2022 IEEE International Conference on Quantum Computing and Engineering (QCE) | 978-1-6654-9113-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/QCE53715.2022.00060

The Capacity Region of Entanglement Switching: Stability and Zero Latency

Wenhan Dai[1,2], Anthony Rinaldi[3] and Don Towsley[1]
[1]College of Information and Computer Science, University of Massachusetts Amherst
[2] Quantum Photonics Laboratory, Massachusetts Institute of Technology
[3]University of Massachusetts Amherst
Email: whdai@cs.umass.edu, aarinaldi@umass.edu, and towsley@cs.umass.edu

***Abstract*—Quantum switches distribute entangled pairs among end nodes by entanglement swapping and are critical components in quantum networks. In this work, we design protocols that schedule entanglement swapping in quantum switches. In contrast to most existing studies, we consider that entanglement requests randomly arrive at the switch, and determine the capacity region of rate vectors that the switch can support stably. For a rate vector inside the capacity region, we develop protocols that not only stabilize the switch, but also achieve zero average latency. Among these protocols, the on-demand protocols are computationally efficient and achieve high fidelity and low latency demonstrated by results obtained using a quantum network discrete event simulator.**

***Index Terms*—quantum switch, entanglement distribution, quantum networking**

I. Introduction

Quantum networks will play a critical role in enabling numerous quantum applications such as quantum key distribution [1]–[4], teleportation [5]–[7], and quantum sensing [8]–[10]. One of the major tasks of quantum networks is distributing quantum entanglement among geographically separated nodes. Such a task usually involves generating Einstein-Podolsky-Rosen (EPR) pairs through quantum channels and then performing entanglement swapping among the generated EPR pairs. For example, consider a star-shape network consisting of a center node and a collection of end nodes. Entanglement swapping is performed at the center node to establish entanglement among end nodes. The center node serves as a quantum switch, a critical building block in quantum networks. See Figure 1 for details.

A key problem in the implementation of a quantum switch is decision-making about which EPR pairs to perform entanglement swapping operations on. The prioritization of entanglement swapping affects the performance of the switch, such as the fidelity of the distributed entanglement, the latency of the entanglement requests, and the throughput of the switch. Existing studies on entanglement swapping generally focus on maximizing entanglement generation rate, and the quantum network establishes entanglement whenever possible. Relevant work is summarized in Section III. In this manuscript, we consider a more generic and practical scenario where entanglement requests randomly arrive at the switch, and the switch aims to address these requests. Instead of maximizing entanglement generation rate, we tackle seemingly more difficult problems that involve the concept of stability. Roughly speaking, a quantum switch being stable implies that the number of unaddressed entanglement requests is not very large with a high probability. Note that a protocol that maximizes the entanglement generation rate may not stabilize the switch since the number of unaddressed entanglement requests may grow to infinity sub-linearly with respect to time. The goal of this manuscript is to solve the following two problems: What is the capacity region of the switch? Is there an entanglement swapping protocol that can stabilize the switch for any workload vector within this capacity region?

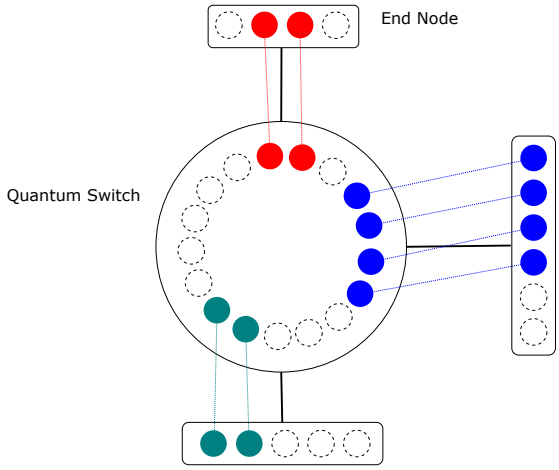

Fig. 1: Illustration for a quantum switch. The big circle represents the switch and the rectangles represent end nodes. Solid colorful dots represent entangled qubits, and empty dots represent empty memory slots in the quantum switch or end nodes. Different colors correspond to different end nodes. The lines connecting the switch and end nodes represent quantum channels. A colorful dashed line implies that the two colorful dots connected by the dashed line consist of an EPR pair.

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/QCE53715.2022.00060

Note that in classical data networks, stability is widely used to investigate protocols for data routing and resource allocation [11]–[13]. These studies cannot be directly applied to quantum networks because entanglement swapping involve two interfaces, whereas data transmission in classical data networks normally involves one link. Moreover, entanglement generated through quantum channels can be stored to satisfy future entanglement requests whereas in classical networks, data have to arrive first and then get transmitted through channels. Though these fundamental differences precludes our directly applying existing results in classical networks, the mathematical tools, such as Lyapunov drift analysis, used to obtain them are useful for the analysis of the entanglement swapping protocols in a quantum switch. With different Lyapunov functions tailored for the quantum switch, we can develop several protocols that can stabilize the switch. The key contributions of this manuscript are as follows:

- We determine the capacity region for the entanglement rates. In particular, we show that if the entanglement rates are outside of this region, no entanglement swapping protocol can stabilize the switch.
- For any entanglement rates being an interior point of the capacity region, we develop stationary protocols that stabilize the switch.
- We develop on-demand protocols that stabilize the switch. These protocols are computationally efficient and do not require statistical knowledge of the entanglement requests and the quantum channels.
- We further show that the stationary protocol and the on-demand protocol with a small modification achieve zero average latency. This means that almost all requests are served immediately when they arrive at the switch.
- We evaluate the proposed protocols with a quantum network discrete event simulator. We compare the protocols according to the fidelity and latency. The on-demand protocol is computationally efficient and achieves high fidelity and low latency demonstrated by numerical results.

II. Background

In this section, we provide some background information used in this manuscript. An EPR pair is a quantum state consisting of two qubits:

$$|\Psi_{AB}\rangle = \frac{1}{\sqrt{2}}(|0_A\rangle\,|0_B\rangle + |1_A\rangle\,|1_B\rangle)$$

where $|0\rangle$ and $|1\rangle$ are qubits represented by two-dimensional vectors, and the subscripts A and B represent two physics systems.

One can use entanglement swapping at an intermediate party C to generate an EPR pair between two parties A and B [14]. In our setup, the quantum switch generates EPR pairs with end nodes through quantum channels, and performs entanglement swapping to generate EPR pairs between end nodes.

Generating EPR pairs between the quantum switch and end nodes requires qubit transmission through quantum channels. One of the widely used mediums for qubit transmission is optical fiber, and correspondingly, the quantum information in qubits are carried by photons. For a single photon that goes through optical fiber, with probability p this photon successfully reaches the receiver, and with probability $1-p$ it is lost. Note that we consider heralded entanglement generation, i.e., the results of entanglement generation, either success or failure, are known to the switch.

Entanglement swapping requires making Bell state measurements on two qubits at the quantum switch. This can be done by performing a CNOT operation and making measurements with standard computation basis. In practice, CNOT operations are not always successful when implemented using linear optics [15]–[17] or photon-spin interaction [18]. Therefore, we model the Bell state measurement as a probabilistic operator: with probability q, it succeeds and the EPR pair between the corresponding two end nodes is generated; with probability $1-q$, no EPR pair is generated between the two end nodes although the two EPR pairs between the quantum switch and the two end nodes are consumed.

III. Related Work

Quantum switches are important components of quantum networks, and have attracted increasing research interest [19]–[22]. In [19], a quantum switch that serves multipartite entanglement to a set of end nodes is analyzed. In [20], a similar setup that focuses on bipartite entanglement distribution is considered. Compared to [19], the model of the quantum switch in [20] is more general, accounting for decoherence of quantum states in the memory and finite memory size. The setup in this manuscript is significantly different from these studies in three aspects. First, entanglement generation between the quantum switch and end nodes in these studies is formulated as a continuous-time Markov chain, i.e., at each time slot, one and only one of entanglement pair is generated through the quantum channels. In this manuscript, instead of the continuous-time Markov chain, we adopt the discrete-time Markov chain, which is shown to be much more challenging for analysis [21]. Second, [19]–[22] implicitly assume that the number of entanglement requests for every pair of users is infinite at any time slot, and when the quantum switch performs entanglement swapping or GHZ projection successfully, the generated bipartite or tripartite entanglement is immediately released from memory to address the entanglement requests. In this manuscript, the entanglement requests randomly arrive at the switch according to a stochastic process model. Correspondingly, the definition of stability is different from [19], and we focus on the unaddressed entanglement requests at the quantum switch. Third, one of the contributions of this manuscript is the design of entanglement swapping protocols, whereas in [19]–[22], the operations of the quantum switch are relatively simple. The reason for such differences is the introduction of entanglement requests, and the objective of the switch is to address these requests instead of maximizing the entanglement switching rate.

Entanglement swapping protocols are proposed for networks with other structures than the star-shaped ones. In a recent paper [23], entanglement distribution for a network consisting of quantum switches and users is considered. Similarly to this

manuscript, entanglement requests are considered. However, the setup in this manuscript is significantly different that in [23]. First, the lifetime of a qubit is assumed to be one cycle in [23], but the lifetime is assumed to be infinite in this manuscript. Second, the protocols proposed in [23] have high computational complexities, whereas most of the protocols, especially the on-demand protocols, proposed in this manuscript are efficient. Third, we show the proposed protocols can achieve zero average latency in addition to stability. Another example of entanglement swapping protocols is [24], where the authors propose an approach to calculate the average waiting time for generating an entangled pair in quantum repeater chains. Similarly to the studies on the quantum switch, this work aims at maximizing entanglement generating rate rather than addressing entanglement requests between end nodes.

IV. System Model

Consider a star-shape network consisting of $K+1$ nodes, where node 0 is a quantum switch and the rest are end nodes. The quantum switch has K interfaces that serve EPR pairs, where interface k serves EPR pairs between the switch and node k, $k \in \mathcal{K} = \{1, 2, \ldots, K\}$. Time is slotted and at each time slot t, three types of events may occur, described as follows.

Entanglement Generation: The quantum switch attempts to generate EPR pairs with end nodes. An EPR pair between the quantum switch and node k is generated with probability p_k, $k \in \mathcal{K}$ using a quantum channel. One qubit of each EPR pair is stored at the switch and the other at the end node. Let $C_{0i}(t)$ denotes the number of EPR pairs $|\Psi_{0i}\rangle$ generated between the quantum switch and node $i \in \mathcal{K}$ at time slot t, and we assume that $\{C_{0i}(t) : t \geq 0\}$, $i \in \mathcal{K}$ are mutually independent Bernoulli processes.[1]

Entanglement Swapping: The quantum switch performs entanglement swapping operations. In particular, an EPR pair $|\Psi_{ij}\rangle$ is created with probability q by consuming two EPR pairs, $|\Psi_{0i}\rangle$ and $|\Psi_{0j}\rangle$.[2]

Entanglement Request: During time slot t, entanglement requests randomly arrive at the switch, and the quantum switch maintains a queue for storing entanglement requests. Let $A_{ij}(t)$ denote the number of entanglement requests between nodes k and j at time slot t, and we assume that $\{A_{ij}(t) : t \geq 0\}$ are mutually independent sequences of random variables. For the entanglement requests $\{A_{ij}(t) : t \geq 0\}$, we assume it is a stationary and ergodic process with rates λ_{ij}.

A. System Dynamics

We now describe the variables and evolution of the switch. Let $E_{ij}(t)$ denote the number of EPR pairs $|\Psi_{ij}\rangle$ stored in nodes $i, j \in \mathcal{K}$ at time $t \geq 0$. Let $U_{ij}(t)$ denote the number of pending entanglement requests for $|\Psi_{ij}\rangle$ at time $t \geq 0$, $i, j \in \mathcal{K}$. At each time slot, the quantum switch makes decisions about what link EPR pairs to perform entanglement swapping operations on. In particular, the quantum switch attempts to create entanglement $|\Psi_{ij}\rangle$ by consuming $F_{ij}(t)$ pairs of $|\Psi_{0i}\rangle$ and $|\Psi_{0j}\rangle$ from the stored entanglement in the quantum switch, $i, j \in \mathcal{K}$.[3] The following constraints need to be satisfied:

$$\sum_{i \in \mathcal{K}} F_{ij}(t) \leq E_{0j}(t) + C_{0j}(t), \quad \forall j \in \mathcal{K}; t \geq 0.$$

Moreover, we consider that at each time slot, the quantum switch can perform at most W entanglement swapping operations, leading to the following constraint:

$$\sum_{i,j \in \mathcal{K}} F_{ij}(t) \leq W; t \geq 0.$$

Outcomes of entanglement swapping operations are independent events, each succeeding with probability q. Let $\mu_{ij}(t)$ denote the number of successfully generated pairs $|\Psi_{ij}\rangle$, $i, j \in \mathcal{K}$. Let $[x]^+ = \max\{x, 0\}$. We assume entanglement swapping is performed at the beginning of each time slot, whereas entanglement requests may arrive at any time during a time slot. Then, $U_{ij}(t)$ and $E_{0i}(t)$ evolve as follows:

$$U_{ij}(t+1) = [U_{ij}(t) + A_{ij}(t) - E_{ij}(t) - \mu_{ij}(t)]^+, \quad \forall i, j \in \mathcal{K}$$
$$E_{ij}(t+1) = [E_{ij}(t) + \mu_{ij}(t) - U_{ij}(t) - A_{ij}(t)]^+, \quad \forall i, j \in \mathcal{K}$$
$$E_{0i}(t+1) = E_{0i}(t) - \sum_{j \in \mathcal{K}} F_{ij}(t) + C_{0i}(t), \quad \forall i \in \mathcal{K}.$$

Without loss of generality, we assume $U_{ij}(0) = E_{ij}(0) = 0$, $i, j \in \mathcal{K}$. Since no entanglement swapping is performed at time 0, we have $\mu_{ij}(0) = 0$.

With the introduction of the system dynamics, we further assume that the second moment of entanglement requests is bounded at every time slot, regardless of history, i.e.,

$$\mathbb{E}\big[A_{ij}^2(t) | \boldsymbol{H}(t) = \boldsymbol{h}\big] \leq A_{\max}^2, \quad \forall i, j \in \mathcal{K}. \tag{1}$$

where $\boldsymbol{H}(t)$ denotes the history of system states.

The goal of the quantum switch is to maintain as small a queue backlog $U_{ij}(t)$ for user pairs i, j as possible and to stabilize the system. For now, we assume that there is no limit to the number of EPR pairs that can be stored in the nodes. Moreover, we assume that a qubit never decoheres. These assumptions will be relaxed in a later section of this manuscript.

B. Stability and Capacity Region

We follow the definition of stability in classical networks [12]. Consider the following function:

$$g_{ij}(V) := \limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{P}[\, U_{ij}(\tau) > V \,], \quad \forall i, j \in \mathcal{K}. \tag{2}$$

This function characterizes the fraction of time that the number of unfinished requests for EPR pairs $|\Psi_{ij}\rangle$ exceeds a certain value V.

[1] If multiplexing techniques can be used, $\{C_{0i}(t)\}$ can be modelled as Binomial random variables. Results in this work can be easily generalized to accommodate multiplexing techniques.

[2] The entanglement swapping probability does not need to be the same for all node pairs. Results in this work can be easily generalized to accommodate different entanglement swapping probabilities among node pairs.

[3] We do not distinguish between the order of nodes i and j, i.e., $F_{ij}(t) = F_{ji}(t)$, $\forall i, j \in \mathcal{K}$.

Notation	Definition
K	Number of end nodes
$\mathcal{K}$	Set of end nodes
p_k	Success probability of generating an EPR pair between the switch and node k
$\|\Psi_{ij}\rangle$	EPR pair between node i and j
q	Success probability of entanglement swapping
$A_{ij}(t)$	Number of entanglement requests between nodes i and j at time t
$E_{ij}(t)$	Number of $\|\Psi_{ij}\rangle$ stored in nodes i and j at time t
$E_{0j}(t)$	Number of $\|\Psi_{0j}\rangle$ stored in the switch and node j at time t
$U_{ij}(t)$	Number of requests for $\|\Psi_{ij}\rangle$ at time t
$F_{ij}(t)$	Number of $\|\Psi_{0i}\rangle$ and $\|\Psi_{0j}\rangle$ consumed to create $\|\Psi_{ij}\rangle$ at time t
W	Maximum number of entanglement swaps per time slot
$\mu_{ij}(t)$	Number of successfully generated $\|\Psi_{ij}\rangle$ at time t via entanglement swapping
C_{0i}(t)	Number of successfully generated $\|\Psi_{0i}\rangle$ at time t via the quantum channel
λ_{ij}	Rate of entanglement request $A_{ij}(t)$
$\boldsymbol{H}(t)$	All the history information at time t
$A_{\max}$	Upper bound for the second moment of $A_{ij}(t)$ defined in (1)
$\boldsymbol{\Lambda}$	Capacity region
$\mathcal{M}_{ij}^{i}(t)$	Set of $\|\Psi_{0i}\rangle$ labelled as (i,j) that are not consumed up to time t in π_{stat}

TABLE I: Notations of Important Quantities.

Definition 1: The quantum switch is stable if $g_{ij}(V) \to 0$ as $V \to \infty$, for all $i, j \in \mathcal{K}$.

We can then define the capacity region for the quantum switch system as follows.

Definition 2: The capacity region for the quantum switch is the closure of the set of matrices $(\lambda_{ij})_{i,j\in\mathcal{K}}$ such that there exists an entanglement swapping algorithm that stabilizes the switch.

Theorem 1 (Capacity Region): With given parameters p_k, $k \in \mathcal{K}$, q, and W, the capacity region $\boldsymbol{\Lambda}$ is the set of all matrices $(\lambda_{ij})_{i,j\in\mathcal{K}}$ for which there exist non-negative variables $\{f_{ij}\}_{i,j\in\mathcal{K}}$ satisfying:[4]

$$\sum_{i\in\mathcal{K}} f_{ij} \le p_j, \quad \forall j \in \mathcal{K} \tag{3}$$

$$\sum_{i,j\in\mathcal{K}} f_{ij} \le W \tag{4}$$

$$\lambda_{ij} \le q\, f_{ij}, \quad \forall i, j \in \mathcal{K}. \tag{5}$$

We skip the proof here due to space constraints and refer to the full version [25] for the detailed proof.

V. Protocol Design

In this section, we design several protocols, namely, the stationary and the on-demand protocol, for entanglement swapping in a quantum switch. More importantly, we show that these protocols stabilize the switch when the rate matrix is within the capacity region. For simplicity, we assume that the quantum switch can perform an arbitrary number of entanglement swapping operations (i.e., $W = \infty$), but all of the protocols can be easily modified to satisfy the constraint on the number of entanglement swapping operations per time slot. The stationary protocol requires knowing the rate matrix and channel parameters, and does not depend on the states of the system. The on-demand protocols do not require any knowledge of rate matrix or channel parameters. Though they need to solve an optimization problem every time slot, this optimization problem can be efficiently solved.

[4]We do not distinguish between the order of nodes i and j, i.e., $f_{ij} = f_{ji}$, $\forall i, j \in \mathcal{K}$.

A. Stationary Protocol

We first design a stationary protocol π_{stat} as follows.

Stationary Protocol: Suppose the rate matrix $(\lambda_{ij})_{i,j\in\mathcal{K}}$ is known to the quantum switch and there exists $\epsilon > 0$ such that

$$(\lambda_{ij} + \epsilon)_{i,j\in\mathcal{K}} \in \boldsymbol{\Lambda}. \tag{6}$$

Then there exists a set of variables $\{\tilde{f}_{ij}\}_{i,j\in\mathcal{K}}$ such that

$$\sum_{i\in\mathcal{K}} \tilde{f}_{ij} \le p_j, \quad \forall j \in \mathcal{K} \tag{7}$$

$$\lambda_{ij} + \epsilon \le q\, \tilde{f}_{ij}, \quad \forall i, j \in \mathcal{K}. \tag{8}$$

In fact, determining $\{\tilde{f}_{ij}\}_{i,j\in\mathcal{K}}$ is straightforward since we can set $\tilde{f}_{ij} = (\lambda_{ij} + \epsilon)/q$ and the conditions (7) and (8) hold if $(\lambda_{ij} + \epsilon)_{i,j\in\mathcal{K}} \in \boldsymbol{\Lambda}$. For an EPR pair $|\Psi_{0i}\rangle$ generated at time slot t, $i \in \mathcal{K}$, label it as (i,j) with probability $\tilde{f}_{ij}/p_i$. Let $\mathcal{M}_{ij}^{i}(t)$ denote the set of EPR pairs $|\Psi_{0i}\rangle$ labelled as (i,j) that are not consumed for entanglement swapping up to time slot t. If $\mathcal{M}_{ij}^{i}(t) \neq \varnothing$ and $\mathcal{M}_{ij}^{j}(t) \neq \varnothing$, the quantum switch performs entanglement swaps to create EPR pair $|\Psi_{ij}\rangle$ by consuming EPR pairs $|\Psi_{0i}\rangle$ and $|\Psi_{0j}\rangle$ until either $\mathcal{M}_{ij}^{i}(t)$ or $\mathcal{M}_{ij}^{j}(t)$ is empty.

Theorem 2: The quantum switch is stable using the stationary discard protocol π_{stat} if $(\lambda_{ij})_{i,j\in\mathcal{K}}$ is an interior point in $\boldsymbol{\Lambda}$, i.e., (6) holds for some $\epsilon > 0$.

We skip the proof here due to space constraints and refer to the full version [25] for the detailed proof.

Though the stationary protocol can stabilize the switch, it requires knowing not only the parameters p_k, $k \in \mathcal{K}$, but also the rate matrix $(\lambda_{ij})_{i,j\in\mathcal{K}}$ in order to find $\{\tilde{f}_{ij}\}_{i,j\in\mathcal{K}}$. Such knowledge of all the rates may not be available in practice, and we may need protocols that requires less information. This motivates the design of on-demand protocols.

B. On-demand Protocol

We next develop an on-demand protocol that has the flexibility to prioritize entanglement requests as long as they satisfy certain constraints.

On-demand Protocols: At each time slot, the quantum switch attempts to create entanglement $|\Psi_{ij}\rangle$ using F_{ij} pairs of entanglement $|\Psi_{0i}\rangle$ and F_{ij} pairs of entanglement $|\Psi_{0j}\rangle$. The decisions $\{F_{ij}\}_{i,j\in\mathcal{K}}$ need to satisfy the following constraints:

$$\sum_{i\in\mathcal{K}} F_{ij} \le E_{0j}(t) + C_{0j}(t), \quad j \in \mathcal{K} \tag{9}$$

$$F_{ij} \le U_{ij}(t) + A_{ij}(t), \quad i,j \in \mathcal{K} \tag{10}$$

$$F_{ij} = F_{ji} \in \mathbb{N}, \quad i,j \in \mathcal{K}$$

$$\big(E_{0i}(t) + C_{0i}(t) - \sum_{k\in\mathcal{K}} F_{ik}\big)\big(E_{0j}(t) + C_{0j}(t) - \sum_{k\in\mathcal{K}} F_{kj}\big) \cdot \big(U_{ij}(t) + A_{ij}(t) - F_{ij}\big) = 0, \quad i,j \in \mathcal{K}. \tag{11}$$

Then the quantum switch attempts to create entanglement $|\Psi_{ij}\rangle$ using F_{ij} pairs of entanglement $|\Psi_{0i}\rangle$ and F_{ij} pairs of entanglement $|\Psi_{0j}\rangle$. An on-demand protocol is denoted by π_{od}.

Remark 1: The intuition of (10) in on-demand protocols is that the quantum switch creates entanglement only to serve existing requests. Moreover, the quantum switch should not "waste" any opportunities to create entanglement, in the sense that for any i,j, it should attempt to create as many pairs of $|\Psi_{ij}\rangle$ as possible provided that (10) holds. This gives the condition (11). Note that there may be multiple choices of $\{F_{ij}\}_{i,j\in\mathcal{K}}$ that satisfy the constraints. The quantum switch can select any one of them for entanglement swapping.

Remark 2: One way to satisfy constraints (9)-(11) is to first set $F_{ij} = 0$, $i,j \in \mathcal{K}$ and then check the unfinished entanglement requests $\{U_{ij}(t)\}_{i,j\in\mathcal{K}}$ in any order. Details are given in Algorithm 1. Note that the complexity is $O(K^2)$ since the iteration is over $i,j \in \mathcal{K}$. Comparatively, the stationary protocol also has a complexity $O(K^2)$ in average since the switch has to randomly label every EPR generated between the switch and nodes. In practice the complexity is much lower. In fact, as shown later in Appendix I, $F_{ij}(t) = U_{ij}(t) + A_{ij}(t)$ with a high probability, and there is no need to solve any optimization problem.

In order to establish stability of the on-demand protocols, we need to make the following assumption with on-demand protocols.

Algorithm 1 On-demand Protocol

Input: $E_{0j}(t)$, $C_{0j}(t)$, $\forall j \in \mathcal{K}$, $U_{ij}(t)$, $A_{ij}(t)$, $\forall i,j \in \mathcal{K}$
Output: F_{ij} $\forall i,j \in \mathcal{K}$ that satisfy (9) to (11)
1: Initialization: $F_{ij} = 0$ $\forall i,j \in \mathcal{K}$
2: **for** $i,j \in \mathcal{K}$ **do**
3: $F_{ij} \leftarrow \min\{U_{ij}(t) + A_{ij}(t), E_{0i}(t) + C_{0i}(t) - \sum_{k\in\mathcal{K},k\neq i} F_{kj}, E_{0j}(t) + C_{0j}(t) - \sum_{k\in\mathcal{K},k\neq i} F_{kj}\}$
4: **end for**

Assumption 1: For an arbitrary $\epsilon_0 > 0$, consider the following event:

$$B_0 : \exists i,j, \ \left|\frac{1}{t+1}\sum_{\tau=0}^{t} A_{ij}(\tau) - \lambda_{ij}\right| > \epsilon_0. \tag{12}$$

There exists a $c_1(\epsilon_0) < \infty$ independent of t such that $\{A_{ij}(t)\}_{i,j\in\mathcal{K}}$ satisfies

$$\mathbb{E}\Big[\sum_{\tau=0}^{t} A_{ij}(\tau)\Big|B_0\Big]\mathbb{P}[B_0] \le c_1(\epsilon_0). \tag{13}$$

Note that Assumption 1 holds for many random processes. For example, if $A_{ij}(t)$ is i.i.d over time and $\mathbb{V}\text{ar}\Big[A_{ij}(t)\Big] = \sigma^2$, where σ is a constant irrelevant of t, then one can use Chebyshev's inequality to verify that Assumption 1 holds.

Theorem 3: The quantum switch system is stable using an on-demand protocol π_{od} if $(\lambda_{ij})_{i,j\in\mathcal{K}}$ is an interior point in $\boldsymbol{\Lambda}$ under Assumption 1.

The proof is found in Appendix I.

VI. Zero Average Latency

In previous sections, we showed that the stationary and on-demand protocols stabilize the switch provided the rate matrix is an interior point of the capacity region. In this section, we take a step further and show that with slight modification, these protocols achieve zero average latency.

It suffices to show for all $i,j \in \mathcal{K}$

$$\lim_{t\to\infty} \mathbb{E}[U_{ij}(t)] = 0. \tag{14}$$

Therefore, we only need to show the expected queue length $\mathbb{E}[U_{ij}(t)]$ converges to 0. This coupled with Little's law [26] allows us to conclude that average latency is zero. To do this, we need a stronger assumption than Assumption 1.

Assumption 2: For an arbitrary $\epsilon_0 > 0$, consider the following event:

$$C_0 : \exists i,j, \ \frac{1}{t+1}\sum_{\tau=0}^{t} A_{ij}(\tau) > \lambda_{ij} + \epsilon_0. \tag{15}$$

Then entanglement requests $\{A_{ij}(t)\}_{i,j\in\mathcal{K}}$ satisfies

$$\lim_{t\to\infty} \mathbb{E}\Big[\sum_{\tau=0}^{t} A_{ij}(\tau)\Big|C_0\Big]\mathbb{P}[C_0] = 0. \tag{16}$$

Note that though stronger than Assumption 1, Assumption 2 still holds for many random processes. For example, if $A_{ij}(t)$ is i.i.d over time, its support has an upper bound, and

$$\mathbb{P}\left[\frac{1}{t+1}\sum_{\tau=0}^{t} A_{ij}(t) > \lambda_{ij} + \epsilon_0\right] \sim o(1/t).$$

Assumption 2 allows us to show that the stationary protocol achieves zero average latency.

Theorem 4: The quantum switch system achieves zero average latency using the stationary protocol if $(\lambda_{ij})_{i,j\in\mathcal{K}}$ is an interior point in $\boldsymbol{\Lambda}$ under Assumption 2.

Proof: For ϵ in (6), define events:

$$\widetilde{C}_0 : \frac{1}{t}\sum_{\tau=0}^{t-1} A_{ij}(\tau) > \lambda_{ij} + \epsilon/2 \tag{17}$$

$$\widetilde{C}_1 : \frac{1}{t}\sum_{\tau=0}^{t-1} \mu_{ij}(\tau) \leq \lambda_{ij} + \epsilon/2. \tag{18}$$

Note that if neither $\widetilde{C}_0$ nor $\widetilde{C}_1$ occurs, then $\Big(\sum_{\tau=0}^{t-1} A_{ij}(\tau) - \sum_{\tau=0}^{t-1}\mu_{ij}(\tau)\Big)^+ = 0$. As a consequence,

$$\begin{aligned}\mathbb{E}\big[U_{ij}(t)\big] \leq\ & \mathbb{E}\Big[\sum_{\tau=0}^{t-1} A_{ij}(\tau)|\widetilde{C}_0\Big]\mathbb{P}[\widetilde{C}_0] \\ & + \mathbb{E}\Big[\sum_{\tau=0}^{t-1} A_{ij}(\tau)|\widetilde{C}_1\Big]\mathbb{P}[\widetilde{C}_1].\end{aligned} \tag{19}$$

The first term in (19) converges to zero due to Assumption 2. Regarding the second term in (19),

$$\mathbb{E}\Big[\sum_{\tau=0}^{t-1} A_{ij}(\tau)|\widetilde{C}_1\Big]\mathbb{P}[\widetilde{C}_1] = \mathbb{E}\Big[\sum_{\tau=0}^{t-1} A_{ij}(\tau)\Big]\mathbb{P}[\widetilde{C}_1] = \lambda_{ij} t \mathbb{P}[\widetilde{C}_1].$$

Using Chernoff bounds on $\sum_{\tau=0}^{t-1} F_{ij}(\tau)$, $\sum_{\tau=0}^{t-1} X_{ij}^{(i)}$, and $\sum_{\tau=0}^{t-1} X_{ij}^{(j)}$, one can easily verify that $\mathbb{P}[\widetilde{C}_1]$ decays exponentially with respect to t. Therefore, (19) converges to zero as t goes to infinity. This concludes the proof. □

As mentioned in the previous section, the stationary protocol requires knowing the rate matrix $(\lambda_{ij})_{i,j\in\mathcal{K}}$, which may not be available in practice. To address this issue, we develop on-demand protocols with *virtual requests* that provide for zero latency. They allow the switch to create and store end-to-end entanglement that can be used to some future requests, so latency can be made arbitrarily small. In particular, select $\alpha \in (1/2, 1)$ and define

$$\widetilde{A}_{ij}(t) = A_{ij} + \lceil (t+1)^\alpha \rceil - \lceil t^\alpha \rceil \tag{20}$$

where the term $\lceil (t+1)^\alpha \rceil - \lceil t^\alpha \rceil$ is the virtual request. In this section, we also define

$$\widetilde{U}_{ij}(t) = U_{ij}(t) - E_{ij}(t) + \lceil t^\alpha \rceil. \tag{21}$$

Then one can verify that

$$\begin{aligned}\widetilde{U}_{ij}(t+1) &= \widetilde{U}_{ij}(t) + \widetilde{A}_{ij}(t) - \mu_{ij}(t) \\ U_{ij}(t) &= \big[\widetilde{U}_{ij}(t) - \lceil t^\alpha \rceil\big]^+.\end{aligned}$$

On-demand Protocols with Virtual Requests: At each time slot, the quantum switch attempts to create entanglement $|\Psi_{ij}\rangle$ using F_{ij} pairs of entanglement $|\Psi_{0i}\rangle$ and F_{ij} pairs of entanglement $|\Psi_{0j}\rangle$. The decisions $\{F_{ij}\}_{i,j\in\mathcal{K}}$ need to satisfy the following constraints:

$$\sum_{i\in\mathcal{K}} F_{ij} \leq E_{0j}(t) + C_{0j}(t), \quad j \in \mathcal{K} \tag{22}$$

$$F_{ij} \leq \Big\lceil \big[\widetilde{U}_{ij}(t) + \widetilde{A}_{ij}(t)\big]^+/q \Big\rceil, \quad i,j \in \mathcal{K} \tag{23}$$

$$\begin{aligned}&F_{ij} = F_{ji} \in \mathbb{N}, \quad i,j \in \mathcal{K} \\ &\Big(E_{0i}(t) + C_{0i}(t) - \sum_{k\in\mathcal{K}} F_{ik}\Big)\Big(E_{0j}(t) + C_{0j}(t) - \sum_{k\in\mathcal{K}} F_{kj}\Big) \\ &\cdot \left[\left\lceil \big[\widetilde{U}_{ij}(t) + \widetilde{A}_{ij}(t)\big)/q\big]^+ \right\rceil - F_{ij}\right] = 0, \quad i,j \in \mathcal{K}.\end{aligned} \tag{24}$$

Details are given in Algorithm 2. Compared with constraints (9) to (11), there are two main differences. The first is the presence of virtual requests in $\widetilde{A}_{ij}(t)$, and the second is the factor $1/q$ in (23) and (24). With this factor, the expected value $\mu_{ij}(t)$ is almost the same as $\big(\widetilde{U}_{ij}(t) + \widetilde{A}_{ij}(t)\big)/q$ provided that there is sufficient entanglement $|\Psi_{0i}\rangle$ and $|\Psi_{0j}\rangle$ for swapping. With these two modifications, we can show that an on-demand protocol with virtual requests, denoted by $\widetilde{\pi}_{\rm od}$, can achieve zero average latency.

Algorithm 2 On-demand Protocol with Virtual Requests

Input: $E_{0j}(t)$, $C_{0j}(t)$, $\forall j \in \mathcal{K}$, $E_{ij}(t)$, $U_{ij}(t)$, $A_{ij}(t)$, $\forall i,j \in \mathcal{K}$

Output: F_{ij} $\forall i,j \in \mathcal{K}$ that satisfy (22) to (24)

1: Initialization: $F_{ij} = 0$ $\forall i,j \in \mathcal{K}$
2: Determine the requests (real and virtual ones) $\widetilde{A}_{ij}(t)$ from $A_{ij}(t)$ as in (20), $\forall i,j \in \mathcal{K}$
3: Determine the unfinished requests with virtual requests $\widetilde{U}_{ij}(t)$ from $U_{ij}(t)$ as in (21), $\forall i,j \in \mathcal{K}$
4: **for** $i,j \in \mathcal{K}$ **do**
5: $\quad F_{ij} \leftarrow \min\Big\{\big[\widetilde{U}_{ij}(t) + \widetilde{A}_{ij}(t)\big)/q\big]^+, E_{0i}(t) + C_{0i}(t) - \sum_{k\in\mathcal{K},k\neq i} F_{kj}, E_{0j}(t) + C_{0j}(t) - \sum_{k\in\mathcal{K},k\neq i} F_{kj}\Big\}$
6: **end for**

Theorem 5: The quantum switch system achieves zero average latency using an on-demand protocol with virtual requests $\widetilde{\pi}_{\rm od}$ if $(\lambda_{ij})_{i,j\in\mathcal{K}}$ is an interior point in $\boldsymbol{\Lambda}$ under Assumption 2.

The proof is found in Appendix II.

VII. Numerical Results

In this section, we investigate the performance of the stationary and the on-demand protocols with a quantum network discrete event simulator, NetSquid [27]. Since the stability of these protocols are proven in previous sections, we focus on other performance metrics in practical scenarios. Moreover, we relax several assumptions including the infinite amount of memory for storing EPR pairs and the infinite qubit lifetime.

A. Simulation Setting

A quantum switch can be implemented with different technologies. For example, qubits stored in quantum memories can be realized with electron spins of SiV defect centers [28]. A quantum switch is equipped with photon sources, and each of

them generates a pair of entangled photons in each time slot. One of the photons interacts with the electron spin, and the other photon is sent to one of the end nodes through a quantum channel and interacts with the electron spin at the end node. A photon-spin interaction is essentially a CNOT operation on the photon and the spin. After the photon-spin interaction, photons are measured in the X basis by beamsplitters and photon detectors. Entanglement swapping in the quantum switch can also be realized with photon-spin interactions. To be consistent with practical devices, we assume that qubits stored in the memory suffer from decoherence. Furthermore, we assume that the quantum switch has a finite number of memory slots, and these memory slots are equally distributed among interfaces. With these practical constraints, we can evaluate more performance metrics:

- Average fidelity: for a state ρ shared between node i and j, the fidelity is defined as $F(\rho) = \langle \Psi_{ij} | \rho | \Psi_{ij} \rangle$. Note that a generated EPR pair may decohere in the memory slots before being used to serve entanglement requests.
- Average latency: the latency of an entanglement request is defined as the amount of time to address the entanglement request. Note that the entanglement request may occur at any time in a slot, and we evaluate the latency in units of nanoseconds rather than slots.

The developed protocols do not specify how to prioritize the EPR pairs and unserved requests, but such prioritization impacts fidelity and latency. In this section, we apply the First-In-First-Out (FIFO) method to process entanglement requests, i.e., the oldest request is the first to serve. Regarding the order of EPR pairs, we consider two methods: Oldest-Qubit-First (OQF) and youngest-Qubit-First (YQF), which use the oldest and youngest qubit for entanglement swapping, respectively. All the protocols discard EPR pairs when their fidelities fall below a preset threshold.

Each time slot is 1 μs long. The number of entanglement requests between any two end nodes i and j in a time slot is given by a mixture of two Poisson distributions with different rates. Specifically, $A_{ij}(t)$ is i.i.d over time and given by

$$A_{ij}(t) = Z \cdot Y_1 + (1 - Z) \cdot Y_2$$

where Z is a Bernoulli random varable with mean $1/2$, and Y_1 and Y_2 are independent Poisson random variables with mean λ_1 and λ_2.

B. Performance Analysis

In this subsection, unless otherwise specified, the number of memory slots is 100 per interface and the entanglement swapping probability $q = 0.9$. The channels between the switch and end nodes are lossy optical fibers, and the entanglement generation probability $p = 0.9$ is the same for all interfaces. Note that $p = 0.9$ corresponds to a distance between the switch and end nodes of 2.3 km given a fiber attenuation coefficient of 0.2 dB/km.[5] The qubits suffer from dephasing noise [29] when staying idle in memory slots, and the T2 time for the dephasing noise in each memory slot is set to 1 millisecond. Specifically, the dephasing noise model in a memory slot is modelled as follows:

$$\mathcal{N}_{\text{dephase}}: \quad \rho \to (1 - p_{\text{dephase}})\rho + p_{\text{dephase}} \sigma_Z \rho \sigma_Z$$

where ρ is the density matrix of a qubit, $\sigma_Z = |0\rangle\langle 0| - |1\rangle\langle 1|$ is one of the Pauli operators, and p_{dephase} is the dephasing probability, given by

$$p_{\text{dephase}} = \frac{1}{2}(1 - \exp\{-\Delta t / T_2\})$$

in which Δt denotes the time that a qubit stays idle in the memory slot. If one qubit of an EPR $|\Psi\rangle$ is stored in a memory qubit, then after time Δt, one can verify that its fidelity becomes $(1 + \exp\{-\Delta t / T_2\})/2$.

Protocol	Average Fidelity	Average Latency (μs)
Stationary (YQF)	0.976	15.6
Stationary (OQF)	0.908	14.2
On-demand (YQF)	0.975	12.4
On-demand (OQF)	0.916	14.4

TABLE II: Performance of Entanglement Swapping Protocols: $\lambda_{ij} = 0.2/\mu\text{s}, \forall i, j \in \mathcal{K}$.

Protocol	Average Fidelity	Average Latency (μs)
Stationary (YQF)	0.961	0.092
Stationary (OQF)	0.752	0.089
On-demand (YQF)	0.960	0.080
On-demand (OQF)	0.752	0.067

TABLE III: Performance of Entanglement Swapping Protocols: $\lambda_{ij} = 0.12/\mu\text{s}, \forall i, j \in \mathcal{K}$.

We begin with a comparison of the stationary protocol to an on-demand protocol. For each protocol, we further implement the YQF and OQF methods to prioritize EPR pairs. The number of interfaces is $K = 5$. Tables II and III show the performance of the developed protocols for $\lambda_{ij} = 0.2/\mu$s and $\lambda_{ij} = 0.12/\mu$s, $\forall i, j \in \mathcal{K}$, respectively. First, the on-demand protocols and stationary protocols perform similarly in terms of fidelity. Regarding latency, the on-demand protocols perform the best in most cases. Note that the on-demand protocol does not require any statistical knowledge of the requests or the systems, and that it involves low computational overhead. Therefore, the on-demand protocol is a desirable choice in practice. Second, the average fidelities of the YQF protocols are much higher than those of the OQF protocols, but the average latencies are generally greater than those of the OQF protocols, especially when λ_{ij} is small. This agrees with intuition. YQF tends to use newly generated EPR pairs for entanglement swapping, so the fidelities are higher. Moreover, since the EPR pairs are discarded when their fidelities are low, more EPR pairs are discarded when YQF is used. Since fewer EPR pairs are used, latencies increase. We next use the

[5]When $p = 0.9$ (i.e., the distance between the switch and end nodes is 2.3 km), the link level entanglement rate is 0.9 EPR pair per μs. This is reasonable since ideally photon sources can emit photons at the rate of 10^7 Hz, corresponding to 10 EPR pairs per μs.

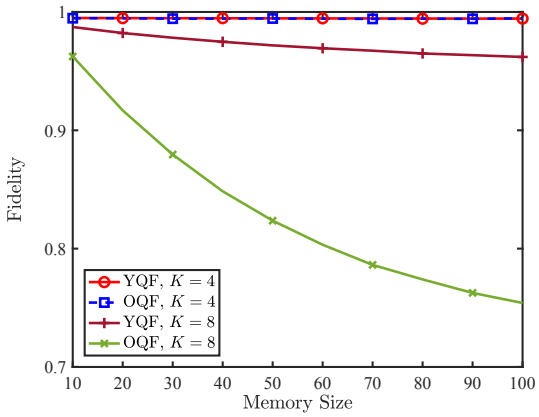

Fig. 2: Average fidelity as a function of the memory size. The x-axis denote the number of memory slots normalized by K.

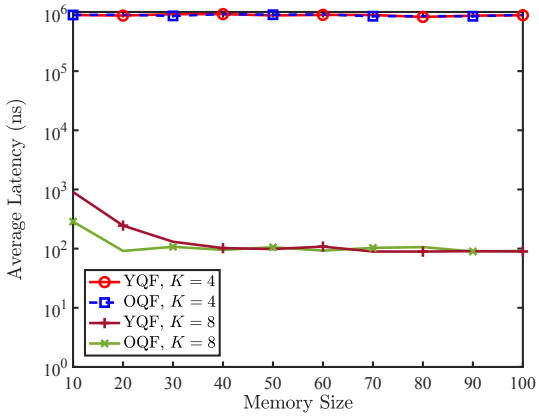

Fig. 3: Average latency as a function of the memory size. The x-axis denote the number of memory slots normalized by K.

on-demand protocol as the default protocol to evaluate the performance of quantum switch in different settings.

Figures 2 and 3 respectively show average fidelity and latency of the quantum switch achieved by the on-demand protocols (YQF and OQF) as functions of the number of memory slots. In these two figures, we set $\sum_{i,j}\lambda_{ij} = 1.2/\mu$s. For $K = 8$, the average fidelity and average latency decrease with memory size; for $K = 4$, the average fidelity and average latency remain constant. This is because for $K = 4$, the rates $(\lambda_{ij})_{1\le i,j\le 4}$ lie outside the capacity region. The quantum switch is then unstable, leading to high latencies. Since there are many unserved entanglement requests, most of the time the generated qubits between the switch and end nodes are used immediately for entanglement swapping, and this leads to a fidelity close to one. For $K = 8$, the rates $(\lambda_{ij})_{1\le i,j\le 8}$ lie inside the capacity region. In many occasions, there are no unserved entanglement requests, and qubits in the memory suffer from decoherence. An increase in memory slots results in qubits staying in the memory longer before consumed, leading to a decrease in fidelity. In addition, more memory slots imply that the quantum switch discards fewer EPR pairs and therefore reduces latency.

Figures 4 and 5 respectively show average fidelity and latency of the quantum switch achieved by the on-demand protocols (YQF and OQF) as functions of the entanglement swapping success probability q. In these two figures, we set $\sum_{i,j}\lambda_{ij} = 2/\mu$s. First, fidelity initially decreases with q and then remains constant. This is because for small q, the rates are outside the capacity region, and the generated qubits between the switch and end nodes are used immediately for entanglement swapping, leading to fidelities close to one. As q increases, the rates gradually move into the capacity region, and some qubits stay in the memory before being consumed for entanglement swapping, leading to a decrease in fidelity. When q is sufficiently large, the memory slots are full most of the time, and increasing q does not significantly improve the memory occupancy rate. In these occasions, the average fidelity does not change with q. Second, the average latency

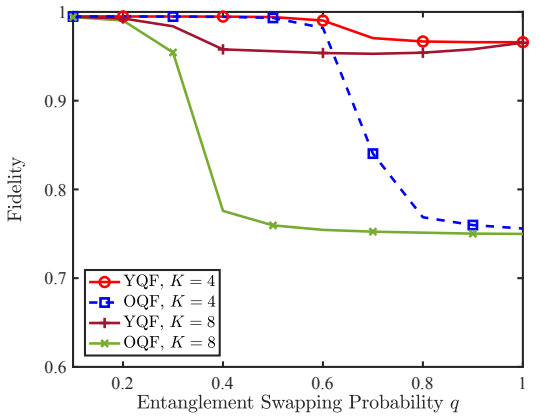

Fig. 4: Average fidelity as a function of entanglement swapping success probability q.

first decreases with q and then remains constant. The reasoning for this behavior is almost the same as that for fidelity. Third, one observes that the sudden change of fidelity and latency occurs around $q = 0.33$ for $K = 8$ and $q = 0.67$ for $K = 4$. These two points exactly correspond to the boundary points for the capacity region. This shows that the switch demonstrate entirely different behavior inside or outside the capacity region, which is consistent with the stability analysis in earlier sections. This figure also shows that if q corresponds to a boundary point of the capacity region, increasing q provides limited performance improvement.

VIII. Conclusion

We develop efficient entanglement swapping protocols for a quantum switch and analyze their performance in terms of stability of the switch, fidelity of EPR pairs and latency of entanglement requests. We determine the capacity region for entanglement rates under the assumption of infinite memory size and infinite qubit lifetime. Specifically, we show that no entanglement swapping protocols can stabilize the switch if

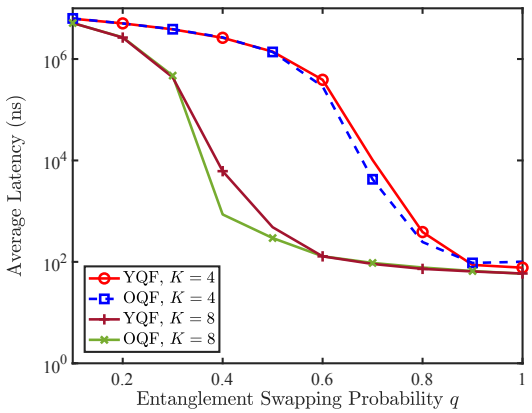

Fig. 5: Average latency as a function of entanglement swapping success probability q.

the entanglement rates lie outside this capacity region, and develop different protocols that stabilize the switch when the entanglement rates correspond to an interior point of the capacity region. The performance of the developed protocols is evaluated using NetSquid, and practical constraints such as decoherence in memory slots are accounted for. We show that stationary protocols and on-demand protocols exhibit high fidelity and low latency. Moreover, the tradeoff between memory size, decoherence rate, fidelity, and latency shown in the simulation results offer guidance in the implementation of quantum switches.

A potential future direction is the design and analysis of switches with a finite memory size and finite lifetime. Note that in this manuscript, simulation results are obtained in practical scenarios, whereas the stability analysis is based on the assumptions that the switch has sufficiently many memory slots and that qubits have infinite lifetime when stored in the memory. The capacity region for a finite memory size and lifetime must be different, and the entanglement swapping protocols must be designed accordingly.

Acknowledgment

The authors thank Philippe Nain for his helpful suggestions and careful reading of the manuscript.

Appendix I
Sketch of the Proof for Theorem 3

We consider a random event $S(t)$ defined as follows:

$$S(t) = \Big\{ \sum_{i \in \mathcal{K}} \big[U_{ij}(t) + A_{ij}(t)\big] \le E_{0j}(t) + C_{0j}(t), \forall j \in \mathcal{K} \Big\} \tag{25}$$

When the event $S(t)$ occurs, then the quantum switch has sufficient entanglement to address the entanglement request, i.e., $F_{ij} \ge U_{ij}(t)$, $i, j \in \mathcal{K}$ when an on-demand protocol is used.

Lemma 1: Under Assumption 1, if an on-demand protocol is used, then there exists a constant c_2 irrelevant of t, such that for any $\tilde{i}$, $\tilde{j} \in \mathcal{K}$

$$\mathbb{E}\Big[\sum_{\tau=0}^{t} A_{\tilde{i}\tilde{j}}(\tau) | \overline{S}(t) \Big] \mathbb{P}\big[\overline{S}(t)\big] \le c_2. \tag{26}$$

We skip the proof for Lemma 1 here due to space constraints and refer to the full version [25] for the detailed proof.

Let $\boldsymbol{u}$ denote a matrix with (i,j)-th element u_{ij}. Define

$$L_{\text{od}}(\boldsymbol{u}) = \sum_{i,j \in \mathcal{K}} u_{ij} \tag{27}$$

as a Lyapunov function of unprocessed entanglement request. For a control policy, we consider the following unconditional 1-step Lyapunov drift

$$\tilde{\Delta}_1^{\pi}(t) = \mathbb{E}\big[L_{\text{od}}(\boldsymbol{U}(t+1)) - L_{\text{od}}(\boldsymbol{U}(t))\big]. \tag{28}$$

Lemma 2: For an on-demand protocol, the 1-step Lyapunov drift at any slot t satisfies

$$\tilde{\Delta}_1^{\pi}(t) \le \sum_{i,j \in \mathcal{K}} \Big[\lambda_{ij} + q c_2 - q\, \mathbb{E}\Big[U_{ij}(t)\Big] \Big]$$

where c_2 is the constant in (26).

Proof: Recall the definition of $S(t)$ in (25). Note that

$$\begin{aligned} \tilde{\Delta}_1^{\pi}(t) = \mathbb{P}\big[S(t)\big] \sum_{i,j \in \mathcal{K}} \mathbb{E}\Big[U_{ij}(t+1) - U_{ij}(t) | S(t)\Big] \\ + \mathbb{P}\big[\overline{S}(t)\big] \sum_{i,j \in \mathcal{K}} \mathbb{E}\Big[U_{ij}(t+1) - U_{ij}(t) | \overline{S}(t)\Big]. \end{aligned} \tag{29}$$

Note that if $S(t)$ occurs, then the control variable $F_{ij}(t) = U_{ij}(t) + A_{ij}(t)$, for all $i, j \in \mathcal{K}$. This leads to

$$\begin{aligned} &\mathbb{E}\Big[U_{ij}(t+1) - U_{ij}(t) | S(t)\Big] \\ &\quad \le (1-q)\mathbb{E}\Big[A_{ij}(t) | S(t)\Big] - q\, \mathbb{E}\Big[U_{ij}(t) | S(t)\Big]. \end{aligned} \tag{30}$$

If $\overline{S}(t)$ occurs, we have

$$\mathbb{E}\Big[U_{ij}(t+1) - U_{ij}(t) | \overline{S}(t)\Big] \le \mathbb{E}\Big[A_{ij}(t) | \overline{S}(t)\Big]. \tag{31}$$

Combining (29),(30), and (31), we have

$$\tilde{\Delta}_1^{\pi}(t) \le \sum_{i,j \in \mathcal{K}} \Big[\lambda_{ij} - q \mathbb{P}[S(t)]\, \mathbb{E}\Big[U_{ij}(t) | S(t)\Big] \Big]. \tag{32}$$

On the other hand,

$$\begin{aligned} \mathbb{E}\big[U_{ij}(t)\big] &= \mathbb{P}[\overline{S}(t)] \mathbb{E}\Big[U_{ij}(t) | \overline{S}(t)\Big] \\ &\quad + \mathbb{P}[S(t)] \mathbb{E}\Big[U_{ij}(t) | S(t)\Big] \\ &\le c_2 + \mathbb{P}[S(t)] \mathbb{E}\Big[U_{ij}(t) | S(t)\Big] \end{aligned} \tag{33}$$

where the last inequality is because of Lemma 1.

Combining (32) and (33), we have

$$\tilde{\Delta}_1^{\pi}(t) \le \sum_{i,j \in \mathcal{K}} \Big[\lambda_{ij} + q c_2 - q\, \mathbb{E}\big[U_{ij}(t)\big] \Big]$$

which completes the proof of Lemma 2. □

Lemma 2 implies

$$\sum_{i,j\in\mathcal{K}}\Big[\mathbb{E}\big[U_{ij}(t+1)\big]-\mathbb{E}\big[U_{ij}(t)\big]\Big] \le \sum_{i,j\in\mathcal{K}}\Big[\lambda_{ij}+qc_2-q\,\mathbb{E}\big[U_{ij}(t)\big]\Big]. \tag{34}$$

Summing (34) over t from 0 to $T-1$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i,j\in\mathcal{K}} q\mathbb{E}\big[U_{ij}(t)\big] \le \sum_{i,j\in\mathcal{K}}\Big(\lambda_{ij}+qc_2+\mathbb{E}\big[U_{ij}(0)\big]/T\Big).$$

Then for any $i_0, j_0 \in \mathcal{K}$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{P}\big[U_{i_0j_0}(t)>V\big] \le \frac{\sum_{i,j\in\mathcal{K}}(\lambda_{ij}+qc_2+\mathbb{E}\big[U_{ij}(0)\big]/T)}{qV}$$

Taking $\limsup_{t\to\infty}$, we see that $g_{i_0j_0}(V)$ defined in (2) is on the order of $O(1/V)$. Taking limits as $V\to\infty$, we have $g_{ij}(V)\to 0$, $i,j\in\mathcal{K}$.

Appendix II
Sketch of the Proof for Theorem 5

We consider a random event $S_0(t)$ defined as follows:

$$S_0(t)=\Big\{\sum_{i\in\mathcal{K}}\Big\lceil\big[\widetilde{U}_{ij}(t)+\widetilde{A}_{ij}(t)\big]/q\Big\rceil \le E_{0j}(t)+C_{0j}(t), \forall j\in\mathcal{K}\Big\} \tag{35}$$

where $\widetilde{U}_{ij}(t)$ and $E_{0j}(t)$ are the number of backlog and entanglement requests $|\Psi_{ij}\rangle$ and the number of entanglements $|\Psi_{0j}\rangle$ achieved by the used on-demand protocol. If event $S_0(t)$ occurs, then the quantum switch has sufficient entanglement to address the requests, i.e., $F_{ij}(t)\ge\Big\lceil\big[\widetilde{U}_{ij}(t)+\widetilde{A}_{ij}(t)\big]/q\Big\rceil$, $i,j\in\mathcal{K}$.

Lemma 3: Under Assumption 2, if an on-demand protocol is used, then for any $\tilde{i}$, $\tilde{j}\in\mathcal{K}$

$$\lim_{t\to\infty}\mathbb{E}\Big[\sum_{\tau=0}^{t}\widetilde{A}_{\tilde{i}\tilde{j}}(\tau)|\overline{S_0}(t)\Big]\mathbb{P}\big[\overline{S_0}(t)\big]=0. \tag{36}$$

The proof for Lemma 3 is similar to that for Lemma 1, and we skip it here due to space constraints.

We now consider $\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+\Big]$. Note that

$$\begin{aligned}&\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+\Big]\\ &=\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+|\overline{S_0}(t)\Big]\mathbb{P}\big[\overline{S_0}(t)\big]\\ &+\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil,0\big]^+|C_0\Big]\mathbb{P}\big[C_0\big]\\ &+\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+|S_0(t)\cap\overline{C}_0\Big]\mathbb{P}\big[S_0(t)\cap\overline{C}_0\big].\end{aligned} \tag{37}$$

We next show that the three terms in (37) converge to 0 as t goes to infinity and thus complete the proof of Theorem 5.

The first term in (37) converges to 0 because

$$\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+|\overline{S_0}(t)\Big]\mathbb{P}\big[\overline{S_0}(t)\big] \le \mathbb{E}\Big[\sum_{\tau=0}^{t}\widetilde{A}_{ij}(\tau)|\overline{S_0}(t)\Big]\mathbb{P}\big[\overline{S_0}(t)\big]$$

which converges to 0 because of Lemma 3.

Similarly, the second term in (37) converges to 0 because

$$\begin{aligned}&\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+|C_0\Big]\mathbb{P}\big[C_0\big]\\ &\le\mathbb{E}\big[\widetilde{U}_{ij}(t+1)|C_0\big]\mathbb{P}\big[C_0\big]\le\mathbb{E}\Big[\sum_{\tau=0}^{t}\widetilde{A}_{ij}(\tau)|C_0\Big]\mathbb{P}\big[C_0\big]\end{aligned}$$

which converges to 0 because of Assumption 2.

We now consider the third term in (37). The third term is a conditional expectation when the event $S_0(t)\cap\overline{C}_0$ occurs. Note that $\widetilde{U}_{ij}(t+1)=\widetilde{U}_{ij}(t)+\widetilde{A}_{ij}(t)-\mu_{ij}(t)$. Let

$$D_1=\big\{\widetilde{U}_{ij}(t)+\widetilde{A}_{ij}(t)>\lceil(t+1)^\alpha\rceil\big\}.$$

If $\overline{D}_1$ occurs, then $\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+=0$. We next consider that D_1 occurs. Let $\beta\in(0,\alpha-0.5)$ and consider the following event

$$D_2=\big\{\widetilde{U}_{ij}(t+1)\ge\big[\big(\widetilde{U}_{ij}(t)+\widetilde{A}_{ij}(t)\big)/q\big]^{0.5+\beta}\big\}.$$

Note that $\mathbb{P}[D_2\,|\,D_1]$ can be bounded by

$$\begin{aligned}\mathbb{P}[D_2\,|\,D_1]&\le 2\exp\Big\{-2\big[\big(\widetilde{U}_{ij}(t)+\widetilde{A}_{ij}(t)\big)/q\big]^{2\beta}\Big\}\\ &\le 2\exp\big\{-2(t+1)^{\alpha\beta}\big\}\end{aligned}$$

where the first inequality is based on Chernoff's bound and the last inequality is because D_1 occurs. If $\overline{D}_2$ occurs, then

$$\begin{aligned}&\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]^+\\ &\le\Big[\big[\big(\widetilde{U}_{ij}(t)+\widetilde{A}_{ij}(t)\big)/q\big]^{0.5+\beta}-\lceil(t+1)^\alpha\rceil\Big]^+\\ &\le\Big[\big[(\lambda_{ij}+\tilde{\epsilon})t/q\big]^{0.5+\beta}-\lceil(t+1)^\alpha\rceil\Big]^+.\end{aligned}$$

Then by discussing whether D_1 and D_2 occur, one can bound the third term in (37) as

$$\begin{aligned}&\mathbb{E}\Big[\big[\widetilde{U}_{ij}(t+1)-\lceil(t+1)^\alpha\rceil\big]|S_0(t)\cap\overline{C}_0\Big]\mathbb{P}\big[S_0(t)\cap\overline{C}_0\big]\\ &\le 2(\lambda_{ij}+\tilde{\epsilon})t\exp\big\{-2(t+1)^{\alpha\beta}\big\}\\ &+\Big[\big[(\lambda_{ij}+\tilde{\epsilon})t/q\big]^{0.5+\beta}-\lceil(t+1)^\alpha\rceil\Big]^+\end{aligned}$$

which converges to 0 as t goes to infinity because $\beta<\alpha-0.5$. This shows that the last term in (37) converges to 0.

REFERENCES

[1] P. W. Shor and J. Preskill, "Simple proof of security of the bb84 quantum key distribution protocol," *Phys. Rev. Lett.*, vol. 85, no. 2, p. 441, 2000.

[2] H.-K. Lo and H. F. Chau, "Unconditional security of quantum key distribution over arbitrarily long distances," *Science*, vol. 283, no. 5410, pp. 2050–2056, 1999.

[3] D. Gottesman, H.-K. Lo, N. Lütkenhaus, and J. Preskill, "Security of quantum key distribution with imperfect devices," in *Proc. IEEE Int. Symp. on Inf. Theory*, Chicago, USA, 2006, p. 135.

[4] G. L. Long and X. S. Liu, "Theoretically efficient high-capacity quantum-key-distribution scheme," *Phys. Rev. A*, vol. 65, no. 3, p. 032302, 2002.

[5] 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.

[6] X.-S. Ma, T. Herbst, T. Scheidl, D. Wang, S. Kropatschek, W. Naylor, B. Wittmann, A. Mech, J. Kofler, E. Anisimova *et al.*, "Quantum teleportation over 143 kilometres using active feed-forward," *Nature*, vol. 489, no. 7415, p. 269, 2012.

[7] W. Pfaff, B. J. Hensen, H. Bernien, S. B. van Dam, M. S. Blok, T. H. Taminiau, M. J. Tiggelman, R. N. Schouten, M. Markham, D. J. Twitchen, and R. Hanson, "Unconditional quantum teleportation between distant solid-state quantum bits," *Science*, vol. 345, no. 6196, pp. 532–535, Aug. 2014.

[8] G. M. D'Ariano, P. L. Presti, and M. G. A. Paris, "Using entanglement improves the precision of quantum measurements," *Phys. Rev. Lett.*, vol. 87, no. 27, p. 270404, Dec. 2001.

[9] Z. Huang, C. Macchiavello, and L. Maccone, "Usefulness of entanglement-assisted quantum metrology," *Phys. Rev. A*, vol. 94, no. 1, p. 012101, Jul. 2016.

[10] R. Demkowicz-Dobrzański and L. Maccone, "Using entanglement against noise in quantum metrology," *Phys. Rev. Lett.*, vol. 113, no. 25, p. 250801, Dec. 2014.

[11] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[12] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[13] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 466–478, Mar. 1993.

[14] M. Żukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert, "'Event-ready-detectors' Bell experiment via entanglement swapping," *Phys. Rev. Lett.*, vol. 71, no. 26, pp. 4287–4290, December 1993.

[15] F. Ewert and P. van Loock, "3/4-efficient bell measurement with passive linear optics and unentangled ancillae," *Phys. Rev. Lett.*, vol. 113, no. 14, p. 140403, Oct. 2014.

[16] W. P. Grice, "Arbitrarily complete bell-state measurement using only linear optical elements," *Phys. Rev. A*, vol. 84, no. 4, p. 042331, 2011.

[17] P. Kok, W. J. Munro, K. Nemoto, T. C. Ralph, J. P. Dowling, and G. J. Milburn, "Linear optical quantum computing with photonic qubits," *Rev. Mod. Phys.*, vol. 79, no. 1, pp. 135–174, 2007.

[18] L.-M. Duan and H. J. Kimble, "Scalable photonic quantum computation through cavity-assisted interactions," *Phys. Rev. Lett.*, vol. 9, no. 12, p. 127902, Mar. 2004.

[19] P. Nain, G. Vardoyan, S. Guha, and D. Towsley, "On the analysis of a multipartite entanglement distribution switch," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 4, no. 2, Jun. 2020.

[20] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the stochastic analysis of a quantum entanglement distribution switch," *IEEE Trans. Quantum Engineering*, vol. 2, p. 4101016, Feb. 2021.

[21] ——, "On the exact analysis of an idealized quantum switch," *Perform. Eval.*, vol. 144, no. 102141, Dec. 2020.

[22] ——, "On the capacity region of bipartite and tripartite entanglement switching," *ACM SIGMETRICS Performance Evaluation Review*, vol. 48, no. 3, pp. 45–50, Dec. 2020.

[23] T. Vasantam and D. Towsley, "Stability analysis of a quantum network with max-weight scheduling," *arXiv*, 2021. [Online]. Available: https://arxiv.org/abs/2106.00831

[24] E. Shchukin, F. Schmidt, and P. van Loock, "Waiting time in quantum repeaters with probabilistic entanglement swapping," *Phys. Rev. A*, vol. 100, no. 3, p. 032322, 2019.

[25] W. Dai, A. Rinaldi, and D. Towsley, "Entanglement swapping in quantum switches: Protocol design and stability analysis," *arXiv*, 2021. [Online]. Available: https://arxiv.org/pdf/2110.04116

[26] J. D. C. Little, "A proof for the queuing formula: $L = \lambda W$," *Oper. Res.*, vol. 9, no. 3, pp. 296–435, May 1961.

[27] T. Coopmans *et al.*, "NetSquid, a discrete-event simulation platform for quantum networks," *Commun Phys*, vol. 4, no. 164, 2021.

[28] M. Bhaskar, R. Riedinger, B. Machielse *et al.*, "Experimental demonstration of memory-enhanced quantum communication," *Nature*, vol. 580, pp. 60–64, 2020.

[29] J. Preskill, "Lecture notes for Physics 219," 2015.