^{B+1}}.$$

Note that as $B \to \infty$, C for the finite buffer case approaches C for the infinite buffer case. The expected number of stored qubits is

$$E[Q] = \sum_{i=1}^{B} i\pi_i = \frac{k\left(B + (k-1)^{B+1} - (B+1)(k-1)\right)}{(2(k-1)^{B+1} - k)(k-2)}.$$

As for the infinite-buffer case, for a homogeneous-link system with a finite-size buffer, there is no dependence in $E[Q]$ on the entanglement generation rates (in contrast to a heterogeneous-link system).

4.3 Decoherence

Assume now that quantum states in our system are subject to decoherence with an associated cut-off policy for qubit storage as described in Section 3. Further, assume that all states decohere at the same rate α, even in the case of heterogeneous links; see Section 6 for a discussion on relaxing this assumption for the case of link-dependent coherence or cut-off times. Under the assumption that coherence time is exponentially distributed with rate $1/\alpha$, incorporating decoherence does not change the structure of the CTMC; it merely increases "backward" transition rates. Specifically, in the homogeneous case, the transition from any state $j \geq 1$ to state $j-1$ now has rate $(k-1)\mu + j\alpha$, where $j\alpha$ represents the aggregate decoherence rate of all j stored qubits. In the heterogeneous case, the transitions are modified in a similar manner for any state $j\boldsymbol{e}_l$, $l \in \{1, \ldots, k\}$, $j \geq 1$. The derivations of stationary distributions, capacities, and expected number of qubits stored are very similar to those for models without decoherence; we present the final relevant expressions here and leave details to Appendix 9. All expressions below can be computed numerically. **Heterogeneous Links:** For finite buffer size $B < \infty$,

$$\begin{aligned}
\pi_0 &= \left(1 + \sum_{l=1}^{k}\sum_{j=1}^{B}\prod_{i=1}^{j}\frac{\mu_l}{\gamma - \mu_l + i\alpha}\right)^{-1}, \\
C &= q\pi_0 \sum_{l=1}^{k}\sum_{j=1}^{B}(\gamma - \mu_l)\prod_{i=1}^{j}\frac{\mu_l}{\gamma - \mu_l + i\alpha}, \\
E[Q] &= \pi_0 \sum_{j=1}^{B} j \sum_{l=1}^{k}\prod_{i=1}^{j}\frac{\mu_l}{\gamma - \mu_l + i\alpha}.
\end{aligned}$$

For infinite-size buffer, let $B \to \infty$ in all expressions above. **Homogeneous Links:** For finite buffer size $B < \infty$,

$$\begin{aligned}
\pi_0 &= \left(1 + k\sum_{i=1}^{B}\prod_{j=1}^{i}\frac{\mu}{((k-1)\mu + j\alpha)}\right)^{-1}, \\
C &= q(k-1)\mu(1-\pi_0), \\
E[Q] &= \pi_0 k \sum_{i=1}^{B} i \prod_{j=1}^{i}\frac{\mu}{((k-1)\mu + j\alpha)}.
\end{aligned}$$

For infinite-size buffer, let $B \to \infty$ in all expressions above.

5 Numerical Observations

In this section, we investigate the capacity and buffer requirements of a bipartite entanglement switch based on our model. In particular, we are interested in how buffer capacity B and number of users k affect capacity and $E[Q]$. We then examine the effect of decoherence and qubit storage cut-off times on homogeneous and heterogeneous switches with finite as well as infinite buffer capacities. Next, via simulation we look at some examples of a deterministic cut-off policy for qubit storage and compare the results to our probabilistic one; we also validate our analytical expressions for decoherence using both types of simulations. Last, we compare the result of our CTMC model to another, discrete-time Markov chain (DTMC) model of the switch studied in [25].

Throughout this section, we denote the distance of user l from the switch as L_l (measured in km). It is implicitly assumed in our model of a quantum switch that in addition to the B quantum memories used solely to store entangled qubits, each link has available to it another set of memories which are used solely to assist with the entanglement generation protocol. Specifically, for a link of length L_l and speed of light c_f in fiber, there would be an initial delay of approximately $T = 2L_l/c_f$ for the switch to receive a notification from the user of whether the first entanglement generation attempt was successful. For subsequent entanglement generation attempts, however, the switch receives a notification every τ seconds, which is the time between entanglement attempts at a given link (see more discussion on the repetition rate below). Thus, for the system to be at all operational, the quantum memories that are assisting in the entanglement generation protocol would need coherence times of at least T, which we assume to be the case from now on. Further, the number of these additional memories per link is $\lceil T/\tau \rceil$, so that at the switch, the time between notification arrivals is τ. In summary, T affects only the initial latency, the number of additional memories needed at each link, and the initial fidelity of entanglement immediately before the qubit is moved to one of the B storage memories, but it does not affect the successful entanglement generation rate of a link, nor the capacity of the switch. Thus, T does not enter into our steady-state analyses and we may disregard it henceforth.

We assume that each user is connected to the switch with single mode optical fiber of loss coefficient $\beta = 0.2$ dB/km. We also assume that the switch is equipped with a photonic entanglement source with a raw (local) entanglement generation rate of 1 Mega-ebits[3] per second. So, in every (1 μs long) time slot, one photon of a Bell state is loaded into a memory local to the switch, and the other photon is transmitted (over a lossy optical fiber) to a user, who loads the received photon into a memory (held by the user), which has a trigger which lets the user know the time slots in which their memory successfully loads a photon. We choose a 1 MHz clock rate because is not far from near-term realizations; *e.g.*, in [18] a similar rate was achieved with silicon vacancy color centers in diamond. Let us denote $\tau = 1$ μs as the time duration of one qubit of each entangled pair, and the entanglement generation rate between the switch and the user l, $\mu_l = c\eta_l/\tau$ ebits per second. Here, we take $c = 0.1$ to account for various losses other than the transmission loss in fiber, for example inefficiencies in loading the entangled photon pair in the two memories (at the switch and at the user), and any inefficiency in a detector in the memory at the user used for heralding the arrival of a photon (*e.g.*, by doing a Bell measurement over the received photon pulse and one photon of a locally-generated two-photon entangled state produced by the user). Here, η_l, the transmissivity of the optical fiber connecting user l and the switch is given by $\eta_l = 10^{-0.1\beta L_l}$. Channel loss to user l, measured in dB, is $10\log_{10}(1/\eta_l)$. Unless otherwise stated, all μ_l discussed in this section have units of Kilo-ebits/sec.

5.1 Effect of Buffer Size: Homogeneous Links

In homogeneous-link systems, all users are equidistant from the switch (*i.e.*, $L_l = L_m$, $\forall l, m \in \{1, \ldots, k\}$). In Figure 4, we compare models with infinite and finite buffer sizes as the number

[3] An ebit is one unit of bipartite entanglement corresponding to the state of two maximally entangled qubits, the so-called Bell or EPR state.

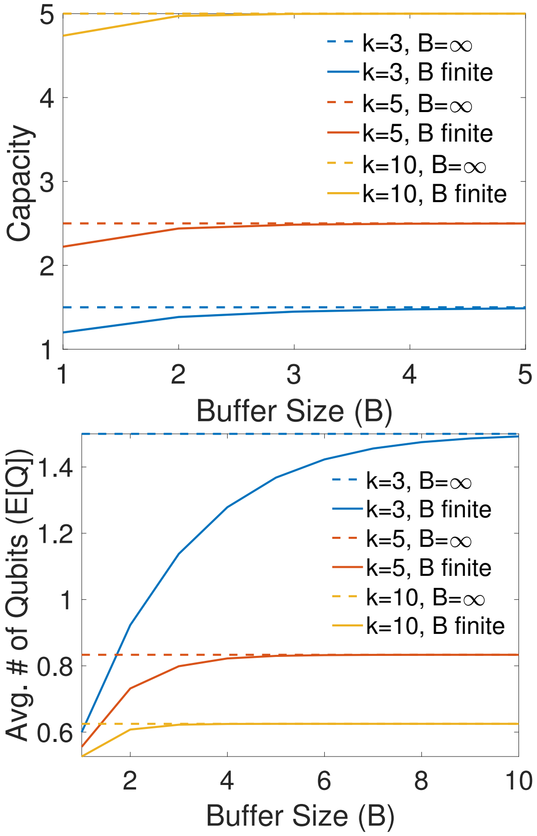


Figure 4: The effect of buffer size on capacity (top) and on the expected number of stored qubits (bottom) in systems with homogeneous links. Capacity is in Kilo-ebits/sec.

of links k is varied. Recall that when links are homogeneous, $q\mu$ is simply a multiplicative factor in the expressions for C, and does not factor into formulas for $E[Q]$. Hence, we set $q\mu = 1$ for Figure 4 (top), and with $\mu = 1$, the links are 100 km long. For the finite buffer models, B is varied from one to five. Recall from Section 4.2.2 that as $B \to \infty$, the capacity of the finite-buffer model approaches that of the infinite-buffer model, as expected, and note that the same is true when $k \to \infty$. Interestingly, this convergence occurs rapidly, even for the smallest value of k (3), and the maximum relative difference between the two capacities is 0.25 (even as μ increases). From this, we conclude that buffer does not play a major role in the capacity of a homogeneous-link system under the switching policy described in Section 3 and only a small quantum memory is required.

Figure 4 (bottom) shows the behavior of $E[Q]$ for infinite and finite buffer sizes and different values of k. As with capacity, the effect of buffer capacity on $E[Q]$ diminishes as k grows, and the largest relative difference occurs for $k = 3$ and $B = 1$, and equals $1.5 -$

less than two qubits. Note from the expressions for $E[Q]$ in Sections 4.2.1 and 4.2.2 that as $k \to \infty$, $E[Q] \to 1/2$. Numerically, we observe that convergence to this value occurs quickly: even for $k = 25$, $E[Q]$ is already 0.54 for both the infinite and finite buffer models.

In Figure 4, we also observe that C increases, but $E[Q]$ decreases with k. The reason for the higher capacity is that in a homogeneous system, as the number of links grows, so does the rate of successfully-generated link-level entanglement (when viewed across all links), creating more opportunities for the switch to perform a BSM. At the same time, these extra BSM opportunities result in entangled qubits spending less time in storage – hence the decrease in $E[Q]$.

5.2 Effect of Buffer Size: Heterogeneous Links

Figure 5 illustrates how buffer size and number of users affect C and $E[Q]$ for a set of heterogeneous systems. We vary the number of links from three to nine. For each value of k, the links are split into two classes: links in the first class successfully generate entanglement at rate μ_1 and those in the second class at rate μ_2. We set $\mu_1 = 1.9\mu_2$ and $\mu_2 = 1$. This setting corresponds to links in class one having lengths 86 km and links in class two having lengths 100 km. Values of μ_1 and μ_2 are chosen in a manner that satisfies the stability condition for heterogeneous systems: recall from Section 4.1.1 that for all $l \in \{1, \ldots, k\}$, μ_l must be strictly less than half the aggregate entanglement generation rate. For all experiments, $q = 1$ since it only scales capacity.

For each value of k, the ratio of class 1 to class 2 links is 1:2 (so $k = 3, 6, 9$ have one, two, and three class 1 links, respectively). As with the homogeneous-link systems, we observe that the slowest convergence of the finite-buffer metrics C and $E[Q]$ to corresponding infinite-buffer metrics is for smaller values of k and the largest relative difference is for smaller values of B. However, the rate of convergence speeds up quickly as k increases from 3 to 6: with the latter, convergence is already observed for $B < 10$. Meanwhile, when $k = 9$, there is little benefit in having storage for more than two qubits. Another interesting observation is that quantum memory usage is large when $k = 3$ but not for larger values of k. This is due to the system operating closer to the stability constraints for $k = 3$ than for larger values of k. In the next section, we will see another example of a system that operates near the boundary of its stability region. In such cases, C and $E[Q]$ can be affected significantly as B is varied.

5.3 Effect of Decoherence

In this section, we study the effect of decoherence and associated qubit storage cut-off times on capacity and expected number of stored qubits $E[Q]$. We set $q = 1$ for all experiments since it only scales capacity. Figure 6 presents C and $E[Q]$ for a homogeneous system with $\mu = 1$ (corresponding to 100 km long links), $B = \infty$ and different values of k, as decoherence rate α varies from 0 (the equivalent of previous models that did not incorporate decoherence) to 1, which is equal to μ. Note that in practice, α is expected to be much smaller than μ.

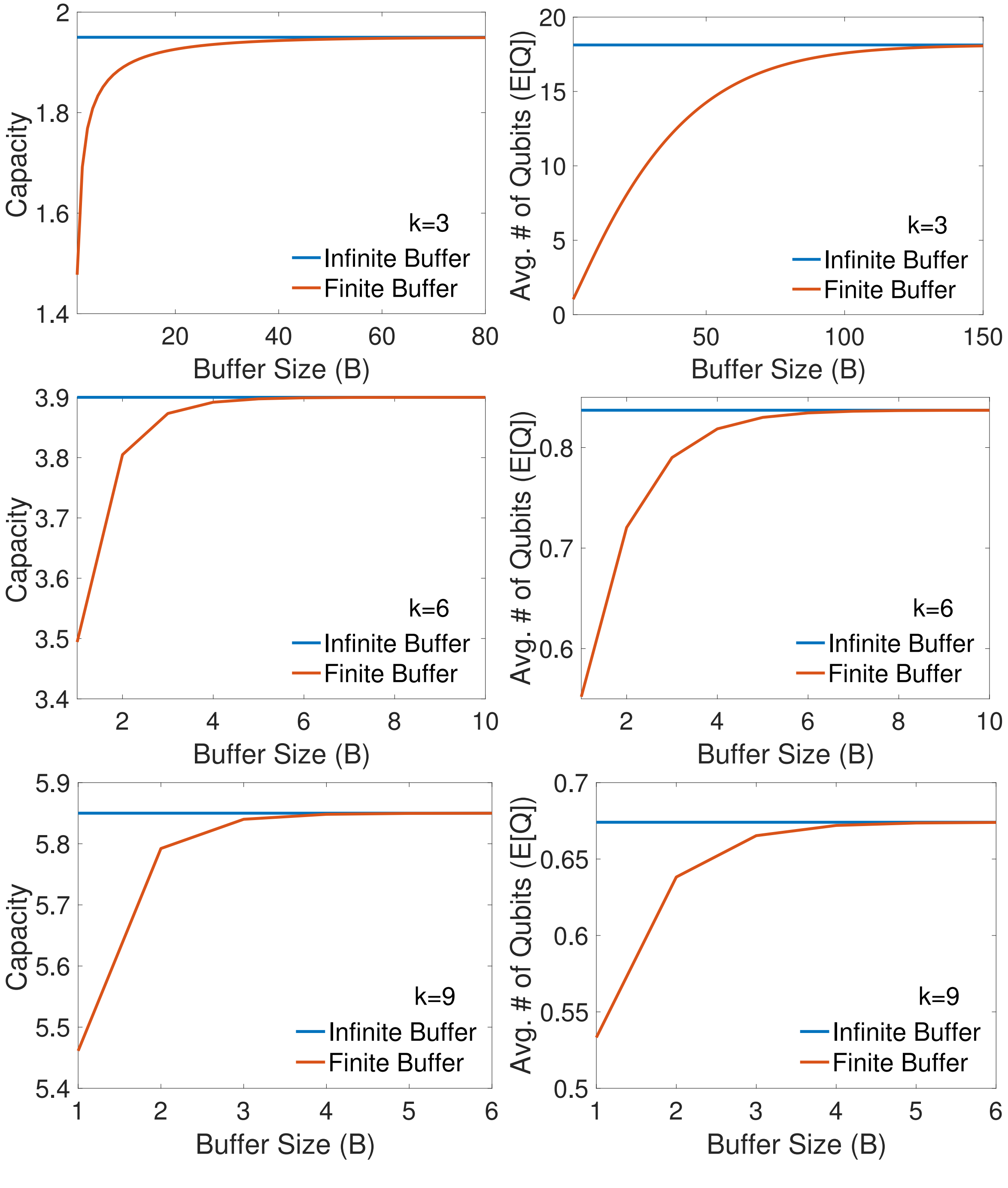


Figure 5: Capacity (Kilo-ebits/sec) and expected number of qubits in memory $E[Q]$ for heterogeneous systems with varied number of links and buffer sizes. Links are divided into two classes: one class generates entanglement approximately twice as quickly as the other class.

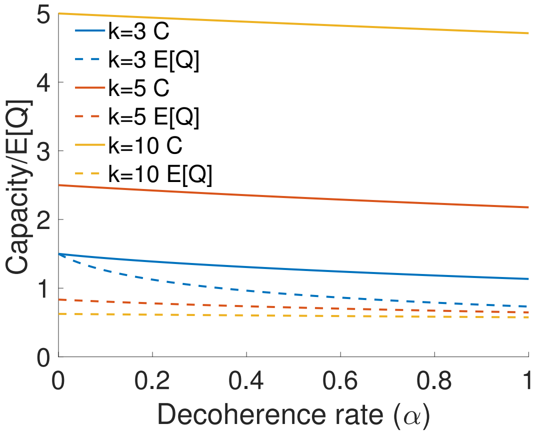


Figure 6: Effect of decoherence on capacity (Kilo-ebits/sec) and expected number of stored qubits $E[Q]$, for varying number of users k. For all experiments, $B = \infty$ and the entanglement generation rate is $\mu = 1$ for all links.

We observe that even as α approaches μ decoherence does not cause major degradation in capacity for homogeneous systems, and likewise does not introduce drastic variations in $E[Q]$.

Figure 7 presents the effect of α on the performance of a heterogeneous system with infinite-size buffer. In these experiments, entanglement generation rates are set in a similar manner to that of Section 5.2, with two classes of links configured so that the first class generates entanglements almost twice as fast as the second class (here, $\mu_1 = 0.99$ and $\mu_2 = 0.5$, corresponding to 100.2 km and 115 km long links for class one and two, respectively), and the number of links in class one to those in class two is 1:2. In these experiments, for each value of k, capacity behaves much as it would in a homogeneous system with μ set as the average of the μ_l from the heterogeneous system. Note that for $k = 3$, $E[Q]$ is very large when $\alpha = 0$; similar to the experiment in Figure 5 (see panel with $k = 3$) this is because the system is operating near the boundary of its stability region. In all other cases, $E[Q]$ is close to 0.

In Figure 8(a), we focus on a heterogeneous system that operates near the boundary of its stability region and observe the effects of both decoherence and buffer size on C and $E[Q]$. There are five links, with entanglement generation rates (35 15 15 3 3) Kilo-ebits/sec, corresponding to link lengths of 22.8, 41.2, 41.2, 76, and 76 km, respectively. For this system, $\gamma/2 = 35.5$, so the fastest link is just below the constraint when $\alpha = 0$. The average of the μ_l is 14.2, so α is varied from 0 to this value. B is varied from 1 to 100, with the latter being close enough to mimic infinite buffer behavior for C and $E[Q]$. Figure 8(b) presents the performance of a homogeneous system with $k = 5$ and $\mu = 14.2$ for a comparison. We observe that the homogeneous system achieves higher capacity for all values of B, even though the average entanglement generation rate is the same for both systems. Further, the

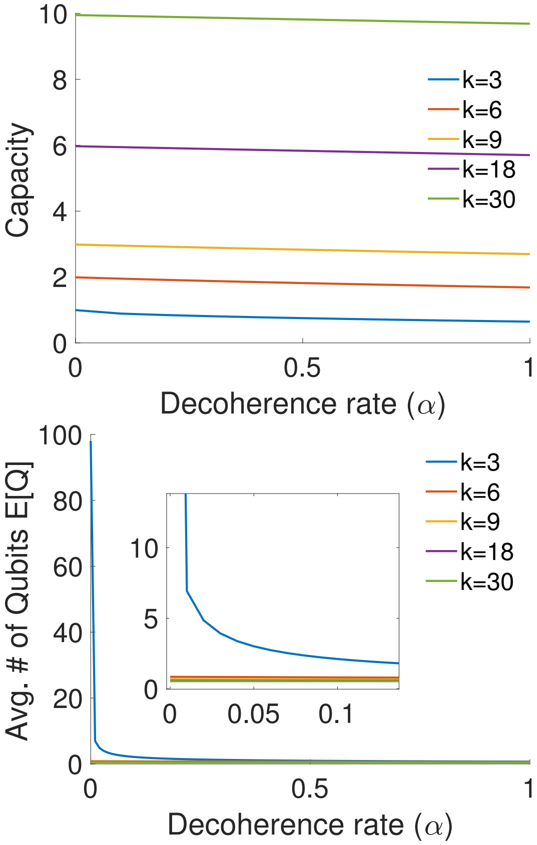
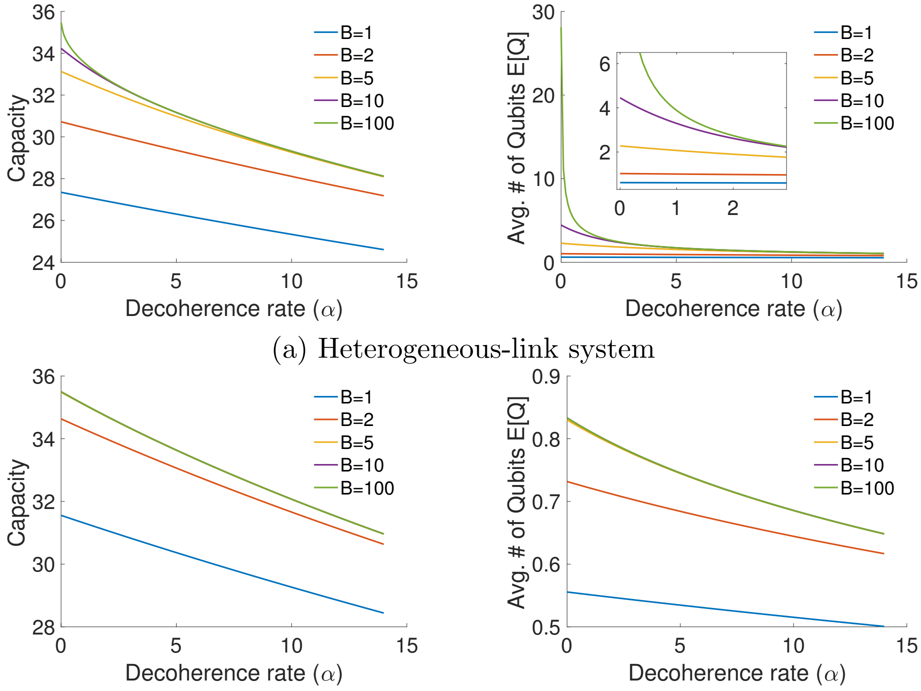


Figure 7: Effect of decoherence and associated storage cut-off times on capacity (Kilo-ebits/sec) and expected number of stored qubits $E[Q]$, for varying number of users k. In all experiments, the links are heterogeneous and the buffer size is infinite. The inset in the bottom figure zooms into the area near the origin.

homogeneous system is more robust to changes in buffer size than the heterogeneous system: for the former, $B = 5, 10$ are equivalent to $B = 100$. Further, note that for $B = 100$ and $\alpha = 0$ the heterogeneous system performs almost as well as the homogeneous system in terms of capacity, but the memory usage is much higher for the former. Finally, for this buffer size, as α increases, the homogeneous system is more robust to the effects of decoherence: capacity degrades by 7.35 Kilo-ebits/sec for the heterogeneous system between $\alpha = 0$ and $\alpha = 14$, while it degrades by 4.54 Kilo-ebits/sec for the homogeneous system.



(a) Heterogeneous-link system

(b) Homogeneous-link system

Figure 8: Effect of decoherence and associated storage cut-off times on capacity (Kilo-ebits/sec) and expected number of stored qubits $E[Q]$ for $k = 5$ links and varying buffer sizes B. The inset in the first $E[Q]$ plot zooms into the area near the origin. In (a), μ_l are (35 15 15 3 3), and in (b), μ is the average of μ_l, $l = 1, \ldots, 5$, i.e., 14.2. For all plots above, $B = 100$ curves behave equivalently to $B = \infty$.

5.4 Deterministic vs Probabilistic Cut-Off Policy for Qubit Storage

Recall from Section 3, that we approximate deterministic storage cut-off times for entangled qubits using a probabilistic model, where the cut-off (or coherence) times are exponentially-distributed with mean $1/\alpha$. In this section, we simulate both cut-off time implementations. For the deterministic cut-off policy, we keep all entangled qubits in storage for an equal amount of time $1/\alpha$. For additional validation, we compare both simulations with our analytical expressions. For all experiments in this section, each datapoint that is obtained via simulation is an average of five simulation runs.

Figure 9(a) presents a comparison for a homogeneous-link system with $B = 5$ and entanglement generation rates of 1 Kilo-ebits/sec for all links. The maximum relative error for the capacity, defined as

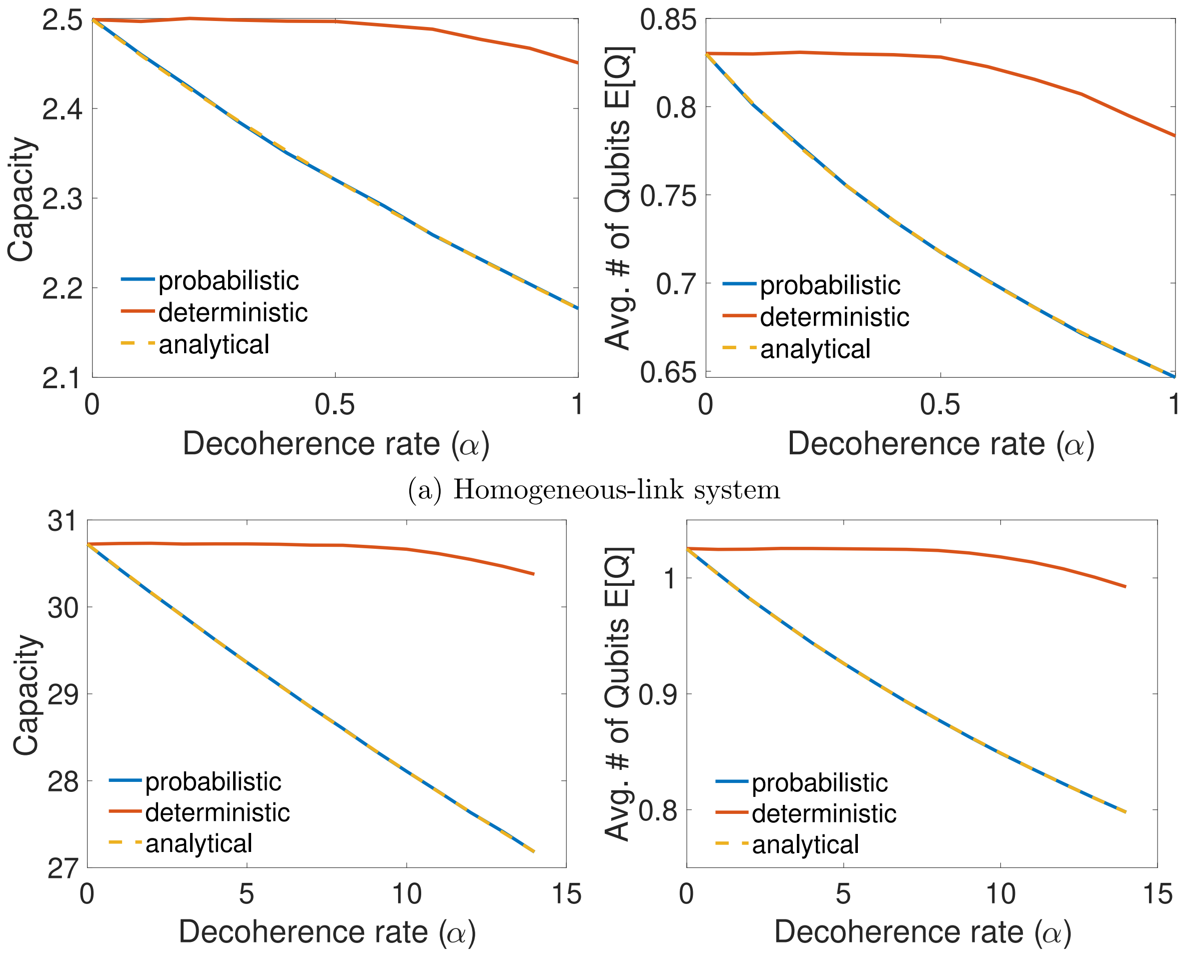
$$maxRelErr_C = \max_{\alpha} \frac{|C_{\text{det}(\alpha)} - C_{\text{prob}}(\alpha)|}{C_{\text{det}}(\alpha)}, \tag{5}$$

is 11% and the maximum relative error for $E[Q]$ is 17%. Note, however, that the maximum of the errors occurs for $\alpha = 1$, a decoherence rate that may be considered exceedingly high for a real implementation. For more realistic values of α (an order of magnitude smaller than μ), the relative errors appear acceptable, as they would yield a difference of less than 0.3 Kilo-ebits/sec in capacity predictions and a difference of well under one qubit in $E[Q]$ predictions.

Figure 9(b) presents a comparison for a heterogeneous-link system with $B = 2$ and entanglement generation rates of (35 15 15 3 3). Decoherence/cut-off rate is varied from $\alpha = 0$ to $\alpha = 14$, the latter approximately the average of the entanglement generation rates. The maximum relative errors for for C and $E[Q]$ are 10.5% and 19.6%, respectively. As with the homogeneous-link experiment, these maxima correspond to the highest value of α (14). The average relative errors for C and $E[Q]$, taken over all values of $\alpha \in \{0, 1, \ldots, 14\}$, are 5.8% and 11.7%, respectively. Overall, the predictions in C differ by 3 Kilo-ebits/sec in the worst case, and the predictions in $E[Q]$ differ by far less than a qubit even in the worst case. Thus, the approximation appears reasonable for this heterogeneous-link example as well.

5.5 Comparison of DTMC model with CTMC models

Until now, we have only employed CTMCs to model and analyze variants of the system described in Section 3. A more accurate way to model such a system is to instead construct a DTMC on the appropriate state space, as done in [25]. To do so, we assume that at each time step of length τ seconds, all k users attempt to generate link-level entanglements. Link l succeeds in generating an entanglement with probability p_l. In [25], we show that unfortunately, this method is not the most scalable (in terms of k or n) and is not the easiest to analyze even in the simple setting of homogeneous links and infinite switch buffer size. Further obstacles arise when one considers, for example, accounting for decoherence



(a) Homogeneous-link system

(b) Heterogeneous-link system

Figure 9: Comparisons of the deterministic and probabilistic qubit storage cut-off time policies (simulations), against analytical results. For the homogeneous-link system, $\mu = 1$ and $B = 5$. For the heterogeneous-link system, μ_l are (35 15 15 3 3) and $B = 2$.

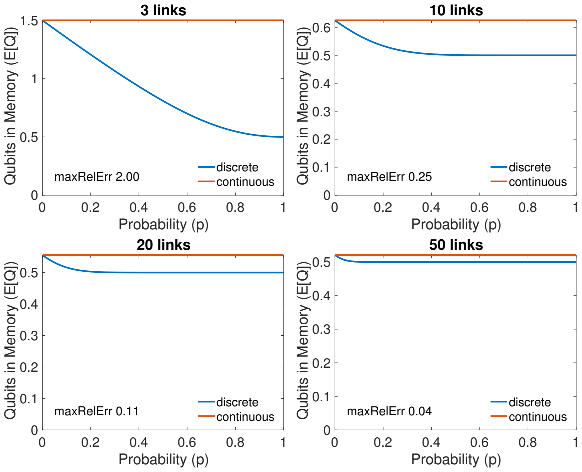


Figure 10: Comparison of the expected number of qubits in memory $E[Q]$ for the DTMC and CTMC models, as the number of links is varied $\in \{3, 10, 20, 50\}$ and for entanglement generation probabilities $p \in (0, 1)$. maxRelErr is the maximum relative error between discrete and continuous values for $E[Q]$.

in a DTMC model. When using a CTMC, we approximate the operation of the switch by viewing link-level entanglements as exponential random variables with generation rate equal to $\mu_l = p_l/\tau$ for link l, instead of viewing them as Bernoulli trials. The analysis is significantly less challenging with CTMCs.

We will now compare the results of the DTMC and CTMC for a homogeneous system with infinite buffer and no decoherence, as this is the only result we were able to obtain for the former in [25]. Note that in the discrete model, the amount of time it takes to successfully generate a link entanglement is given by τ/p. In the continuous model, the rate of successful entanglement generation is μ, so the time to generate an entanglement is $1/\mu$. Hence, $\tau/p = 1/\mu$ or equivalently, $\mu = p/\tau$. The DTMC capacity of $qkp/2$ that we derived in [25] is the capacity per time slot of length τ seconds. Therefore, in order to make a comparison against the CTMC capacity, we must perform a unit conversion: divide the discrete capacity by τ in order to obtain the number of entanglement pairs per *second*, as opposed to per *time slot*. This yields

$$C_{\text{DTMC}} = \frac{qkp}{2\tau} = \frac{qk\mu}{2} = C_{\text{CTMC}}.$$

We conclude that the capacities produced by the DTMC and CTMC models match exactly.

Next, we compare the expected number of qubits in memory at the switch, $E[Q]$ as predicted by the DTMC and the CTMC models. Figure 10 compares numerically the discrete and continuous $E[Q]$'s as the number of users k and probability p vary. For each value of p and k, we numerically solve for the discrete $E[Q]$, since we do not have a closed-form expression for it due to being unable to analytically solve for the stationary distribution of the DTMC. For each value of k, we report the maximum relative error, defined as

$$maxRelErr(k) = \max_{p \in (0,1)} \frac{|E[Q]_{\mathrm{DTMC}}(k,p) - E[Q]_{\mathrm{CTMC}}(k)|}{E[Q]_{\mathrm{DTMC}}(k,p)},$$

where $E[Q]_{\mathrm{DTMC}}$ and $E[Q]_{\mathrm{CTMC}}$ are the discrete and continuous functions for $E[Q]$, respectively. We observe that the error is largest when p is close to 1. In [25], we argue that as $k \to \infty$, $E[Q]_{\mathrm{DTMC}}$ and $E[Q]_{\mathrm{CTMC}}$ both approach 1/2. We conclude that as $k \to \infty$, $maxRelErr \to 0$, which can be observed in Figure 10. Also, the largest $maxRelErr$ occurs for the lowest value of $k = 3$, when $p \to 1$. But even in this (worst case), although the error is $maxRelErr(3) = 2$, it corresponds to discrete and continuous versions of $E[Q]$ differing by a prediction of only a single qubit. From these analytic and numerical observations, we conclude that the CTMC model is sufficiently accurate so as to be used to explore issues such as decoherence, link heterogeneity, and switch buffer constraints.

6 Relaxing Modeling Assumptions

In this work, we only study the effects that our decoherence model and storage cut-off time policy have on the capacity of the switch, but not the effects on the quality of entanglement. Thus, our model is applicable in rather general settings, where the initial entanglement fidelities at the link level may differ from link to link and are given by F_l, $l = 1, \ldots, k$. In our model of decoherence in Section 4.3, the cut-off time is configured such that the entangled qubits are held in memory for an amount of time $1/\alpha$ on average. We note that the model may be easily extended to include a link-dependent cut-off time $t_l^\star \equiv 1/\alpha_l$, to enable configurable cut-off times for each link's quantum memory storage, in scenarios where an application requests a minimum fidelity of F_{thresh} for each link (or alternatively, a minimum final end-to-end fidelity $F'_{\mathrm{thresh}} < F_{\mathrm{thresh}}$; without loss of generality, we focus here on F_{thresh}, the fidelity at the link level, to simplify the discussion). To make this change, one would simply compute the time $t_l^\star$ that it would take for the initial fidelity of entanglement at link l, F_l, to degrade to F_{thresh}, under a suitable decoherence model for a given platform. Then, within the model, set $\alpha_l{:=}1/t_l^\star$ and modify each transition in the CTMC accordingly: the aggregate decoherence rate from state $j\boldsymbol{e}_l$ is now $j\alpha_l$. Intuitively, the less time that is needed for the entangled link's fidelity to degrade to F_{thresh} (meaning, either F_l is close to F_{thresh} or $t_l^\star$ is small for a given noise model), the faster the decoherence rate α_l.

It is also possible to extend our model to account for a simple entanglement purification scheme. Suppose that the switch imposes a minimum fidelity F_{thresh} requirement on each link-level entanglement involved in a BSM. This means that any link with an initial entanglement fidelity $F_l < F_{\mathrm{thresh}}$ must first run a purification protocol. We may assume without loss of generality that F_l is sufficiently high so as to allow for a purification protocol that is implementable on the physical architecture of the switch to boost F_l to F_{thresh} with a non-zero probability. If no such purification protocol exists, then link l's qubits may never participate in a BSM, and therefore we may trivially ignore the link in all calculations. To incorporate the purification scheme into our model, it suffices to compute or estimate the rate of entanglement generation at the link level, of entanglement fidelity F_{thresh}; note that such rates may be link-dependent in the case of heterogeneous-link systems. Call this rate μ'_l. The final step is to substitute each parameter value μ_l with μ'_l; no further changes are required to the model.

7 Conclusion

In this work, we examined variants of a system with k users who are being served bipartite entangled states by a quantum entanglement distribution switch in a star topology. Each user is connected to the switch via a dedicated link; we considered both the case of homogeneous and heterogeneous links. We also analyzed cases in which the switch has finite or infinite buffer space for storing entangled qubits. We obtained simple and intuitive expressions for switch capacity, as well as for the expected number of qubits in memory when the switch operates at or near capacity.

We made numerical comparisons of these two metrics while varying the number of users k and buffer sizes B. We observed that in most cases, little memory is required to achieve the performance of an infinite-memory system. We also made numerical observations for models that incorporate decoherence and associated qubit storage cut-off times, and concluded that in homogeneous systems these phenomena have little effect on performance metrics, while they can have more significant consequences in heterogeneous systems that operate near the boundaries of their stability regions.

Acknowledgment

The work was supported in part by the National Science Foundation under grant CNS-1617437. Saikat Guha acknowledges support of an NSF subaward of a Yale University led project, grant number 1640959, “EFRI ACQUIRE: Integrated nanophotonic solid state memories for telecom wavelength quantum repeaters”.

Appendix

8 Capacity for Heterogeneous Systems with $B = \infty$

Throughout this appendix, assume that the stability conditions for the CTMC are met; *i.e.*, that for all l, $\mu_l < \gamma/2$.

Proof of the last equality in Eq. (1)

From the first part of this equation, we have

$$\begin{aligned}
C &= q\sum_{l=1}^{k}\sum_{j=1}^{\infty}\pi_l^{(j)}(\gamma-\mu_l)\\
&= q\sum_{l=1}^{k}\sum_{j=1}^{\infty}\pi_0\rho_l^j(\gamma-\mu_l)\\
&= q\pi_0\sum_{l=1}^{k}\frac{(\gamma-\mu_l)\rho_l}{1-\rho_l}\\
&= q\pi_0\sum_{l=1}^{k}\left(\frac{\gamma}{2}\frac{\rho_l}{1-\rho_l}+\left(\frac{\gamma}{2}-\mu_l\right)\frac{\rho_l}{1-\rho_l}\right)\\
&= q\pi_0\sum_{l=1}^{k}\left(\frac{\gamma}{2}\frac{\rho_l}{1-\rho_l}+\left(\frac{\gamma-2\mu_l}{2}\right)\frac{\mu_l(\gamma-\mu_l)}{(\gamma-\mu_l)(\gamma-2\mu_l)}\right)\\
&= q\pi_0\sum_{l=1}^{k}\left(\frac{\gamma}{2}\frac{\rho_l}{1-\rho_l}+\frac{\mu_l}{2}\right)\\
&= q\pi_0\frac{\gamma}{2}\left(\sum_{l=1}^{k}\frac{\rho_l}{1-\rho_l}+1\right)\\
&= \frac{q\gamma}{2}.
\end{aligned}$$

Proof that $C_l = q\mu_l$

Letting $B\to\infty$ in Eq. (4),

$$\begin{aligned}
C_l &= q\pi_0\left((\gamma-\mu_l)\frac{\rho_l}{1-\rho_l}+\mu_l\sum_{\substack{m=1,\\ m\neq l}}^{k}\frac{\rho_m}{1-\rho_m}\right)\\
&= q\pi_0\mu_l\left(\frac{1}{1-\rho_l}+\sum_{\substack{m=1,\\ m\neq l}}^{k}\frac{\rho_m}{1-\rho_m}+\frac{\rho_l}{1-\rho_l}-\frac{\rho_l}{1-\rho_l}\right)\\
&= q\pi_0\mu_l\left(1+\sum_{m=1}^{k}\frac{\rho_m}{1-\rho_m}\right)=q\mu_l.
\end{aligned}$$

9 Decoherence

Throughout this appendix, for systems with infinite buffer, assume that the corresponding stability conditions are satisfied, *i.e.*, $k > 2$ in homogeneous-link systems and $\mu_l < \gamma/2$, for all l, in heterogeneous-link systems.

Homogeneous, Infinite Buffer

For this system, the balance equations are as follows:

$$\begin{aligned}
\pi_0 k\mu &= \pi_1(\alpha + (k-1)\mu), \\
\pi_{i-1}\mu &= \pi_i(i\alpha + (k-1)\mu), \quad i = 2, 3, \dots, \\
\sum_{i=0}^{\infty} \pi_i &= 1.
\end{aligned}$$

Solving for the stationary distribution, we have:

$$\begin{aligned}
\pi_1 &= \frac{k\mu}{(k-1)\mu + \alpha}\pi_0, \\
\pi_2 &= \frac{\mu\pi_1}{(k-1)\mu + 2\alpha} = \frac{k\mu^2\pi_0}{((k-1)\mu + 2\alpha)((k-1)\mu + \alpha)},
\end{aligned}$$

and so on. In general, for $i = 1, 2, \dots$ we can write

$$\pi_i = \frac{\pi_0 k\mu^i}{\prod_{j=1}^{i}((k-1)\mu + j\alpha)} = \pi_0 k \prod_{j=1}^{i} \frac{\mu}{((k-1)\mu + j\alpha)}.$$

Using the normalizing condition, we have

$$\begin{aligned}
\pi_0 + k\pi_0 \sum_{i=1}^{\infty}\prod_{j=1}^{i} \frac{\mu}{((k-1)\mu + j\alpha)} &= 1, \quad \text{so that} \\
\pi_0 &= \left(1 + k\sum_{i=1}^{\infty}\prod_{j=1}^{i} \frac{\mu}{((k-1)\mu + j\alpha)}\right)^{-1}.
\end{aligned}$$

The capacity and $E[Q]$ can be computed numerically using the following formulas:

$$\begin{aligned}
C &= \sum_{i=1}^{\infty} \pi_i (k-1)\mu = (k-1)\mu(1 - \pi_0), \\
E[Q] &= \sum_{i=1}^{\infty} i\pi_i = \pi_0 k \sum_{i=1}^{\infty} i \prod_{j=1}^{i} \frac{\mu}{((k-1)\mu + j\alpha)}.
\end{aligned}$$

Homogeneous, Finite Buffer

The derivations are very similar to the previous case, with the only difference being that the balance equations are truncated at state $i = B$. The resulting expressions are almost identical to those above, with the exception of i being in $\{1, \dots, B\}$ instead of $\{1, 2, \dots\}$:

$$\begin{aligned}
\pi_0 &= \left(1 + k \sum_{i=1}^{B} \prod_{j=1}^{i} \frac{\mu}{((k-1)\mu + j\alpha)}\right)^{-1}, \\
C &= \sum_{i=1}^{B} \pi_i (k-1)\mu = (k-1)\mu(1-\pi_0), \\
E[Q] &= \sum_{i=1}^{B} i\pi_i = \pi_0 k \sum_{i=1}^{B} i \prod_{j=1}^{i} \frac{\mu}{((k-1)\mu + j\alpha)}.
\end{aligned}$$

Heterogeneous, Infinite Buffer

The balance equations are:

$$\begin{aligned}
&\pi_0 \mu_l = \pi_l^{(1)}(\gamma - \mu_l + \alpha),\ l \in \{1, \dots, k\}, \\
&\pi_l^{(j-1)} \mu_l = \pi_l^{(j)}(\gamma - \mu_l + j\alpha),\ l \in \{1, \dots, k\},\ j \in \{2, 3, \dots\}, \\
&\pi_0 + \sum_{l=1}^{k} \sum_{j=1}^{\infty} \pi_l^{(j)} = 1.
\end{aligned}$$

For $j = 1, 2, \dots,$ we can write

$$\pi_l^{(j)} = \pi_0 \prod_{i=1}^{j} \frac{\mu_l}{\gamma - \mu_l + i\alpha}.$$

Using the normalizing condition, we obtain

$$\pi_0 = \left(1 + \sum_{l=1}^{k} \sum_{j=1}^{\infty} \prod_{i=1}^{j} \frac{\mu_l}{\gamma - \mu_l + i\alpha}\right)^{-1}.$$

The capacity and $E[Q]$ can be computed numerically using

$$\begin{aligned} C &= \sum_{l=1}^{k}\sum_{j=1}^{\infty}\pi_l^{(j)}(\gamma-\mu_l) \\ &= \pi_0\sum_{l=1}^{k}\sum_{j=1}^{\infty}(\gamma-\mu_l)\prod_{i=1}^{j}\frac{\mu_l}{\gamma-\mu_l+i\alpha}, \\ E[Q] &= \sum_{j=1}^{\infty} jP(Q=j) = \sum_{j=1}^{\infty} j\sum_{l=1}^{k}\pi_l^{(j)} \\ &= \pi_0\sum_{j=1}^{\infty} j\sum_{l=1}^{k}\prod_{i=1}^{j}\frac{\mu_l}{\gamma-\mu_l+i\alpha}. \end{aligned}$$

Heterogeneous, Finite Buffer

The derivations are similar to the previous case, with the only difference being that j is now in $\{1,\dots,B\}$ instead of in $\{1,2,\dots\}$. The resulting relevant expressions are:

$$\begin{aligned} \pi_0 &= \left(1+\sum_{l=1}^{k}\sum_{j=1}^{B}\prod_{i=1}^{j}\frac{\mu_l}{\gamma-\mu_l+i\alpha}\right)^{-1}, \\ C &= \pi_0\sum_{l=1}^{k}\sum_{j=1}^{B}(\gamma-\mu_l)\prod_{i=1}^{j}\frac{\mu_l}{\gamma-\mu_l+i\alpha}, \\ E[Q] &= \pi_0\sum_{j=1}^{B} j\sum_{l=1}^{k}\prod_{i=1}^{j}\frac{\mu_l}{\gamma-\mu_l+i\alpha}. \end{aligned}$$

References

[1] Artur K. Ekert. Quantum cryptography based on Bell's theorem. *Physical Review Letters*, 67(6):661–663, aug 1991. doi: 10.1103/physrevlett.67.661.

[2] Charles H. Bennett and Gilles Brassard. Quantum cryptography: Public key distribution and coin tossing. *Theoretical Computer Science*, 560:7–11, dec 2014. doi: 10.1016/j.tcs.2014.05.025.

[3] Charles H. Bennett, Gilles Brassard, and N. David Mermin. Quantum cryptography without Bell's theorem. *Physical Review Letters*, 68(5):557–559, feb 1992. doi: 10.1103/physrevlett.68.557.

[4] Feihu Xu, Bing Qi, Zhongfa Liao, and Hoi-Kwong Lo. Long distance measurement-device-independent quantum key distribution with entangled photon sources. *Applied Physics Letters*, 103(6):061101, aug 2013. doi: 10.1063/1.4817672.

[5] Juan Yin, Yuan Cao, Yu-Huai Li, Ji-Gang Ren, Sheng-Kai Liao, Liang Zhang, Wen-Qi Cai, Wei-Yue Liu, Bo Li, Hui Dai, Ming Li, Yong-Mei Huang, Lei Deng, Li Li, Qiang Zhang, Nai-Le Liu, Yu-Ao Chen, Chao-Yang Lu, Rong Shu, Cheng-Zhi Peng, Jian-Yu Wang, and Jian-Wei Pan. Satellite-to-Ground Entanglement-Based Quantum Key Distribution. *Physical Review Letters*, 119(20), nov 2017. doi: 10.1103/physrevlett.119.200501.

[6] Anne Broadbent, Joseph Fitzsimons, and Elham Kashefi. Universal Blind Quantum Computation. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, oct 2009. doi: 10.1109/focs.2009.36.

[7] Liang Jiang, Jacob M. Taylor, Anders S. Sørensen, and Mikhail D. Lukin. Distributed quantum computation based on small quantum registers. *Physical Review A*, 76(6), dec 2007. doi: 10.1103/physreva.76.062323.

[8] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Advances in quantum metrology. *Nature Photonics*, 5(4):222–229, mar 2011. doi: 10.1038/nphoton.2011.35.

[9] Yi Xia, Wei Li, William Clark, Darlene Hart, Quntao Zhuang, and Zheshen Zhang. Demonstration of a Reconfigurable Entangled Radio-Frequency Photonic Sensor Network. *Physical Review Letters*, 124(15), apr 2020. doi: 10.1103/physrevlett.124.150502.

[10] D. Leibfried. Toward Heisenberg-Limited Spectroscopy with Multiparticle Entangled States. *Science*, 304(5676):1476–1478, jun 2004. doi: 10.1126/science.1097576.

[11] Quntao Zhuang and Zheshen Zhang. Physical-Layer Supervised Learning Assisted by an Entangled Sensor Network. *Physical Review X*, 9(4), oct 2019. doi: 10.1103/physrevx.9.041023.

[12] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem. *IEEE Transactions on Information Theory*, 48(10):2637–2655, oct 2002. doi: 10.1109/tit.2002.802612.

[13] Haowei Shi, Zheshen Zhang, and Quntao Zhuang. Practical Route to Entanglement-Assisted Communication Over Noisy Bosonic Channels. *Physical Review Applied*, 13(3), mar 2020. doi: 10.1103/physrevapplied.13.034029.

[14] Stefano Pirandola. End-to-end capacities of a quantum communication network. *Communications Physics*, 2(1), may 2019. doi: 10.1038/s42005-019-0147-3.

[15] Mihir Pant, Hari Krovi, Don Towsley, Leandros Tassiulas, Liang Jiang, Prithwish Basu, Dirk Englund, and Saikat Guha. Routing entanglement in the quantum internet. *npj Quantum Information*, 5(1), mar 2019. doi: 10.1038/s41534-019-0139-x.

[16] Axel Dahlberg, Matthew Skrzypczyk, Tim Coopmans, Leon Wubben, Filip Rozpędek, Matteo Pompili, Arian Stolk, Przemysław Pawełczak, Robert Knegjens, Julio de Oliveira Filho, Ronald Hanson, and Stephanie Wehner. A link layer protocol for quantum networks. In *Proceedings of the ACM Special Interest Group on Data Communication*, pages 159–173. ACM, aug 2019. doi: 10.1145/3341302.3342070.

[17] Rodney Van Meter. *Quantum Networking*. John Wiley & Sons, Ltd, apr 2014. doi: 10.1002/9781118648919.

[18] M. K. Bhaskar, R. Riedinger, B. Machielse, D. S. Levonian, C. T. Nguyen, E. N. Knall, H. Park, D. Englund, M. Lončar, D. D. Sukachev, and M. D. Lukin. Experimental demonstration of memory-enhanced quantum communication. *Nature*, 580(7801):60–64, mar 2020. doi: 10.1038/s41586-020-2103-5.

[19] Yuan Lee, Eric Bersin, Axel Dahlberg, Stephanie Wehner, and Dirk Englund. A quantum router architecture for high-fidelity entanglement flows in multi-user quantum networks. *arXiv preprint arXiv:2005.01852*, 2020.

[20] Ruoyu Li, Luca Petit, David P. Franke, Juan Pablo Dehollain, Jonas Helsen, Mark Steudtner, Nicole K. Thomas, Zachary R. Yoscovits, Kanwal J. Singh, Stephanie Wehner, Lieven M. K. Vandersypen, James S. Clarke, and Menno Veldhorst. A crossbar network for silicon quantum dot qubits. *Science Advances*, 4(7):eaar3960, jul 2018. doi: 10.1126/sciadv.aar3960.

[21] Seiji Armstrong, Jean-François Morizur, Jiri Janousek, Boris Hage, Nicolas Treps, Ping Koy Lam, and Hans-A. Bachor. Programmable multimode quantum networks. *Nature Communications*, 3(1), jan 2012. doi: 10.1038/ncomms2033.

[22] I. Herbauts, B. Blauensteiner, A. Poppe, T. Jennewein, and H. Hübel. Demonstration of active routing of entanglement in a multi-user network. *Optics Express*, 21(23):29013, nov 2013. doi: 10.1364/oe.21.029013.

[23] Matthew A. Hall, Joseph B. Altepeter, and Prem Kumar. Ultrafast Switching of Photonic Entanglement. *Physical Review Letters*, 106(5), feb 2011. doi: 10.1103/physrevlett.106.053901.

[24] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2009. doi: 10.1017/cbo9780511976667.

[25] Gayane Vardoyan, Saikat Guha, Philippe Nain, and Don Towsley. On the exact analysis of an idealized quantum switch. *Performance Evaluation*, 144:102141, dec 2020. doi: 10.1016/j.peva.2020.102141.

[26] Charles H. Bennett, Gilles Brassard, Sandu Popescu, Benjamin Schumacher, John A. Smolin, and William K. Wootters. Purification of Noisy Entanglement and Faithful Teleportation via Noisy Channels. *Physical Review Letters*, 76(5):722–725, jan 1996. doi: 10.1103/physrevlett.76.722.

[27] Gayane Vardoyan, Saikat Guha, Philippe Nain, and Don Towsley. On the Stochastic Analysis of a Quantum Entanglement Switch. *ACM SIGMETRICS Performance Evaluation Review*, 47(2):27–29, dec 2019. doi: 10.1145/3374888.3374899.

[28] Tim Coopmans, Robert Knegjens, Axel Dahlberg, David Maier, Loek Nijsten, Julio Oliveira, Martijn Papendrecht, Julian Rabbie, Filip Rozpędek, Matthew Skrzypczyk, et al. NetSquid, a discrete-event simulation platform for quantum networks. *arXiv preprint arXiv:2010.12535*, 2020.

[29] Stefano Pirandola, Riccardo Laurenza, Carlo Ottaviani, and Leonardo Banchi. Fundamental limits of repeaterless quantum communications. *Nature Communications*, 8(1), apr 2017. doi: 10.1038/ncomms15043.

[30] E. Shchukin, F. Schmidt, and P. van Loock. Waiting time in quantum repeaters with probabilistic entanglement swapping. *Physical Review A*, 100(3), sep 2019. doi: 10.1103/physreva.100.032322.

[31] Gayane Vardoyan, Saikat Guha, Philippe Nain, and Don Towsley. On the Capacity Region of Bipartite and Tripartite Entanglement Switching. *arXiv preprint arXiv:1901.06786*, 2019.

[32] Philippe Nain, Gayane Vardoyan, Saikat Guha, and Don Towsley. On the Analysis of a Multipartite Entanglement Distribution Switch. *ACM SIGMETRICS Performance Evaluation Review*, 48(1):49–50, jul 2020. doi: 10.1145/3410048.3410077.

[33] Saikat Guha, Hari Krovi, Christopher A. Fuchs, Zachary Dutton, Joshua A. Slater, Christoph Simon, and Wolfgang Tittel. Rate-loss analysis of an efficient quantum repeater architecture. *Physical Review A*, 92(2), aug 2015. doi: 10.1103/physreva.92.022357.

[34] Fabian Ewert and Peter van Loock. 3/4-Efficient Bell Measurement with Passive Linear Optics and Unentangled Ancillae. *Physical Review Letters*, 113(14), sep 2014. doi: 10.1103/physrevlett.113.140403.

[35] W. P. Grice. Arbitrarily complete Bell-state measurement using only linear optical elements. *Physical Review A*, 84(4), oct 2011. doi: 10.1103/physreva.84.042331.

[36] O. A. Collins, S. D. Jenkins, A. Kuzmich, and T. A. B. Kennedy. Multiplexed Memory-Insensitive Quantum Repeaters. *Physical Review Letters*, 98(6), feb 2007. doi: 10.1103/physrevlett.98.060502.

[37] Wojciech Kozlowski, Axel Dahlberg, and Stephanie Wehner. Designing a quantum network protocol. In *Proceedings of the 16th International Conference on emerging Networking EXperiments and Technologies*. ACM, nov 2020. doi: 10.1145/3386367.3431293.

[38] Sumeet Khatri, Corey T. Matyas, Aliza U. Siddiqui, and Jonathan P. Dowling. Practical figures of merit and thresholds for entanglement distribution in quantum networks. *Physical Review Research*, 1(2), sep 2019. doi: 10.1103/physrevresearch.1.023032.

[39] Boxi Li, Tim Coopmans, and David Elkouss. Efficient Optimization of Cut-offs in Quantum Repeater Chains. In *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, oct 2020. doi: 10.1109/qce49297.2020.00029.

[40] F. Rozpędek, K. Goodenough, J. Ribeiro, N. Kalb, V. Caprara Vivoli, A. Reiserer, R. Hanson, S. Wehner, and D. Elkouss. Parameter regimes for a single sequential quantum repeater. *Quantum Science and Technology*, 3(3):034002, apr 2018. doi: 10.1088/2058-9565/aab31b.

[41] Filip Rozpedek, Raja Yehia, Kenneth Goodenough, Maximilian Ruf, Peter C. Humphreys, Ronald Hanson, Stephanie Wehner, and David Elkouss. Near-term quantum-repeater experiments with nitrogen-vacancy centers: Overcoming the limitations of direct transmission. *Physical Review A*, 99(5), may 2019. doi: 10.1103/physreva.99.052330.

[42] D. Gottesman and Hoi-Kwong Lo. Proof of security of quantum key distribution with two-way classical communications. *IEEE Transactions on Information Theory*, 49(2):457–475, feb 2003. doi: 10.1109/tit.2002.807289.

[43] Leonard Kleinrock. *Queueing Systems, Volume I: Theory*. Wiley New York, 1975.