Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/TQE.2020.DOI

On-demand resource allocation for a quantum network hub

SCARLETT GAUTHIER[1,2], THIRUPATHAIAH VASANTAM[3], AND GAYANE VARDOYAN[1,2,4]

[1]QuTech, Delft University of Technology
[2]Quantum Computer Science, Electrical Engineering, Mathematics and Computer Science, Delft University of Technology
[3]Department of Computer Science, Durham University
[4]Manning College of Information and Computer Sciences, University of Massachusetts, Amherst

Corresponding author: Scarlett Gauthier (email: s.s.gauthier@tudelft.nl).

SG acknowledges support from the Quantum Internet Alliance (QIA). QIA has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101102140. GV acknowledges support from NWO QSC grant BGR2 17.269. TV acknowledges support from the Engineering and Physical Sciences Research Council [Grant Ref: EP/Y028732/1]. GV is now at (4) with the University of Massachusetts, Amherst and was affiliated with (1) and (2) at Delft University of Technology while working on this research.

ABSTRACT

To effectively support the execution of quantum network applications for multiple sets of user-controlled quantum nodes, a quantum network must efficiently allocate shared resources. We study traffic models for a type of quantum network hub called an Entanglement Generation Switch (EGS), a device that allocates resources to enable entanglement generation between nodes in response to user-generated demand. We propose an on-demand resource allocation algorithm, where a demand is either blocked if no resources are available or else results in immediate resource allocation. We model the EGS as an Erlang loss system, with demands corresponding to sessions whose arrival is modeled as a Poisson process. To reflect the operation of a practical quantum switch, our model captures scenarios where a resource is allocated for batches of entanglement generation attempts, possibly interleaved with calibration periods for the quantum network nodes. Calibration periods are necessary to correct against drifts or jumps in the physical parameters of a quantum node that occur on a timescale that is long compared to the duration of an attempt. We then derive a formula for the demand blocking probability under three different traffic scenarios using analytical methods from applied probability and queueing theory. We prove an insensitivity theorem which guarantees that the probability a demand is blocked only depends upon the mean duration of each entanglement generation attempt and calibration period, and is not sensitive to the underlying distributions of attempt and calibration period duration. We provide numerical results to support our analysis. Our numerical results suggest that there exist parameter regimes where it is beneficial for nodes to relinquish control of EGS resources during their calibration periods. This benefit is quantified by the blocking probability and the total entanglement generated in a fixed period of time. Our work is the first analysis of traffic characteristics at an EGS system and provides a valuable analytic tool for devising performance driven resource allocation algorithms.

INDEX TERMS quantum networks, entanglement, quantum switch, queuing theory

I. INTRODUCTION

Quantum networks enable a variety of distributed applications that are not realizable via classical means alone. Among these are quantum key distribution (QKD) [1], [2], blind quantum computation (BQC) [3]–[5], and several entanglement-based quantum sensing techniques [6]–[9]. A quantum network consists of end nodes equipped with quantum hardware, as well as intermediate nodes – quantum repeaters or switches – whose main function is to enable the end nodes to carry out quantum communication tasks. In first- and second-generation quantum networks [10], these intermediate nodes use methods such as entanglement swapping as shown in Figure 1 to provide entanglement as a resource to user-run applications – either to be consumed directly, or used in teleportation-based transport of quantum information [11], [12]. When multiple applications have simultaneous demand for shared and limited resources (e.g., quantum memories, links, measurement modules), contention may arise, and the network must enact a resource allocation scheme.

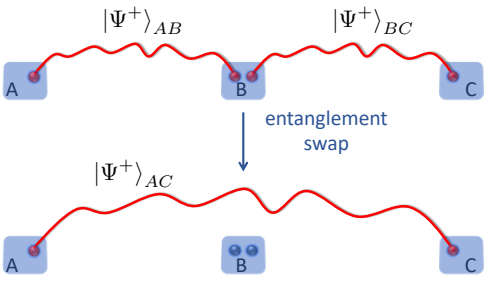


Figure 1: A simple quantum network with end nodes A and C wishing to share entanglement, and an intermediate node B assisting them with the task. Initially, two entangled links – $|\Psi^+\rangle_{AB}$ between A and B and $|\Psi^+\rangle_{BC}$ between B and C – are established. B then performs a swapping operation to directly entangle A and C's qubits. Depending on the distance between A and C, direct generation of entanglement (without an intermediate node), may not be feasible.

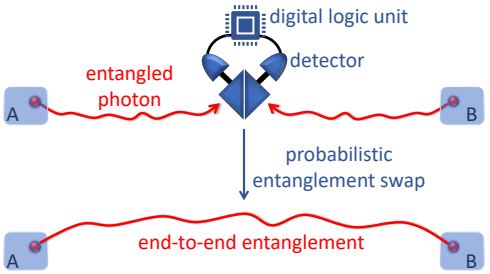


Figure 2: Entanglement generation with a Bell state analyzer. Entangled photons travel towards the BSA, which carries out a probabilistic entanglement swap. At the BSA, photons pass through a beam splitter whose output ports are connected to a pair of detectors. The digital logic unit reads out measurement results and determines if a swap was successful. Results are classically communicated to nodes A and B, whose qubits become entangled upon a successful event.

In this work, we study a type of quantum network hub previously referred to as an Entanglement Generation Switch (EGS) [13]. An EGS is a type of quantum switch with control over a pool of resources which, when allocated to a set of nodes, enable a task such as entanglement generation. Unlike its memory-equipped counterpart (sometimes referred to as an entanglement distribution switch, or EDS), the EGS is relatively easy to fabricate since it has no memories: it possesses a number of resources such as Bell state analyzers (BSAs) [14]–[16], which serve as a means of performing probabilistic optical entanglement swapping on incoming photons (each entangled with a qubit at an end node), and upon a

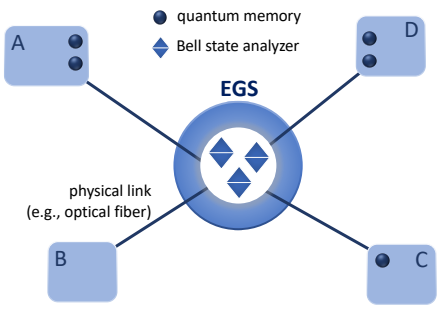


Figure 3: End nodes being serviced by an EGS with a pool of three BSAs.

successful swap generating end-to-end entanglement. This method of generating entanglement is illustrated in Figure 2, and the process is explained in more detail in Section III. In principle, an EGS can serve any number of nodes with a single shared BSA, but more BSAs can ameliorate contention for this resource; see Figure 3 for an example of an EGS with three BSAs to service four nodes. Some EDS proposals on the other hand additionally place BSAs in the middle of each physical link that connects the device to other nodes in the network. While these BSAs assist with entanglement generation at the link level, the resulting architecture is fairly demanding in the number and type of hardware components: K links translate into K dedicated BSAs and at least an equal number of quantum memories at the EDS. In contrast, the EGS architecture places all BSAs at the central hub along with a switching fabric for reconfiguration so as to serve any set of end nodes, resource limitations permitting. An enabling technology for an EGS is an optical switch linking N inputs to a pool of C resources (e.g., BSAs), such that any pair of inputs can be allocated any resource. Architectures for such switches exist, see e.g., [17] for the presentation and evaluation of optimal switching architectures based on 2×2 optical switches, or [18] for foundational work on rearrangeable non-blocking switches.

With these properties, the EGS is poised to be an excellent candidate for a scalable and straightforwardly implementable quantum network component, especially in the Noisy Intermediate Scale Quantum (NISQ) era [19]. While the EGS can be used to directly connect end nodes, as shown in Figure 3, it can also provide entanglement to other intermediate quantum network nodes, e.g., quantum repeaters/switches equipped with quantum memories of sufficiently long coherence time, each servicing a quantum local area network, as shown in Figure 4. The versatility of the EGS warrants investigation into its practical operation within a quantum network; we provide a detailed explanation of this in Section III.

We model the EGS as a fixed-capacity facility operating as an Erlang loss system. This is a classic queueing model (see e.g., [20], [21]) where arriving sessions are immediately

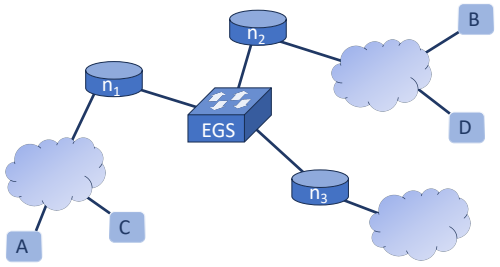


Figure 4: Intermediate quantum network nodes leveraging the EGS to interconnect their respective quantum local area networks.

rejected (not queued) when all resources are busy. This blocking behavior minimizes the response time with which service decisions are communicated, as queueing sessions naturally introduces latency before a session is served. A trade-off in performance arises, in that blocked sessions do not receive service, whereas queued sessions may typically receive service, albeit following an arbitrary delay. Response time is a critical performance metric for on-line entanglement routing protocols [22]. For example, in sequential entanglement swapping schemes [23], [24] end nodes must store entanglement while the end-to-end path between sender and receiver is constructed hop-by-hop, making the process highly sensitive to the latency of routing decisions [22]. Such routing based use cases for the on-demand resource allocation schemes we propose are particularly compatible with EGS systems operating as interconnects between several local areas, as illustrated in Figure 4.

We assume that entanglement requests arrive at the EGS as sessions according to a Poisson process, where each session consists of multiple entanglement generation attempts. We then analyze three operational modes of the device, while heeding physical capabilities and limitations of both the EGS and the nodes connected to it. Namely, inspired by realistic expectations of hardware characteristics in the near term, we equip the system with two important and pragmatic features: (i) batching of entanglement generation attempts due to generally high rates of failure; and (ii) provisioning for calibration periods necessitated by the quantum communication qubits of nodes that are served by the EGS. In the first operational mode scenario we model strict resource reservation, where an accepted demand is scheduled for several batches of entanglement generation interleaved with calibration periods and does not release the resource until all attempts are complete or entanglement generation succeeds, whatever happens first. The second scenario we study is a variation of the first in that a successful attempt does not trigger release of the EGS resource; all attempts are carried out, thus possibly resulting in multiple successfully-established entangled states. In the final scenario, nodes relinquish EGS resources during calibration periods, and must re-obtain them afterwards – we opt for jump-over blocking to describe the retrial behavior of this system. In all scenarios, an incoming request is blocked (dropped) if upon arrival it sees no free resource at the switch. While studying the EGS in the context of these different operation modes, we make the following contributions:

- We provide a comprehensive description of EGS operational details, with system specifications rooted in practical considerations of the underlying physical architecture;
- We model the EGS as an Erlang loss system with sessions of demands arriving as a Poisson process, which we subsequently analyze to obtain (i) the stationary distribution of the number of active requests being served at the switch; (ii) request blocking probabilities; and (iii) an insensitivity result that highlights the broad applicability of our model to practical systems; We highlight that in the third EGS operation mode (with resource relinquishment and retrials), the blocking probability is identical for the first attempt of a session as well as all retrials within that session. This property simplifies application of the result, as only one formula needs to be evaluated.
- We develop an extensive simulation framework capable of enacting sequences of events that model the operation of a real EGS – one that operates in discrete time – in a variety of configurations. Simulation code will be made available to enable future studies of the EGS.

The numerical results obtained from our continuous-time switch model closely match the more realistic discrete-time simulation of the switch. The physical relevance of our model, both in the hardware design we consider, as well as system control protocols we propose, set this work apart from much of the previous literature, wherein hardware limitations are frequently understated. The wide scope of our framework moreover enables one to model arbitrary traffic patterns and a wide variety of hardware settings, including ones where nodes have multiple communication qubits.

The rest of this manuscript is organized as follows: in Section II, we provide the relevant queueing-theoretic and quantum switching background. In Section III, we outline the system description, including physically-motivated operation settings. In Section IV, we introduce the model of the EGS and state our assumptions. Section V presents the analysis, while Section VI provides a numerical evaluation of the system. We make concluding remarks in Section VII.

II. BACKGROUND

An EGS serves a role analogous to a central hub in a system of telephone lines, managing the connection of pairs of quantum nodes to resources, much like a telephone exchange directs and facilitates communication between sets of callers. Traditionally, telephone systems are studied using the Erlang loss model, wherein calls arrive according to a Poisson

process to a server with a total of C telephone lines. An incoming call will be blocked if all C lines are occupied upon arrival. The blocking probability of a call is computed using the well-known Erlang formula [20]. It has been shown that this model exhibits insensitivity to the type of service time distribution of calls, as the blocking formula depends only on the *average* service time of calls [25]. This result follows from the argument that the underlying Markov process describing the system is a partially reversible one, which is a necessary and sufficient condition to have insensitivity [26]. A common method to prove insensitivity is to first study the given queueing model assuming service time distributions are Coxian (these are dense in the class of distributions with nonnegative support [27]), and then show that the stationary distribution or blocking formula is the same as in the case of an exponential service time distribution with the same mean. Then, by using the continuity of queueing models to service time distributions [28], the insensitivity property also holds for generally distributed service times, which can be approximated by Coxian distributions to an arbitrary degree of accuracy. The insensitivity property is a useful tool to dimension a practical system with a general service time distribution by studying the same system with the simpler case of an exponential service time distribution with the same mean.

In [29], Bonald studied the scenario where requests are generated as sessions that arrive according to a Poisson process, with each session containing several calls. It was shown that even in this case, the Erlang model is insensitive to service time distributions. In our model, calls also arrive as sessions, with each session consisting of several attempts for entanglement generation to describe practical quantum systems where entanglement requests arrive in batches from an application. We also assume sessions arrive according to a Poisson process, which will be a valid assumption when a large number of users or applications trigger entanglement requests. The analysis of this paper is based on the analysis of [29], albeit our aim here is to analyze a quantum system which has distinctive features when compared to classical systems. Our analysis is significantly different from [29] due to the presence of new parameters and characteristics of quantum systems, and subsequently analytical expressions for blocking probabilities for our model are different from those presented in [29]. We show that a quantum switch that can be modelled as an Erlang model also exhibits insensitivity to service time distributions under certain necessary conditions.

The EGS architecture was initially introduced in [13], where the authors highlighted the scalable properties of this type of hub. The authors then proposed and studied a Rate Control Protocol whose aim is to modulate user demand rates to the switch based on the EGS's capacity to serve users, as well as on overall traffic trends. The focus of this work is mainly on fair resource allocation, achieved through a network utility maximization-based [30] framework. In contrast, the protocols proposed in our work use request blocking instead of rate control as means of resource management. Furthermore, our work aims to accurately represent the EGS in a discrete setting, with concrete descriptions of request structure and procedures for request handling.

Memory-equipped quantum switches (EDSs) have been extensively investigated from queueing-theoretic and request scheduling perspectives, see e.g., [31]–[34]. In [32] the authors modeled an idealized EDS as a discrete-time Markov chain and computed its capacity under a simple swapping protocol. While this study shares similarities with ours in terms of the focus on discrete-time system evolution, important differences exist. As briefly outlined in the previous section, EDS-based models often assume the presence of quantum memories both at the end nodes and at the switch ([32], for instance assumes infinitely-sized buffers). Each node may even have dedicated infrastructure connecting it to the hub, including a pre-allocated quantum switch buffer and midpoint stations located at each link to herald entanglement. This enables switch-node entanglement to be attempted independently by each link, and successfully generated states to be stored until a swapping opportunity arises. With such an abundance of resources, contention is minimal, and resource allocation is not a primary focus. In another line of work on EDS-based models, contention among entanglement requests of different flows for using switch-node entanglement was studied in [34], [35]. In [35], a throughput optimal scheduling policy was designed for a quantum switch model when switch-node entanglement has a short lifetime of one time slot. They proved that the queues of entanglement requests are stable under their max-weight scheduling policy for all feasible entanglement request rates. A similar model with the scenario where switch-node entanglement has infinite lifetime was studied in [34].

In contrast to the EDS models, the EGS lacks memories, necessitating resource solicitation by nodes, followed by entanglement generation attempts executed in a synchronized manner to ensure nearly simultaneous photon arrival at the hub. Furthermore, our EGS protocols involve batched attempts interleaved with periods of EGS inactivity, effectively constituting extended "sessions" of engagement with EGS resources. The system studied in our work thus exhibits both architectural and algorithmic differences to the EDS, requiring novel and tailored analytical methodologies.

Quantum switches, both of the EGS and EDS types may be compatible with functional integration with quantum repeaters. Indeed, as with the EGS implementation we describe, several repeater architectures (see e.g., [36]–[40]) rely on the interference and measurement of photons by a BSA located at a midpoint between other nodes of the network. Quantum switches differ physically and operationally from repeaters because they require a switching fabric and protocols to effectively mediate contention for shared resources, whereas quantum repeaters are not subject to these requirements.

III. SYSTEM DESCRIPTION

An EGS consists of three main components: (1) A pool of allocatable *resources* such as BSAs; (2) A *switch* capable of allocating any resource to any pair of nodes connected to the EGS; (3) and a *processor* capable of scheduling the allocation of resources to pairs of nodes, controlling the operation of the switch, and sending and receiving classical messages. An EGS with control of a pool of three BSA-type resources is illustrated in Figure 3. Nodes are connected to an EGS by physical links, such as optical fiber connection. To gain access to an EGS resource, pairs of nodes send a message to the EGS called a *request*.

Definition 1. *Request. Nodes n_i and n_j who wish to communicate issue a* request *to the EGS. A request is a demand for the generation of one or more EPR pairs between n_i and n_j. The exact composition of a request depends both on the nodes' physical capabilities as well as on application-focused goals.*

For every distinct pair of nodes (n_i, n_j), one node is designated the *initiator* and the other the *secondary node*. Requests are communicated to the EGS by the initiator node. A node consists of the following set of components: (i) one or more *communication qubits*, each capable of preparing a quantum state and emitting one or more photons[1]; (ii) devices needed to manipulate the state(s) of the communication qubit(s) – examples include lasers, waveform generators and microwave sources; (iii) devices needed to measure a communication qubit; (iv) possibly one or more quantum memories to which the quantum state of a communication qubit may be swapped, capable of storing the state for a finite period of time; (v) and a classical processor capable of controlling the states prepared in communication qubits, triggering swaps to memory, triggering measurement of a communication qubit, and sending and receiving classical messages.

In Figure 3, the nodes connected to the EGS are assumed to be limited in that they either have few or no quantum memories, and may have a restricted number of communication qubits. Figure 4 illustrates a scenario where the EGS may be connected to more powerful nodes, potentially with access to many communication qubits and quantum memories per node. As in Figure 4, nodes can function as relays or intermediates interfacing with several quantum local area networks.

A. PHYSICAL OPERATION SETTINGS

Bipartite Heralded Entanglement (HE) generation and generation of Correlated Information (CI) are two ways in which a pair of nodes can interact via the EGS. In Appendix G, we describe a protocol for the latter, while here we describe in detail a protocol for bipartite HE generation [36], [37] that is based on a single-click scheme [42]. This method of producing entanglement has been successfully demonstrated in several experimental platforms, including Color Centers [42], [43], Ion Traps [44], [45], Atomic Ensembles [46], [47] and Neutral Atoms [48]. Applications of HE generation include BQC, teleportation and clock-synchronization [7], and an application of CI generation is Measurement Device Independent QKD (MDI-QKD) [49], [50]. Each of these tasks can be enabled by an EGS where the shareable resource is a BSA. The EGS can be equipped with multiple BSAs (Figure 3), or more generally with a different type of resource, in order to support other interaction protocols between quantum network nodes. Each of the protocols we describe is compatible with a BSA that consists of a 50/50 beam splitter with two input channels; each of the two output ports of the 50/50 beam splitter is connected to a photon detector; the outputs of the photon detectors are connected to a digital logic unit such as a Field Programmable Gate Array (FPGA) which processes the measurement outcomes and can communicate a success/failure flag back to the nodes of a flow. Such a BSA is depicted in Figure 2. In Table 3 (Section VI) we provide an inventory of physical and protocol parameters motivated by implementations of single click bipartite HE generation based on experimental demonstrations with a Nitrogen Vacancy (NV) center in diamond [42], [51].

1) Heralded entanglement generation

The goal of node pair (n_i, n_j) running the bipartite HE generation protocol is to entangle a communication qubit of node n_i with a communication qubit of node n_j. The term "bipartite" thus refers to the property of the protocol that the resulting entanglement involves two qubits. We will sometimes refer to such states as Einstein–Podolsky–Rosen (EPR) pairs or Bell states. The protocol is called *heralded* because for every attempt to generate entanglement that the nodes make, a success or failure flag is generated and converted into a message that is sent to the nodes, thereby indicating if the attempt succeeded or failed. Triggering a subsequent attempt after a success would destroy the entanglement that was created. To prevent wasting entanglement, the protocol includes a wait time for the heralding flag to arrive before triggering subsequent attempts. In certain EGS operation modes an attempt may not be triggered if a success flag is received (see Section III-A3). In a setting where node pairs use the entanglement generated between them to perform some application, this protocol is beneficial because it allows the nodes to condition commencement of the application on the successful generation of entanglement.

At a high level, a single-click heralded entanglement generation protocol consists of four stages. First, each node performs a sequence of calibration operations and prepares a communication qubit in a known state. Second, each node locally triggers the generation of entanglement between the state of their communication qubit and the presence/absence of a travelling photon. Third, the presence/absence encoded photons are sent to a BSA, at which a Bell-State Measurement (BSM) (entanglement swap) is attempted between the

[1] In some simple realizations of a node, the communication qubits may be replaced by all photonic state preparation devices. See [41] for an example.

encoded photons. Fourth, if the BSM succeeds the communication qubits of the two quantum processing nodes will have become entangled and a success flag is sent to the nodes. If the measurement is not successful a failure flag is communicated to the nodes. The second, third and fourth stages occurring sequentially constitute a single HE generation attempt. For the example physical platform of the NV center in diamond, the calibration operations correspond to a Charge and Resonance (CR) check [51].

Attempts can be repeated in batches of $N_{\text{attempts in batch}}$ sequential attempts that are interleaved with repetition of the first step – calibration of the communication qubit – following the last attempt in a batch. The main limitation on the batched attempt repetition rate is the Round Trip Time (RTT) of communication associated with the third and fourth stages of an attempt. The need to wait for the arrival of the heralding flag especially limits the rate, yet this is necessary to prevent the destruction of created entanglement by triggering a new attempt. This aspect of the protocol motivates an assumption in our mathematical models that entanglement generation attempts are non-overlapping. For any system where an individual attempt to generate entanglement has a low probability of success, it is beneficial to allow such batching of attempts, which increases the probability a success will occur within any finite amount of time.

A communication sequence between node pair (n_i, n_j), which requests and is allocated use of an EGS resource to perform HE generation, is included in Appendix G in Figure 18b.

2) Probability of success

We model experimental implementations of HE generation where the state of the communication qubit is reset (stage one) at the start of each attempt and every attempt in a batch corresponds to an identical experimental sequence. Furthermore we assume that the characteristics of devices used in triggering entanglement generation attempts, such as laser pump power and frequency, remain constant. Similarly, we assume that the optical characteristics of the path from the communication qubit to an allocated resource, including any loss introduced by the switch at the EGS, remain constant. Therefore, the probability of entanglement generation may only change over attempts if there are physical parameters that drift or jump over a batch of attempts. For any system where attempts have a fixed mean duration that is short in comparison to the parameter drift/jump timescales, such effects may be accounted for by assuming that the probability of successful entanglement generation is a function of the jth attempt in a batch of attempts, $p_{\text{gen}}(j)$.

For an implementation in the NV colour center in diamond, one may assume that the outcomes of sequential attempts in a batch are identically and independently distributed (IID), with a fixed probability of success p_{gen} [42]. This assumption is valid as long as calibration periods are performed frequently enough between batches of attempts to prevent slow effects – such as the spectral diffusion which affects solid state quantum emitters – from corrupting the state of the communication qubit. The assumption that the outcomes of sequential attempts are IID with a fixed probability of success p_{gen} also applies to other experimental platforms, such as Trapped Ions [44], [45], where the mean attempt duration is significantly shorter than sources of parameter drift. The assumption that p_{gen} is constant and is independent of attempt duration distributions but depends only on the mean attempt duration is the necessary condition that we use in proving the insensitivity result discussed in the introduction section (this is Theorem 2 in Section V).

3) Single vs multiple entanglement generation

For a limited quantum node, such as a node with one communication qubit and possibly a memory, it may be most practical to engage in single entanglement generation. That is, if an attempt to generate entanglement succeeds, no further attempts in a batch will be executed. Physically, successful entanglement generation renders the communication qubit of the device unavailable for further attempts until that entanglement can be used or transferred to memory. Transfers to memory are not instantaneous and have a finite time cost, thus communication qubits can not be freed instantly even in a system with memory. Moreover, if a communication qubit is coupled to a memory, attempts to generate entanglement while a state is stored in memory may damage the stored state due to induced decoherence [51]. This effect results from a persistent non-zero coupling to the memory. Single entanglement generation may be the preferential operation mode of limited quantum devices to account for these effects [51]. In contrast, a quantum node with multiple communication qubits may leverage them to generate multiple entangled states, possibly by multiplexing photon emission from the node.

IV. MODEL AND ASSUMPTIONS

In this section, we lay the groundwork for the analysis of a star-topology system with the EGS at its center, as shown in Figure 3. To this end, we first state the model and introduce the bulk of our notation. We then present several abstractly defined terms for the periods that comprise a session (a request for EGS resource access). These terms bridge the gap between the physical and queueing-theoretic models of the EGS. We then discuss various EGS operation modes, which we call service models. A service model describes how the EGS handles requests, including:

(1) resource reservation, specifying the amount and duration of resource allocation to a pair of communicating nodes;
(2) retrial behavior, specifying actions taken upon blocked service events; and
(3) termination behavior, specifying events that trigger the EGS to end service to a pair of nodes.

Modeling assumptions are introduced throughout, and we conclude with a table of notation.

A. SYSTEM MODEL AND NOTATION

1) Network Components

The network consists of a star-topology system with the EGS at its center, in which the EGS serves K quantum nodes. Node n_i, $\forall i \in \{1, \cdots, K\}$ has a total of c_i communication qubits which can be used for entanglement generation, while the EGS has C resource modules (e.g., Bell State Analyzers).

2) Sessions

A request issued from nodes n_i, n_j to access an EGS resource triggers the creation of a *session*.

Definition 2. *Session. A* session *between nodes n_i and n_j, denoted by the tuple (n_i, n_j, t) is a specification of the batches of entanglement generation attempts that are required to satisfy a request of type t. The specification includes a number of attempts per batch and a number of batches requested. The batches may be interleaved with idle and/or calibration periods. A session makes use of a single communication qubit at each of nodes $n_i,\ n_j$.*

The definition above makes a reference to a session type: as in [29], we permit the existence of differently-structured sessions within one system. Physically, these may encode the technical capabilities of nodes, such as the maximum number of attempts in a batch before calibration operations need to be repeated. Alternatively, session types may correspond to different applications, entanglement distribution algorithms, or even application instantiations. An example of the latter is a pair of nodes (n_1, n_2) that assist in carrying out two QKD instantiations across the network shown in Figure 4: one for the user pair (A, B) for a requested key size M_1, and the other for the user pair (C, D) for a requested key size $M_2 > M_1$. The session type corresponding to C and D's request would require more entanglement generation attempts over the course of all batches, and possibly more calibration or idle periods, depending on physical system restrictions or the algorithmic design of the EGS control protocol, respectively. Another example is the use of session types to accommodate application-dependent fidelity requirements: an application with a high minimum fidelity threshold can for instance request a session with more attempts in the hopes of producing enough states from which to distill higher-fidelity entanglement [52], [53].

3) Flows

The distinct possibilities for the sources of session requests are known as flows. A flow is associated with a pair of nodes, (n_i, n_j) and a session type t. The set $\mathcal{F}$ of all possible flows is

$$\mathcal{F} \equiv \{f^t_{i,j} : i, j \in \{1, \ldots, K\}, t \in \{1, \ldots, T\}\}, \quad (1)$$

where K is the number of nodes connected to the EGS and T is the total number of possible session types. The cardinality of $\mathcal{F}$ is $F = \binom{K}{2} \times T$, since each node-pair could run any of the T session types.[2] To streamline notation, we denote flows based on their index in $\mathcal{F}$, so the k^{th} flow is denoted by $f_k,\ k \in \{1, \cdots, F\}$. We note that while flows uniquely identify a type of session between a pair of nodes, sessions need not be unique: n_i and n_j can have multiple concurrent sessions of type t, as long as resources (communication qubits and EGS resources) are available. We introduce the notation $n_k \in f$ to mean that node n_k partakes in flow f.

4) Session Request Arrivals

Requests for a new session by flow f are triggered according to a Poisson process with rate ν_1^f, where the subscript indicates arrival to the first period of the session (see Appendix A for greater detail). However, each request is only actually submitted to the EGS if both participating nodes have an available communication qubit with which to execute a resulting session (one of total c_k communication qubits at node n_k). No request is submitted if one or both nodes does not have an available communication qubit. Once admitted, a session occupies one communication qubit from each node for its entire duration. We assume that individual sessions for a flow are independent.

5) System state and request blocking

The system state is characterized by the number of active sessions for each flow:

$$\mathbf{q} = [q_1, \ldots, q_F], \quad (2)$$

where $q_j \in \mathbb{Z}_+$ is the number of ongoing sessions of flow f_j, and $\mathbb{Z}_+$ indicates the set of non-negative integers. A request to reserve an EGS resource is blocked (rejected) if admitting it would violate the EGS capacity constraint (C resources).

6) Traffic Intensity

For every flow $f \in \mathcal{F}$, we define a quantity known as the *traffic intensity*,

$$\rho^f = \nu_1^f \times \mathbb{E}[\text{session duration}], \quad (3)$$

where the session duration includes all periods in which a communication qubit of each node in the flow remains occupied (details to follow in Section IV-B). The traffic intensity is a quantity which often arises in queuing theory. It is more typically defined as the ratio of the arrival rate of a session to the service rate of a session. As the mean session duration is equal to the inverse of the mean session service rate, (3) is equivalent to the standard definition. In Section V we will show that the session blocking probabilities only depend on these flow-level traffic intensities, rather than on the detailed distributions of session components.

[2]It is possible that certain node-pairs do not have the physical capability to carry out sessions of a given type t. The state space can be easily amended to reflect such restrictions; we make the assumption that all node-pairs are capable of all session types merely to simplify notation.

B. SESSION STRUCTURE

A session consists of a sequence of periods that differ in their resource usage patterns. The types of periods in a session form the basis of the service models, developed in Section IV-C.

1) Period Types

Definition 3. *Attempt. An* attempt *is the basic service component for two nodes communicating via the EGS. A attempt involves the active use of an EGS resource to make one entanglement generation attempt. An attempt consumes one EGS resource and one communication qubit from each node.*

An individual attempt period for flow f is denoted $\mathcal{A}_j^f$, where $j \in \{1, \cdots, M_A^f\}$. For the purpose of this work, we establish two additional service component types which are not attempts in that EGS resources are not in active use for their duration. For the first of these, we follow the convention set by [29] to define idle periods.

Definition 4. *Idle period. When nodes n_i and n_j enter an* idle period*, they relinquish all EGS resources, so that a subsequent call would require a new service reservation. Neither EGS resources nor active qubit operations are required. However, one communication qubit of each node remains reserved (unavailable for new session requests) during an idle period.*

We leave relaxations of the assumption that idle periods engage one communication qubit of each node for future work, as this assumption simplifies the analysis.

In our EGS model, there is also a possibility that of two communicating nodes n_i and n_j, one or more necessitates a calibration period after a number of active (attempt) periods. The duration of such calibration periods typically has a finite mean, albeit it can be randomly distributed. Depending on the service model, the nodes may decide to not relinquish resources they have already reserved at the EGS.

Definition 5. *Calibration period. A* calibration period *between nodes n_i and n_j requires node hardware resources from one or both nodes and engages one communication qubit of each node. No entanglement generation attempts occur in a calibration period. Depending on the service model, however, the nodes may continue to hold onto EGS resources for the duration of a calibration period, precluding other nodes from accessing them.*

We assume in this work that a calibration period engages one communication qubit of each node, as opposed to, e.g., all qubits of a node. While the latter scenario may also be of interest, it poses a challenge for analysis since service requests can no longer be treated independently from each other. We also remark that if the nodes relinquish EGS resources at the beginning of a calibration period, then from the perspective of the EGS (and from a modeling perspective) the period is an idle one. We distinguish between calibration and idle period types because calibration periods are always physically motivated: they necessarily engage quantum hardware at nodes. In contrast, idle periods need not stem from quantum hardware restrictions at nodes (even if we assume that communication qubits are unavailable for new service request creation during idle periods). Examples of these are an entanglement generation attempt followed by a classical processing period at the nodes, or a link-layer protocol that imposes a back-off timer between successive entanglement generation attempts.

2) Session Composition

A session of flow type f consists of:

- M_A^f attempt periods, with mean duration $1/\eta_{A,j}^f$, $\forall j \in \{1, \cdots, M_A^f\}$.
- M_C^f calibration periods, each with mean duration $1/\eta_{C,j}^f$, $\forall j \in \{1, \cdots, M_C^f\}$.
- M_I^f idle periods, each with mean duration $1/\eta_{I,j}^f$, $\forall j \in \{1, \cdots, M_I^f\}$.

These periods may be interleaved in various patterns depending on the service model. By convention, sessions always begin and end with attempt periods rather than calibration or idle periods.

We model periods by Coxian phase-type distributions, in which a single period is decomposed into a sequence of N phases, each with exponentially distributed duration such that the mean period duration is preserved. As discussed in Section II, these distributions can approximate generally distributed periods. The accuracy of the approximation can be made arbitrarily accurate by increasing N. It is physically relevant to model periods in this way, to ensure that our results can be applied to real systems, in which periods can follow general distributions. In Appendix A we discuss our model at the phase-level of detail. The key results of our analysis, presented in Section V only include quantities at the flow-level of detail.

3) Mean Session Duration

The duration of a session of flow type f depends on the service model and the mean duration of each period. In the physical systems our model describes, each type of period may have a generally distributed duration. However, we show in Section V, Theorem 2, that the session blocking probabilities are *insensitive* to these distributions. As a result, only the mean durations of periods contribute to the blocking probability, which happens via their contribution to the traffic intensity of flow f, ρ^f.

For example, if a session comprises M_A^f attempt periods each with average duration $1/\eta_A^f$, interleaved by M_C^f calibration periods each with average duration $1/\eta_C^f$ and no periods are skipped, then the mean session duration is

$$\mathbb{E}[\text{session duration}] = \frac{M_A^f}{\eta_A^f} + \frac{M_C^f}{\eta_C^f}.$$

However, some service models allow early termination (see Section IV-C).

C. SERVICE MODELS

We study three distinct service models that differ in their resource reservation policies, retrial behavior, and termination conditions. The key differences between service models are summarized in Table 1.

1) Single EPR Pair Generation with Strict Resource Reservation

In this service model a session consists of batches of attempt periods interleaved with calibration periods. Once a session is admitted at the EGS it holds onto its EGS resource until termination, even during calibration periods. The session terminates either when:

- An entanglement generation attempt succeeds, or
- All M_A^f attempts have been completed.

There are no idle periods in this service model, so $M_I^f = 0$. A calibration period is performed after every m attempts to maintain hardware performance. The strict resource reservation session structure corresponding to this service model is illustrated in Figure 5.

In this service model, the mean duration of a session for flow type f is given by

$$\mathbb{E}[\text{session duration}] = \sum_{j=1}^{M_A^f} \frac{\mathcal{P}_{A,j}^f}{\eta_{A,j}^f} + \sum_{j=1}^{M_C^f} \frac{\mathcal{P}_{C,j}^f}{\eta_{C,j}^f}, \qquad (4)$$

where $\mathcal{P}_{A,j}^f$ and $\mathcal{P}_{C,j}^f$ are the probabilities of reaching the j^{th} attempt and calibration periods, respectively. These probabilities account for the possibility of early termination of the session due to a successful entanglement generation attempt.

2) Multiple EPR Pair Generation with Strict Resource Reservation

This service model is identical to the previous service model, with the exception that a successful entanglement generation attempt does not trigger early termination. Instead, all M_A^f attempts are carried out, potentially producing multiple EPR pairs. The session terminates only when all attempts are complete. As this service model also uses strict resource reservation, the related session structure is again illustrated by Figure 5.

Since early termination cannot occur, the mean session duration simplifies to

$$\mathbb{E}[\text{session duration}] = \sum_{j=1}^{M_A^f} \frac{1}{\eta_{A,j}^f} + \sum_{j=1}^{M_C^f} \frac{1}{\eta_{C,j}^f}. \qquad (5)$$

In Section III we discussed physical considerations pertaining to single vs. multiple entanglement generation operation modes. These considerations imply that there is some subtlety as to what happens physically when entanglement is successfully generated in this service model, as well as in the following service model. If an attempt succeeds, the node will either proceed to consume that entanglement directly for some application or swap the entanglement to memory. Either of these options have a finite time cost. However, in this system model we assume that even if an attempt succeeds the communication qubit of each node associated with the session is available by the start of the next attempt, so that it may be executed. This assumption is an idealization which applies in the limit where the consumption of entanglement or swapping to memory can be completed instantaneously. Practically however, this model is still of interest for nodes which consume entanglement or execute swaps to memory on reasonably short time scales, as compared to the duration of a batch of attempts. In an operational setting such a node may simply skip a small number of subsequent attempts, until these tasks are concluded. Such an action will slightly reduce the expectation value of the number of entangled pairs that the session will produce, but will have no impact on the blocking probability as long as the nodes do not relinquish their EGS resource.

3) Multiple EPR Pair Generation with Resource Relinquishment

In this service model, sessions consist of *active periods*, which make use of EGS resources, interleaved with idle periods, during which resources are relinquished. At the end of each idle period, the session attempts to re-obtain an EGS resource module by submitting a request. The request can be blocked by the EGS if all C resources are occupied. Failure to obtain a resource (either at the beginning of a session or after an idle period) triggers a *jump-over retrial*: the session either transitions to the next idle period, or if one does not exist, terminates. A distinct feature of jump-over retrials is that when a retrial is blocked, the corresponding batch of attempts is skipped entirely. Hence the number of attempts that are actually conducted in a session is not fixed, but depends on the success of retrials. Notably, if the initial request and all retrials of a session are blocked, then a session may terminate from the final idle period without a single entanglement generation attempt having taken place. The structure of sessions corresponding to this service model is illustrated in Figure 6.

As in the other service models, one communication qubit of each node remains reserved throughout the session duration, including during idle periods and retrials. In contrast to the other service models, EGS resources are held only during active periods.

In this service model, the mean session duration depends non-trivially on the request blocking probabilities for the first request of the session, as well as for each subsequent retrial. In Appendix F we prove that the request blocking probability $\overline{\pi}_f$ for a session of flow type f does not depend on whether it is the first request of the session or a retrial. The mean session duration is given by

$$\mathbb{E}[\text{session duration}] = \sum_{j=1}^{M_A^f} \frac{\mathcal{P}_{A,j}^f}{\eta_{A,j}^f} + \sum_{j=1}^{M_I^f} \frac{1}{\eta_{I,j}^f}, \qquad (6)$$

where the probability $\mathcal{P}_{A,j}^f$ of reaching the j^{th} attempt de-

Feature	Single EPR Pair Generation with Strict Resource Reservation	Multiple EPR Pair Generation with Strict Resource Reservation	Multiple EPR Pair Generation with Resource Relinquishment
Idle periods	None	None	Present
EGS resource status during node calibration	Held by session	Held by session	Relinquished by session (calibration takes place in idle periods)
Retrial mechanism	None	None	Jump-over blocking
Session termination trigger	First successful attempt OR all attempts complete	All attempts complete	All retrials complete AND all attempts complete in accepted retrials

Table 1: Summary of the three different service models based on their resource reservation, retrial, and termination behaviors.

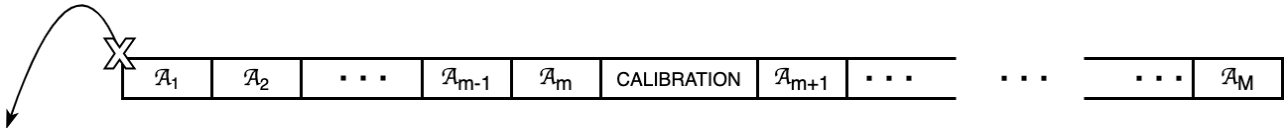


Figure 5: Strict resource reservation service model. A session consists of multiple EPR pair generation attempts, denoted by $\mathcal{A}_i$, $i = 1, \ldots, M_A$. A calibration period is carried out after every m attempts. In the "multiple EPR pair generation" variant of this service model, an admitted session does not relinquish resources for its entire duration, while in the "single EPR pair generation" variant, the session ends after a successful attempt.

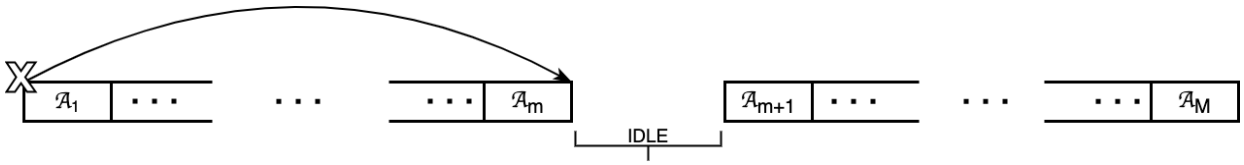


Figure 6: Service model with resource relinquishment and jump-over blocking. A session consists of multiple "active" (periods which make use of an EGS resource module), denoted by $\mathcal{A}_i$, $i = 1, \ldots, M_A$, interspersed with idle periods. In jump-over blocking, a blocked session goes to the beginning of the next idle period, if there is one, or terminates in case no idle periods remain in the session.

pends on the blocking probability $\overline{\pi}_f$ and how the attempts are partitioned into batches by the session structure.

For reference, Table 2 provides an overview of the notation used throughout our analysis.

V. ANALYSIS

In this section, we present the results of our analysis of the three EGS service models: single and multiple EPR pair generation with strict resource reservation, and multiple EPR pair generation with resource relinquishment during idle periods. The main quantities of interest are closed-form expressions for the probability that an arriving request is blocked. This is an event that occurs when said request sees all C EGS resources engaged. In strict resource reservation models, blocking can only occur at the beginning of a session. On the other hand, if sessions contain idle periods, as in the resource relinquishing model, then blocking may also occur throughout the session, after departures from idle periods. For the full details of our analysis see Appendices A-F. For each of the service models our analysis is structured according to the following approach:

1) Model the distribution of periods as Coxian phase-type distributions, to capture the general distributions that can occur in real systems (Appendix A).
2) Construct a Continuous-Time Markov Chain (CTMC) modeling session arrivals, the state of every session at a phase-level of detail, the transitions between phases of sessions, and the transitions leading to session termination. (Appendix A).
3) Derive the stationary distribution of the system using local balance equations. This is a common approach in queuing theory (see e.g. [21]) for deriving the stationary distribution of a CMTC (Appendix C).
4) Use the stationary distribution to obtain the probability that resources are occupied according to a particular state vector, when the system is in steady-state. Use these probabilities to derive the probability that an arriving request is blocked. Simplify by aggregating over phase-level states to obtain results to a period-level of detail. (Appendix D)
5) Prove that an insensitivity result applies (Appendices E and F): the probability that a request is blocked depends only on $\rho^f = \nu_1^f \times \mathbb{E}[\text{session duration}]$, rather than on the detailed distributions of period durations.

Symbol	Definition	Typical Range
System Parameters		
K	Number of nodes connected to the EGS	$\mathbb{Z}^+ = \{0, 1, 2, \cdots\}$
c_k	Number of communication qubits at node n_k	$\{1, 2, 3, \cdots\}$
C	Total number of EGS resources	$\{1, 2, 3, \cdots\}$
T	Number of session types	$\{1, 2, 3, \cdots\}$
F	Number of possible flows, $F = \lvert\mathcal{F}\rvert$, where $\mathcal{F}$ is the set of all possible flows	$\binom{K}{2} \times T$
Flow-level quantities		
$f,\ f_j$	Flow identifier	$f \in \mathcal{F}$
$n_k \in f$	Node n_k partakes in flow f	–
ν_1^f	Session arrival rate for flow f	$[0, \infty)$
ρ^f	Traffic intensity for flow f	$[0, \infty)$
q_j	Number of active sessions for flow f_j	$\mathbb{Z}^+$
$\mathcal{S}$	System state-space (a set of system states possible within a service model)	–
$\mathbf{q}$	System state vector: $[q_1,\ q_2, \cdots,\ q_F]$	$\mathbf{q} \in \mathcal{S}$
$\mathcal{A}_j^f$	The j^{th} attempt period for flow f	$[1, M_A^f]$
$M_A^f / M_C^f / M_I^f$	Number of attempt/calibration/idle periods	$\mathbb{Z}^+$
$1/\eta_{A,j}^f$	Mean duration of the j^{th} attempt period of a session for flow f	$[0, \infty)$
$1/\eta_{C,j}^f$	Mean duration of the j^{th} calibration period of a session for flow f	$[0, \infty)$
$1/\eta_{I,j}^f$	Mean duration of the j^{th} idle period of a session for flow f	$[0, \infty)$
Blocking probabilities		
$\overline{\pi}_f(C)$	Blocking probability for flow f when the EGS has C resources	$[0, 1]$
$\overline{\pi}(C)$	Average blocking probability when the EGS has C resources	$[0, 1]$

Table 2: Summary of the primary, period level notation used to present the results of our analysis. In Appendix A phase level notation is defined, which is used in the proofs of our results.

A. SINGLE EPR PAIR GENERATION WITH STRICT RESOURCE RESERVATION

Recall from Section IV-C1 that in this service model, sessions consist of batches of attempt periods interleaved with calibration periods. These sessions do not have any idle periods, so that $M_I^f = 0\ \forall f \in \mathcal{F}$. Once a session is admitted, it holds onto EGS and local resources until session termination, which either occurs upon the first successful entanglement generation attempt or after all attempts are exhausted. The following theorem states the probability that an arriving request for a session of type f is blocked.

Theorem 1. *For the system with single EPR pair generation operating in strict resource reservation mode, the probability that an arriving request belonging to flow $i \in \{1, \ldots, F\}$ is blocked is given by*

$$\overline{\pi}_i(C) = \left(\sum_{\mathbf{q} \in \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!} \right)^{-1} \sum_{\mathbf{q} \in \mathcal{Q}(C) \bigcap \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!}, \tag{7}$$

where $\mathbf{q} = [q_1, \ldots, q_F]$ *represents the number of active sessions* q_i *from each flow* f_i*, the overall traffic intensities* $\rho^{f_j}\ \forall j$ *are given by* (3) *and* (4)*,*

$$\mathcal{Q}(h) \coloneqq \Bigg\{ \mathbf{q} \in \mathbb{Z}_+^F : \sum_{i=1}^{F} q_i = h, \quad \sum_{i:n_k \in f_i} q_i \leq c_k, \forall k \in \{1, \ldots, K\} \Bigg\}, \text{ and} \tag{8}$$

$$\mathcal{Q}'(i) \coloneqq \Bigg\{ \mathbf{q} \in \mathbb{Z}_+^F : \sum_{j:n_k \in f_j} q_j \leq c_k, \forall n_k \notin f_i, \quad \sum_{j:n_l \in f_j} q_j < c_l, n_l \in f_i \Bigg\}, \tag{9}$$

where $\mathbb{Z}_+$ *indicates the set of non-negative integers.*

The set $\mathcal{Q}(h)$ contains all possible combinations of active sessions such that the total number of active sessions is exactly h, and communication qubit constraints are not violated. The set $\mathcal{Q}'(i)$ contains all combinations of active sessions such that communication qubit constraints are not

violated, with the additional constraint that nodes belonging to flow f_i have at least one unoccupied communication qubit each.

Proof. See Appendix D for the full proof. Here we sketch the idea of the proof.

Let $\mathrm{P}(\mathbf{q})$ denote the probability that EGS resources are occupied according to $\mathbf{q}$. Consider the following two events: $\Omega_1(h)$ is the event that h EGS resources are occupied, and $\Omega_2(i)$ is the event that flow f_i has available communication qubits. By the PASTA (Poisson Arrivals See Time Averages) property, we can write the probability that an arriving request of flow f_i sees h occupied resources (conditioned on the flow having enough qubits to generate a request), i.e., $\mathrm{P}(\Omega_1(h)|\Omega_2(i))$, as

$$\begin{aligned}\overline{\pi}_i(h) &= \frac{\mathrm{P}(\Omega_1(h)\cap\Omega_2(i))}{\mathrm{P}(\Omega_2(i))} \\ &= \left(\sum_{\mathbf{q}\in\mathcal{Q}'(i)}\mathrm{P}(\mathbf{q})\right)^{-1}\sum_{\mathbf{q}\in\mathcal{Q}(h)\cap\mathcal{Q}'(i)}\mathrm{P}(\mathbf{q}). \qquad (10)\end{aligned}$$

Here we use the fact that flow f_i would not be able to initiate a session if any node associated with f_i does not have an unoccupied communication qubit.

For a given flow-level state vector $\mathbf{q}$, there may be several phase-level state vectors $\mathbf{x}$ (introduced in Appendix A) compatible with describing the system configuration. The probability $\mathrm{P}(\mathbf{q})$ that flows occupy EGS resources according to $\mathbf{q}$ can be calculated by summing over the state dependent stationary distributions of the system $\pi(\mathbf{x})$ (derived in Appendix C1) compatible with the flow-level state $\mathbf{q}$. The final result eventually follows by substitution into (10) with $h = C$ and simplification of the resulting expression.

□

Using Theorem 1 we can derive an expression for the average blocking probability for an incoming request by taking an expectation over flow-type of the incoming request. For the following, let $\overline{\pi}_{f_i}(C) \equiv \overline{\pi}_i(C)$, where flow $f_i \in \mathcal{F}$ corresponds to the ith flow label in $\{1,\ldots,F\}$.

Corollary 1. *The average blocking probability of an incoming entanglement request, denoted by $\overline{\pi}(C)$, is given as*

$$\overline{\pi}(C) = \sum_{f\in\mathcal{F}}\left\{\frac{\mathrm{P}(\mathcal{Q}'(f))\nu_1^f}{\sum_{g\in\mathcal{F}}\mathrm{P}(\mathcal{Q}'(g))\nu_1^g}\right\}\overline{\pi}_f(C). \qquad (11)$$

Proof. Recall that $\mathcal{Q}'(f)$, given by (9), is the set of all combinations of active sessions such that communication qubit constraints are not violated, with the additional constraint that nodes belonging to flow f have at least one unoccupied communication qubit each. For a flow f, entanglement requests are generated according to a Poisson process with rate ν_1^f only when all the associated nodes have communication qubits, which happens with probability $\mathrm{P}(\mathcal{Q}'(f))$. If an entanglement request is triggered then the probability that it is of type f is proportional to $\mathrm{P}(\mathcal{Q}'(f))\nu_1^f$. From Theorem 1 we have the expression for the blocking probability for a type f request as $\overline{\pi}_f(C)$. Therefore the average blocking probability is given by (11). □

Finally, it remains to prove the insensitivity of each flow's blocking probability to the traffic characteristics of the system beyond the flow-level traffic intensities.

Theorem 2. *The blocking probabilities*

$$\overline{\pi}_i(C),\ i\in\{1,\ldots,F\}$$

for the system with single EPR pair generation and strict resource reservation depend only on the mean traffic intensities at the flow level, ρ^{f_i}, and are not sensitive to the underlying distributions of attempt and calibration period duration.

To prove the insensitivity theorem we show that the stationary distribution of the Markov chain that describes the system behaviour remains the same when attempt and calibration durations have Coxian or exponential distributions, as long as the means of the distributions match. This implies that the result also holds for a general distribution. A detailed proof of this theorem is presented in Appendix E.

B. MULTIPLE EPR PAIR GENERATION WITH STRICT RESOURCE RESERVATION

When the generation of multiple EPR pairs is permitted in strict resource reservation mode, each session, if admitted for service, traverses all periods. Consequently, session termination only occurs from the final attempt period of the session. The overall system is thus very similar to that of Section IV-C1, and all previous assumptions hold, with the exception of early termination following a successful entanglement generation attempt. The result is that in this service model the probability of reaching the j^{th} period is always 1. This modifies the "overall traffic intensity" from that given by (3) and (4) to

$$\begin{aligned}\rho^f &= \nu_1^f \times \mathbb{E}[\text{session duration}] \\ &= \nu_1^f \times \left(\sum_{j=1}^{M_A^f}\frac{1}{\mu_{A,j}^f} + \sum_{j=1}^{M_C^f}\frac{1}{\mu_{C,j}^f}\right). \qquad (12)\end{aligned}$$

Following these considerations, in the multiple EPR pair generation with strict resource reservation service model the blocking probability for a flow $f_i,\ i\in[1,\cdots,F]$ is given by (7), where $\forall f\in\mathcal{F},\ \rho^f$ is given by (12).

C. MULTIPLE EPR PAIR GENERATION WITH RESOURCE RELINQUISHMENT

Recall from Section IV-C3 that in this service model, sessions consist of batches of attempt periods interleaved with idle periods, at the start of which EGS resources are relinquished. Following an idle period, a session must re-attempt to secure an EGS resource. If no resource is available during such a retrial, then the request is blocked and the session jumps-over the intended batch of attempts and transitions to the next idle period, if there are any remaining. If no idle periods

remain, then such a failed retrial terminates the session. If instead the retrial following the last idle period succeeds, then termination of the session is triggered by completion of the final attempt in this final batch of attempts. These sessions do not have any EGS resource retaining calibration periods, so that $M_C^f = 0 \; \forall f \in \mathcal{F}$.

The blocking probability for this service mode is of the same form as (7) for the two previously discussed service models, with the main differences again manifesting through the traffic intensities ρ^{f_i}. Additional consideration is needed within the analysis to account for the fact that blocking may occur not only at the beginning of a session, but also during – following idle periods. We will show that for any given session, the probability of being blocked at the beginning of the session equals the probability of being blocked immediately after an idle period.

Theorem 3. *For the system with jump-over blocking service mode, the probability that a request belonging to flow $i \in \{1, \ldots, F\}$ is blocked at the beginning of a session or during retrials of a session is given by*

$$\overline{\pi}_i(C) = \left(\sum_{\mathbf{q} \in \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!} \right)^{-1} \sum_{\mathbf{q} \in \mathcal{Q}(C) \bigcap \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!}, \tag{13}$$

where $\mathbf{q} = [q_1, \ldots, q_F]$ represents the number of active sessions q_i from each flow f_i,

$$\mathcal{Q}(h) \coloneqq \left\{ \mathbf{q} \in \mathbb{Z}_+^F : \sum_{i=1}^{F} q_i = h, \; \sum_{i: n_k \in f_i} q_i \leq c_k, \forall k \in \{1, \ldots, K\} \right\}, \text{ and}$$

$$\mathcal{Q}'(i) \coloneqq \left\{ \mathbf{q} \in \mathbb{Z}_+^F : \sum_{j: n_k \in f_j} q_j \leq c_k, \forall n_k \notin f_i, \; \sum_{j: n_l \in f_j} q_j < c_l, n_l \in f_i \right\},$$

where $\mathbb{Z}_+$ indicates the set of non-negative integers.

Proof. We provide a proof in Appendix F. The proof relies on the phase-level system model introduced in Appendix A, and the derivation of the stationary distribution of the system in this service model given in Appendix C. □

Intuitively, (13) denotes the conditional blocking probability that all EGS resources are busy when the users of the flow f_i make an attempt to initiate a session or trigger retrials during a session. Eq. (13) represents the ratio of the probability that both nodes of flow f_i have unoccupied communication qubits and that all EGS resources are occupied (numerator), to the probability that nodes of flow f_i have unoccupied communication qubits to be able to participate in a session (denominator). We remark that *a priori* there is no reason to believe the the blocking probability for the first attempt of a session and subsequent retrials within that session will obey the same formula. Our result, that (13) applies equally to each of these cases, ensures that in practice it is straightforward to evaluate the blocking probability.

Description	Value
Link lengths	10 km
One-way communication time (RTT/2)	50.03 μs
Duration of calibration period (CR check), T_{calib}	1 ms
Probability single attempt success, p_{gen}	1e-5 (a.u.)
Duration single attempt, T_{attempt}	115.072 μs
Attempt batch size	100
# batches (strict allocation) or re-trials (jump-over)	10
# calibration (idle) periods in strict (jump-over) mode	9

Table 3: Physical parameters used in simulations correspond to an EGS supporting batched single click HE generation for NV colour centers in diamond as quantum nodes. We use a pessimistic value of p_{gen} as compared with published values (e.g., [42]) for the purpose of absorbing the additional sources of loss, e.g., that which may be introduced by insertion of an optical switch in the optical paths from nodes to EGS resources.

VI. NUMERICAL EVALUATION

In real-world implementations of entanglement generation, every individual attempt and calibration period has a finite duration. In a demonstration of deterministic HE delivery carried out between two NV nodes [42], for instance, the authors describe entanglement generation attempts as taking a fixed amount of time, $\Delta t_{\text{attempt}}$. On the other hand, the calibration periods take a variable amount of time, of which the mean duration μ_{calib} is known. To model such an experiment one could sample the duration of each calibration period from an exponential distribution with mean μ_{calib} and fix the duration of attempts to $\Delta t_{\text{attempt}}$.

To validate our analysis, we simulate three models of an entanglement generation experiment. These are referred to as *discrete*, *Cox*, and *exponential* and are differentiated by how the duration of entanglement generation attempts and calibration periods are determined. To ensure the simulations are compatible with our analysis two key assumptions are made. First, in all simulation modes and in numeric evaluation of (7) we fix the mean duration of every attempt (calibration period/idle period) to some value T_{attempt} ($T_{\text{calib}}/T_{\text{idle}}$). In discrete simulations the duration of each attempt (calibration period/idle period) is set exactly to these values. Settings (number of phases, duration of phases, transition probabilities between phases) for the Cox distribution, are chosen to ensure (56) is satisfied. Our insensitivity result suggests these settings may be chosen arbitrarily, as long as they obey (56). In Appendix H we include a sample of parameters used for the Cox distribution governing attempts. The second assumption is that the probability any attempt results in successful entanglement generation is a fixed value, p_{gen}. For justification, see Section III-A2. Simulation parameters are detailed in Table 3. Further details on simulation implemen-

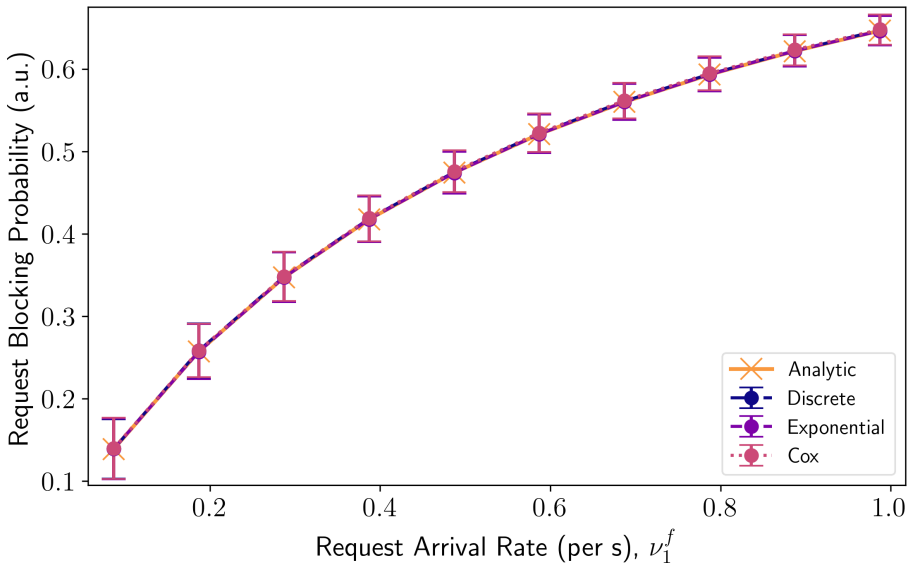


Figure 7: Comparison of the average blocking probability per flow according to (7) with simulations for an EGS with one resource, connected to eight nodes, each with a single communication qubit. The EGS serves $\binom{8}{2} = 28$ flows, one for each possible node pairing. Session traffic is homogeneous. The absolute relative errors are $(\delta_{\text{discrete}}, \delta_{\text{exponential}}, \delta_{\text{Cox}}) = (0.004, 0.001, 0.003)$.

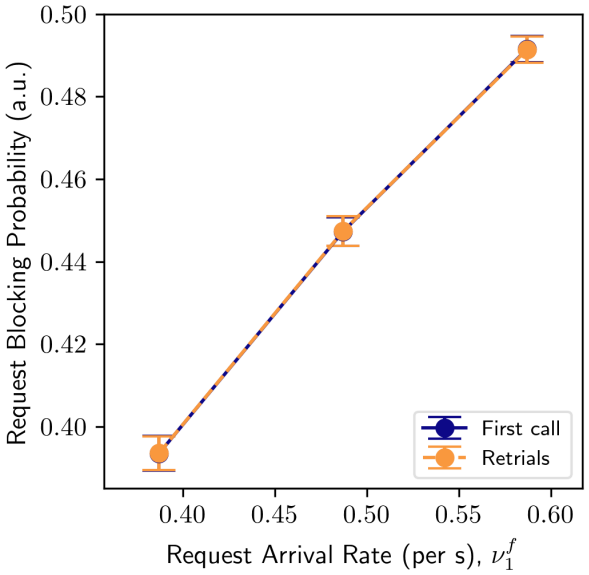


Figure 8: Average blocking probability for the first call of a session in the jump-over service model compared to that of calls which follow idle periods (retrials). Data is obtained from discrete simulations of an EGS with one resource, serving eight nodes via $\binom{8}{2} = 28$ flows. Every node is restricted to a single communication qubit. Session traffic is homogeneous. The maximum absolute relative difference between the average blocking probabilities is 0.00025.

tation are included in Appendix H.

To quantify agreement between numeric evaluation of (7) and simulated results we define error parameters $\delta_{\text{sim. type}}$ based on the maximum absolute relative difference between the points of the analytic and simulated data sets,

$$\delta_{\text{sim. type}} := \frac{|\max_x(y_{\text{Analytic}}[x] - y_{\text{Sim.}}[x])|}{y_{\text{Analytic}}[\arg\max_x(y_{\text{Analytic}}[x] - y_{\text{Sim.}}[x])]}, \quad (14)$$

where y_{Analytic} denotes an analytic data set, $y_{\text{Sim.}}$ denotes a simulated data set, and square brackets denote indexing the data sets. The error parameter reports the difference between the analytic and simulated data point for which the difference is maximum, relative to the analytic value at that point.

Each simulation is run for a duration corresponding to 1e7 iterations of the discrete simulation, equivalent to a simulation of 1150.73 seconds of *simulated real-time*. See Appendix H for details of the expected number of requests placed over the duration of a simulation. For any run of a simulation, the average blocking probability is calculated by averaging the blocking probabilities recorded by each flow. Every data point from a simulation is the result of averaging over 200 independent runs of the simulation. Error bars for request blocking probabilities correspond to the average (over the runs) standard deviation in blocking probabilities between flows.

Traffic in a deployed network will in general be non-homogeneous. The causes of non-homogeneity may stem from differences in the physical parameters of quantum nodes and their connections to an EGS or they may result from differences in request parameters submitted by flows, enabling them to target different applications. In general, our analysis applies to non-homogeneous traffic. However, the simplest

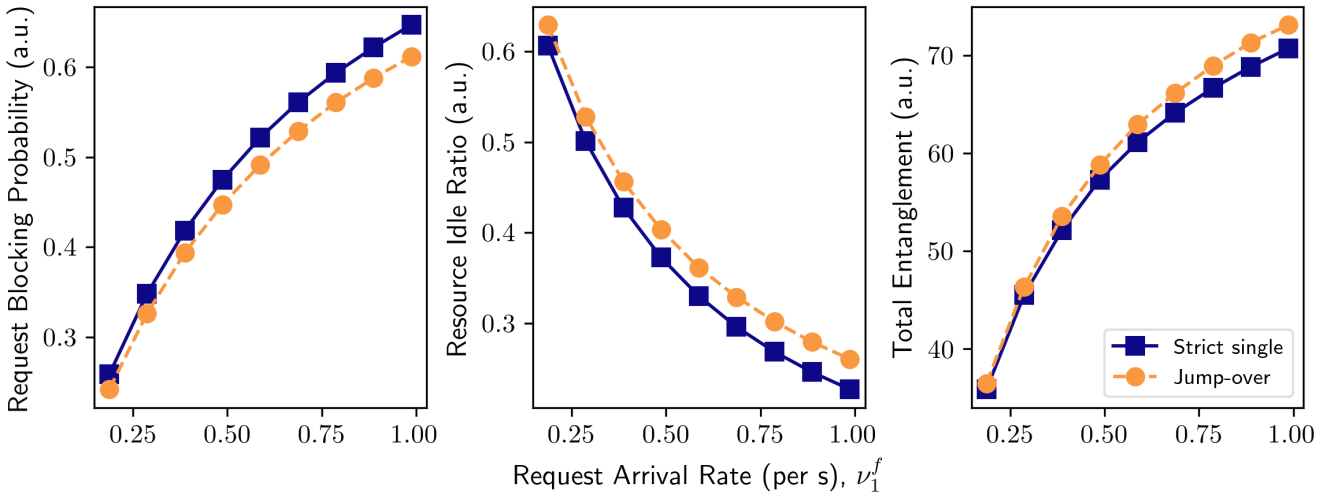


Figure 9: Comparison of the strict single and jump-over service models, for an EGS with one resource, serving eight nodes via $\binom{8}{2} = 28$ flows. Every node is restricted to a single communication qubit. Within each service model, session traffic is homogeneous. Left: blocking probability per request. Middle: proportion of time that the EGS resource is idle, compared to total simulation time. Right: total amount of entanglement generated by all sessions during the time simulated. Data is obtained using discrete simulations.

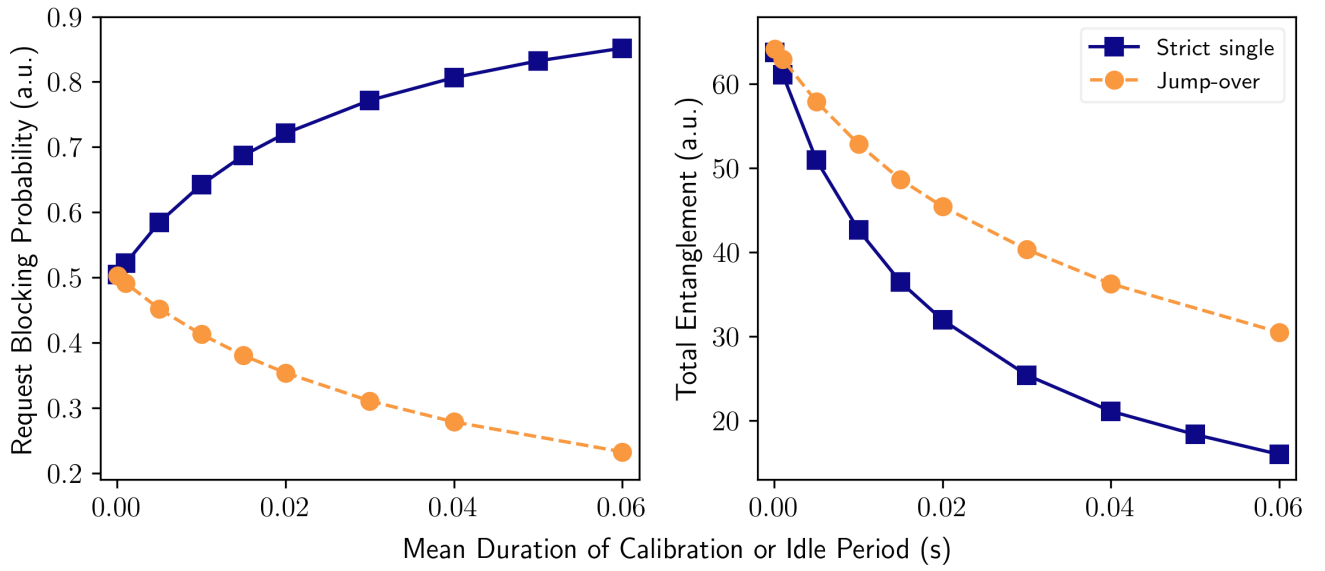


Figure 10: Comparison of the effect of varying the mean duration of calibration periods (strict single service model) and idle periods (jump-over service model). Within each service model, session traffic is homogeneous with a fixed request arrival rate of 0.587 per second. The EGS has one resource and serves eight nodes via $\binom{8}{2} = 28$ flows. Every node is restricted to a single communication qubit. Data is obtained using discrete simulations.

physical implementation of the EGS is one where all nodes are connected to the hub by links of equal length and all flows submit identical requests for sessions with equal rates. This is a homogeneous traffic scenario. In the remainder of this section we first present results from each of the three service models developed in Section IV in a homogeneous traffic scenario. Then, we demonstrate the application of our analysis to a non-homogeneous traffic scenario in the single entanglement generation with strict resource reservation service model.

A. HOMOGENOUS TRAFFIC

In a deployed quantum network, a network operator may be interested in selecting a service model for an EGS based on its performance across a range of metrics. We compare the performance of the three service models of Section IV. In what follows, these service models are referred to simply as *strict single*, *strict multiple* and *jump-over*, respectively. Besides request blocking probabilities, we also study resource utilization and the total entanglement generated in a fixed amount of time. The former provides information on how

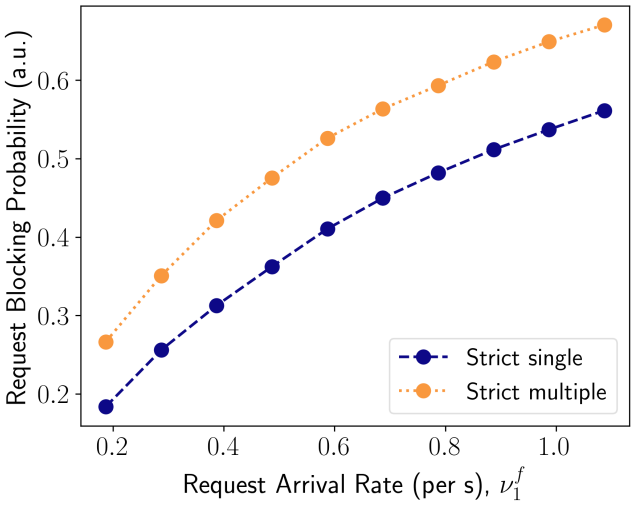


Figure 11: Comparison of the request blocking probabilities of the strict single and strict multiple service modes when there is a high probability ($p_{\text{gen}} = 0.001$) that an attempt to generate entanglement succeeds. Data is obtained from discrete simulations of an EGS with one resource, serving eight nodes with $\binom{8}{2} = 28$ flows. Every node is restricted to a single communication qubit. Session traffic is homogeneous. The absolute relative errors are $(\delta_{\text{discrete}}, \delta_{\text{exponential}}, \delta_{\text{Cox}}) = (0.014, 0.018, 0.003)$ for the strict single service mode and $(0.032, 0.015, 0.002)$ for the strict multiple service mode.

efficiently network resources are used, and the latter gives a measure of the productivity derived from the allocation of network resources. To study the resource utilization in simulation, we define the *resource idle ratio* as the proportion of time that one or more EGS resources is idle, relative to the entire simulated time. To study the total entanglement generated we track and sum the successful generation of entangled pairs by any session over the duration of the entire simulated time.

Before comparing the service models, we validate our analysis of the blocking probability for each of the three service models. Figure 7 compares numeric evaluation of (7) with results from discrete, exponential, and Cox simulations of an EGS operated in the strict single service model, with control of one resource $(C = 1)$, connected to eight nodes, each with a single communication qubit. Load on the EGS network results from requests for resource access by the flows. In alignment with expectations, we observe that the request blocking probability increases as the load on the network increases. Close agreement is observed between the analytic and simulated results, all of which overlap well within one standard deviation for every data point. This is expected due to the insensitivity result (Theorem 2) and supports our analysis. The tightness of the overlap between each simulated data set and the analytic results is captured by the absolute relative errors, defined by (14). These errors are $< 1\%$ for each simulation type. We validate the other service models in the same way and obtain error parameters of $(\delta_{\text{discrete}}, \delta_{\text{exponential}}, \delta_{\text{Cox}}) = (0.061, 0.015, 0.006)$ for the multiple EPR pair service model and $(0.001, 0.003, 0.004)$ for the jump-over service model. To further validate our analysis of the jump-over service model, Figure 8 compares the average probability that the first call of a session is blocked with the probability that any call which follows an idle period (retrial) is blocked. The comparison is made using results from discrete simulations, however similar results are obtained from exponential and Cox simulations. As predicted by our analysis (Theorem 3), the blocking probability is equal for any arbitrary call and does not depend on whether a call is the first one of a session or follows an idle period (retrials). This property simplifies application of our analytic results, as only one formula needs to be evaluated.

The strict reservation service models (strict single, strict multiple), where the batches of entanglement generation attempts in a session are separated by calibration periods during which nodes retain EGS resources, and the jump-over service model, where these batches of attempts are separated by idle periods during which nodes release EGS resources, are first contrasted in Figure 9. There, the performance of the strict single and jump-over service models is reported for various request arrival rates. For the physically motivated simulation parameters used (Table 3) in these evaluations the results for the strict single and strict multiple service models exactly coincide, hence the strict multiple data points are omitted from Figure 9 (and similarly for Figure 10). To facilitate meaningful comparison, the mean duration of the calibration periods of the strict reservation service models and the idle periods of the jump-over service model are all set equal to 1 ms (the same value as in Table 3). Then, to emphasize the impact of resource relinquishment, the performance of the service models is contrasted for various durations of the calibration or idle periods in Figure 10. In that comparison the load on the network was set equally for each service model, by setting the rate of request arrivals from each flow to a value of 0.587 per second. The data in Figures 9 and 10 is obtained from discrete simulations, but exponential and Cox simulations were also conducted. The absolute relative errors as defined by (14) were $\ll 1\%$ for all simulation types of each service model.

For the EGS configuration investigated, the jump-over service mode makes more efficient and productive use of the EGS resources than the strict reservation service modes. With respect to the performance under varied load (Figure 9), when there is a relatively low-load on the EGS, the difference between each performance metric of the jump over and strict reservation service models is marginal. When there is high load on the EGS, the jump-over service model results in lower blocking probabilities, indicating better handling of the high load. The lower blocking probability of the jump-over service model is reflected in the increased idle time ratio. Interestingly, although the EGS resources are idle for a greater proportion of the time in the jump-over mode, a greater total amount of entanglement is produced. The pos-

itive impact of resource relinquishment on the total amount of entanglement produced is further exposed when the calibration and idle period durations are varied (Figure 10). Regardless of the service model, the greatest amount of entanglement produced in a fixed time occurs when the mean duration of calibration or idle periods are zero. As the mean duration of calibration periods in the strict reservation service modes increases, the amount of entanglement produced in a fixed time decreases dramatically and the blocking probability approaches 1 asymptotically. In the strict reservation models, these effects follow from sessions retaining use of the EGS resources for the long durations of calibration periods, preventing the generation of entanglement by other sessions during these periods and resulting in blocking of any requests that arrive during these periods. In contrast, as the mean idle or calibration period duration increases, in the jump-over service mode there is a significantly lower rate of decrease in the total amount of entanglement produced. In complement, the request blocking probability decreases towards 0 asymptotically as the mean duration of an idle period increases. In the jump-over service model, these effects arise because requests for sessions arriving during long idle periods of an active session may be accepted. An increase in the number of active sessions contributes positively to the rate of entanglement generation, partially compensating for the lack of entanglement produced during long idle periods.

For the physically motivated simulation parameters used (Table 3), the performance of the strict multiple service model is not significantly different from that of the strict single service model. With these parameters, the expectation value of successful entanglements per session is 0.01 EPR pairs.

The differences between the strict single and strict multiple service models are more evident if the probability that an attempt to generate entanglement succeeds is significantly increased. In Figure 11 we compare the blocking probabilities resulting from discrete simulation of these service modes when $p_{\text{gen}} = 0.001$, an increase of two orders of magnitude relative to our baseline simulation parameters. The expectation value of successful entanglements per session is 1 EPR pair. In this setting, the strict single service model realizes a much lower blocking probability than the strict multiple service model. This difference can be understood as resulting from a difference in the mean service times of the sessions. The recorded mean service time of sessions in the strict multiple service model is 125 ms. In the strict single service model however, sessions may end earlier if an attempt at entanglement generation succeeds, resulting in a lower observed mean service time of 78 ms. The shorter mean service time of sessions in the strict single service model has the effect that, for the same session request arrival rate, resources are free more often, leading to a lower blocking probability.

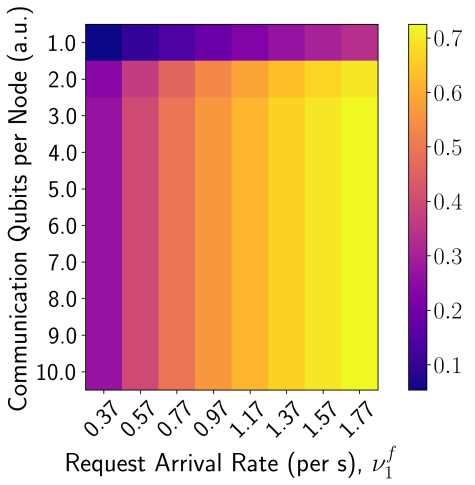
To investigate the impact of the restrictions on communication qubits, we numerically evaluated the blocking probability for an EGS controlled by the strict single service model with two or three resources (Figure 12) as the restriction varies from one to ten communication qubits per node. In each case, for any fixed request rate we observe that an increase in the number of communication qubits per node up to the number of EGS resources results in a large increase in the blocking probability, but further increases in the number of communication qubits have little impact. Numeric evaluation for strict service model scenarios where an EGS with one, two, or three resources is connected to twenty nodes and serves $\binom{20}{2} = 190$ flows (see Appendix I) confirm that the same effect holds for an EGS serving a large number of flows. Recall that a flow does not submit a request if one or more of the nodes of the flow does not have a communication qubit available. Hence for any fixed request rate, a lower blocking probability is expected if nodes do not have enough communication qubits available, due to the lowered effective request rate. As a result of these properties and the observed trends, we conclude that for the EGS configurations studied, when each node is restricted to a number of communication qubits less than or equal to the number of EGS resources (one or two in these configurations), resource contention at the EGS has less of an impact on the blocking probability than does a lack of available communication qubits. In contrast, when nodes are equipped with a number of communication qubits in excess of the number of EGS resources (three or more in these configurations), contention for EGS resources becomes the limiting factor affecting the blocking probability.

Description	Value
Link lengths: users $1 - \lceil K/2 \rceil$	10 km
Link lengths: users $(\lceil K/2 \rceil + 1) - K$	20 km
Single-mode optical fiber attenuation coefficient, σ_{att}	0.2 dB/km
One-way communication time, S_2 (RTT/2)	100.07 μs
Resource allocation duration, S_1	125 ms
Resource allocation duration, S_2	240 ms
Probability photon travelling 1 km optical fiber arrives, $p_a = 10^{-(\sigma_{\text{att}}/10)}$	0.95499 (a.u.)
Probability of single attempt success in S_2: $p_{\text{gen}} \cdot p_a^{10}$	6.31e-6 (a.u.)
Duration of a single attempt, S_2	230.146 μs

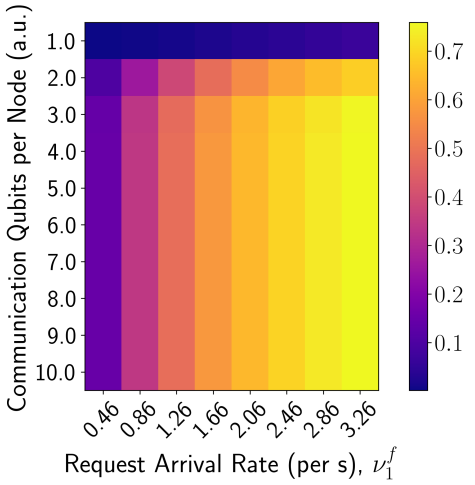
Table 4: Additional physical parameter settings for a non-homogeneous traffic scenario where half of the nodes have doubled link lengths to the EGS. These parameters supplement those of Table 3.

B. NON-HOMOGENEOUS TRAFFIC

Non-homogeneity resulting from network topology will be relevant in any deployed network, as it is unrealistic to expect that nodes may be located in a perfect disk with fixed link lengths to an EGS. We examine a situation where non-homogeneous traffic in the strict single service model results from a network topology where half of the nodes are connected to the EGS by links of length 10 km and the other half of the nodes, are connected by links of double the length, i.e. 20 km. Flows are partitioned into two sets, labelled as *set*



(a) two EGS Resources ($C = 2$)



(b) three EGS Resources ($C = 3$)

Figure 12: Heatmaps of the average blocking probability per flow when the number of communication qubits per node and the request arrival rates are varied. Data results from numeric evaluation of (7) for an EGS with eight nodes and serving $\binom{8}{2} = 28$ flows, one for each possible node pairing. Session traffic is homogeneous.

1 (S_1) and *set 2* (S_2). For every flow in S_2, at least one of the nodes is connected to the EGS by a 20 km link. For every flow in S_1, both nodes are connected to the EGS by 10 km links. The length of a link between a node and the EGS affects the RTT and the probability that a photon sent over the link arrives at the EGS. Table 4 recounts the physical parameters for each set of nodes. See Appendix H for further discussion of the simulation implementation details. Note that within each set of flows, traffic is homogeneous.

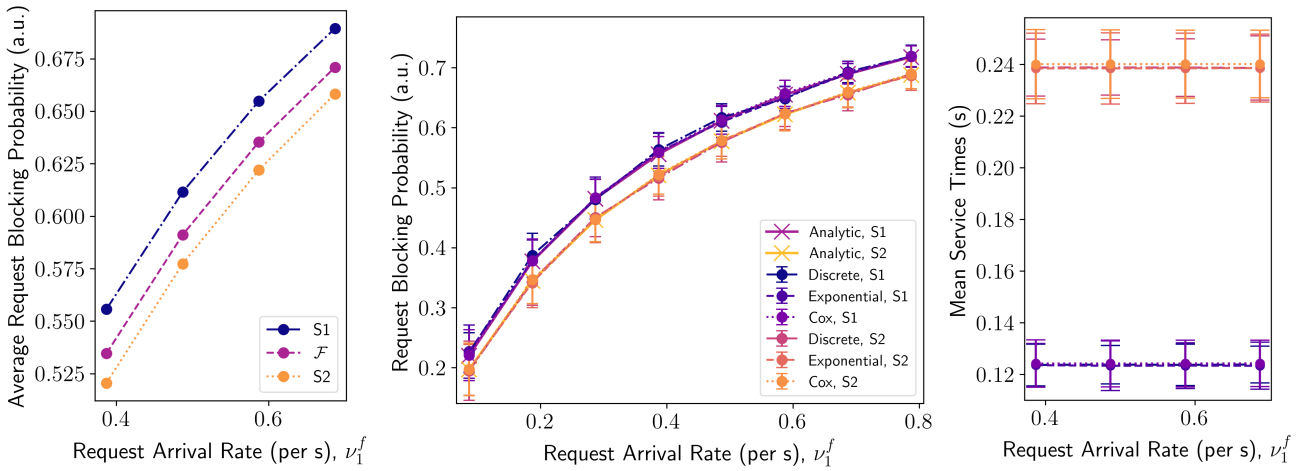
Figure 13 demonstrates that an arbitrary flow receives different service depending on whether it is an element of set S_1 or S_2. Requests from flows in S_1 are blocked with a higher probability and have shorter mean service times than flows in S_2. The blocking probability of any arbitrary flow f from the set of total flows $\mathcal{F}$, indicated in Figure 13a, is the average request blocking probability as given by (11). For the scenario investigated, the blocking probability of an arbitrary flow is greater than for flows in S_2 but less than for flows in S_1. The gap between the blocking probability for flows in S_1 and S_2 is due to the restriction to a single communication qubit per node and the difference in service times of sessions for flows in each set. Although flows from S_1 and S_2 have equal request rates, due to the longer mean service times of flows in S_2, one or more nodes from flows in S_2 will not have a free communication qubit more often than for flows in S_1. As a result, the effective request rate and hence the blocking probability is lower for flows in S_2. The absolute relative errors between the average blocking probability of each simulation type and the analytic blocking probability are $< 4\%$ ($< 6\%$) for flows in S_1 (S_2). These results validate (7) in the presence of non-homogeneity and highlight the complex interplay between the number of communication qubits at a node, mean service times of a session, and blocking probability of an arriving session.

VII. CONCLUSION AND FUTURE WORK

We have proposed an on-demand resource allocation algorithm for an EGS and developed its performance analysis in a variety of traffic scenarios and operation modes. We modeled the system as an Erlang loss one, and discovered an insensitivity property for the request blocking probability. Numerical results validate our analysis. Our work can lead to several new research directions; we provide a non-exhaustive list here. The analytic and simulation frameworks we provide are valuable tools for the development of load-balancing control algorithms for an EGS, which could run at a higher level in the control stack to ensure stable quality of service can be delivered to flows. An important highlight of our model is that it flexibly incorporates restrictions that are very present in NISQ era quantum devices, hence being relevant for the development of a real near-term network. This feature of the model can be used as a tool to investigate efficient resource provisioning schemes – not only for a single EGS serving a number of nodes in a star topology, but also for a more complex network made up of heterogeneous devices. In future work, one could allow a flow to have control over multiple communication qubits. This generalization, along with some of the modeling challenges that might be encountered, were discussed briefly in Section IV. Another natural extension of our model is to generalize from bipartite to multipartite entangled states. This can be accomplished by expanding the definition of a "flow" to include $k \geq 2$ nodes. Last, one could study various back-off mechanisms in lieu of blocking, and determine suitable hardware regimes for each of the schemes.

APPENDIX.

A. PHASE-LEVEL SYSTEM MODEL



(a) (b) (c)

Figure 13: Non-homogeneous traffic, strict single service model: Four nodes are connected to an EGS by 10 km links and four by 20 km links. The EGS has 1 resource. The set $\mathcal{F}$ is partitioned into sets S_1 and S_2. For every flow in S_2, at least one of the nodes is connected to the EGS by a 20 km link. For every flow in S_1, both nodes are connected to the EGS by 10 km links. Traffic within each set is homogeneous. (a) Analytic blocking probability for flows within sets S_1 and S_2 compared to that of an arbitrary flow $f \in \mathcal{F}$, as given by (11). (b) Average request blocking probabilities within sets S_1 and S_2 from discrete, exponential, and Cox simulations compared with numeric evaluation of (7). The absolute relative errors are $(\delta^{S_1}_{\text{discrete}}, \delta^{S_1}_{\text{exponential}}, \delta^{S_1}_{\text{cox}}) = (0.027, 0.032, 0.0032)$ and $(\delta^{S_2}_{\text{discrete}}, \delta^{S_2}_{\text{exponential}}, \delta^{S_2}_{\text{cox}}) = (0.054, 0.0088, 0.0064)$. (c) Comparison of mean service times for sessions in S_1 with those in S_2. Legend is the same as in (b).

1) Motivation for Phase-Level Modeling

In the physical systems which our model describes, attempt, calibration, and idle periods may have generally distributed durations. As discussed in Section II, phase-type Coxian distributions can approximate generally distributed periods by a sequence of exponential phases that preserves the mean period duration. For examples of the distributions determining the duration of attempt and calibration periods, see the discussion based on experiments with the NV center in diamond at the beginning of Section VI.

To derive the blocking probability for each service model (Theorems 1 and 3) and prove our insensitivity result (Theorem 2), we construct a Continuous Time Markov Chain (CTMC) model of the EGS system. This appendix defines the state space of the system for each service model, defines the traffic intensities of each flow with phase-level detail, and defines notation required to prove the results discussed in Section V.

2) Phase-Type (Coxian) Representation

A Coxian distribution with N phases approximates a general distribution by decomposing it into N exponential stages traversed sequentially, with the possibility of exiting the sequence early, based on a transition probability between phases. Each period type in our model is represented this way:

- An attempt period is decomposed into N_A^f exponential phases;
- A calibration period is decomposed into N_C^f exponential phases;
- An idle period is decomposed into N_I^f exponential phases.

We denote individual phases as:

- $A_{i,j}^f$: the i^{th} phase of attempt period j, for a session of flow f, where $i \in \{1, \cdots, N_A^f\}$ and $j \in \{1, \cdots, M_A^f\}$;
- $C_{i,j}^f$: the i^{th} phase of calibration period j, for a session of flow f, where $i \in \{1, \cdots, N_C^f\}$ and $j \in \{1, \cdots, M_C^f\}$;
- $I_{i,j}^f$: the i^{th} phase of idle period j, for a session of flow f, where $i \in \{1, \cdots, N_C^f\}$ and $j \in \{1, \cdots, M_C^f\}$.

The total number of phases in a session of flow type f is

$$L^f \equiv N_A^f \times M_A^f + N_C^f \times M_C^f + N_I^f \times M_I^f. \tag{15}$$

The mean duration of the i^{th} phase of a session for flow type f, where $i \in \{1, \cdots, L\}$ is denoted $1/\mu_i^f$. Sometimes it is convenient to refer instead to the "service rate" or rate of completing the i^{th} phase of a session of flow type f, which is simply μ_i^f. If we refer to the service rate of the i^{th} phase of the j^{th} attempt/calibration/idle period of a session of flow f, the range for i changes to $\{1, \cdots, N_{A/C/I}^f\}$, and the notation $\mu_{i,j}^f$ is used instead.

The possibility of exiting the sequence of phases comprising a period j early, after the i^{th} phase, $i \in \{1, \cdots, N_A^f\}$, is described by a transition probability $1 - r_{i,j}^f$. If this occurs, the period ends, and the session transitions to the subsequent period dictated by the service model, or possibly terminates in the case that the session termination condition is met.

The duration $1/\eta_{A,j}^f$ of an attempt period j for a session of flow f, which appears in the main text, can be simply related

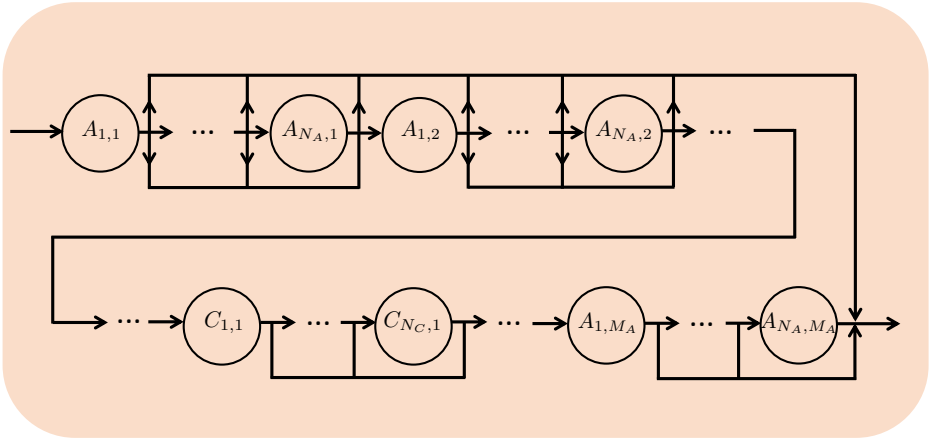


Figure 14: A session from the single EPR pair generation with strict resource reservation service model, shown at the level of periods in Figure 5, decomposed into exponentially-distributed phases so as to result in Coxian-distributed attempt and calibration periods.

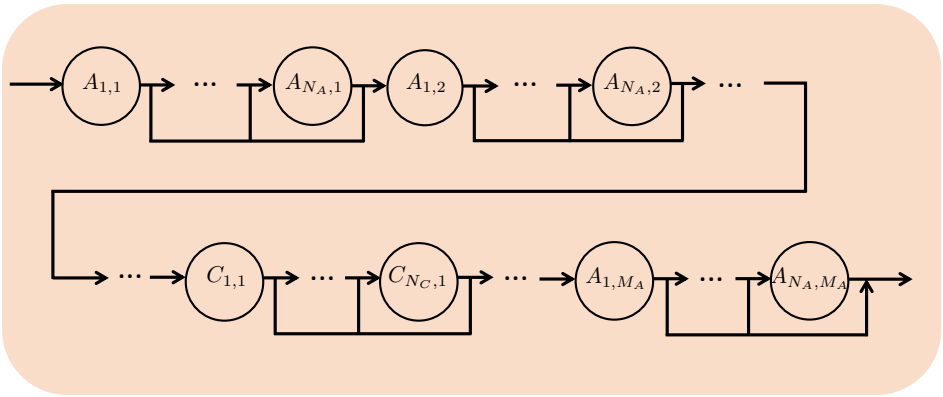


Figure 15: A session from the multiple EPR pair generation with strict resource reservation service model, shown at the level of periods in Figure 5, decomposed into exponentially-distributed phases so as to result in Cox-distributed attempt and calibration periods.

to the duration of the N_A^f exponentially distributed phases which constitute the attempt period,

$$\frac{1}{\eta_{A,j}^f} = \sum_{i=1}^{N_A^f} \frac{r_{i,j}^f}{\mu_{i,j}^f}. \tag{16}$$

Similarly, the same relation holds for the mean duration of calibration and idle periods, with substitution of $1/\eta_{A,j}^f$ and N_A^f for $1/\eta_{C,j}^f$ and N_C^f or $1/\eta_{I,j}^f$ and N_I^f, respectively.

3) Phase-Level State Vector

Irrespective of the service model, the complete state space of the system can be represented using a vector

$$\mathbf{x} = [\mathbf{x}^{f_1}, \ldots, \mathbf{x}^{f_F}], \tag{17}$$

where F is the number of possible flows, and each $\mathbf{x}^f = [x_1^f, \cdots, x_{L_f}^f]$ tracks the number of sessions in each phase for flow f. Specifically, x_i^f is the number of sessions currently in phase i for flow type f. The dimension of this vector $\mathbf{x}$ is

$$L \equiv \sum_{f \in \mathcal{F}} L^f.$$

When we need to identify the specific period and phase within it, we write $x_{i,j}^{f,A}$ for the number of sessions in phase i of attempt period j for flow f (and similarly for calibration and idle periods). This notation is primarily used with reference to transition rates between phases and periods.

We define $\mathbf{e}_i^f, i \in \{1, \cdots, L\}$ to be a vector of dimension L with all entries zero except the one corresponding to the i^{th} component of $\mathbf{x}^f$, which is equal to one. We also use the notation $\mathbf{e}_{i,j}^{f,A}$ when specifically referring to phase i of attempt period j (and similarly for calibration and idle periods).

The flow-level state vector $\mathbf{q}$ used in Sections IV and V can be simply related to the phase-level state vector $\mathbf{x}$. The

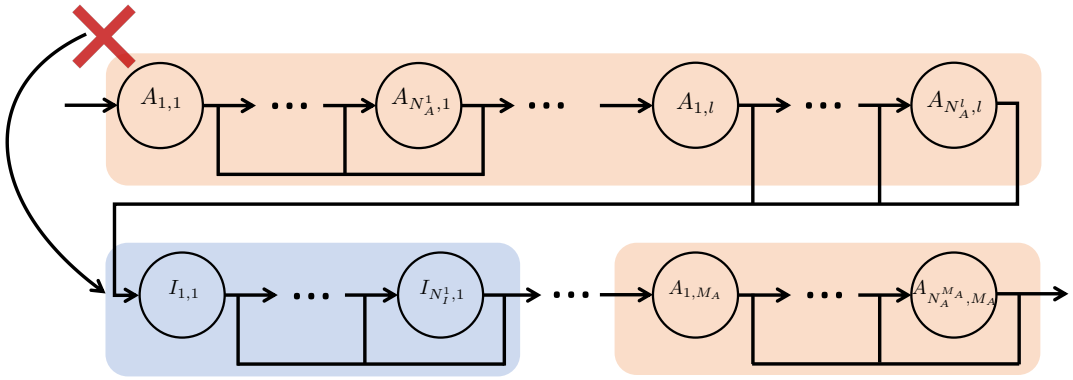


Figure 16: A session with jump-over blocking, shown at the level of periods in Figure 6, decomposed into exponentially-distributed phases so as to result in Cox-distributed active and idle periods. Transitions to outside the queueing network are permitted only from the last period of the session.

number of active sessions of flow f when the system is in state $\mathbf{x}$ is

$$q^f(\mathbf{x}^f) = \sum_{i=1}^{L^f} x_i^f. \tag{18}$$

For service models with resource relinquishment, we distinguish between the total number of sessions and the number of active sessions, which are those currently making active use of EGS resources. The number of active sessions of flow f when the system is in state $\mathbf{x}$ is

$$\tilde{q}^f(\mathbf{x}) \equiv \sum_{j=1}^{M_A^f} \sum_{i=1}^{N_A^f} x_{i,j}^{f,A}. \tag{19}$$

4) State Space Constraints

The admissible state space depends on the service model.

(i) *Strict Resource Reservation Models:*
In all service models, the EGS must heed its resource capacity constraint (C total resource modules) and each network node must heed its own communication qubit limit (c_k communication qubits at node n_k). The following considerations apply to both the Single EPR Pair Generation with Strict Resource Reservation (Section IV-C1) and Multiple EPR Pair Generation with Strict Resource Reservation (Section IV-C2) service models. In these service models, all active sessions (including those in calibration periods) occupy EGS resources. Every session of flow type f occupies one communication qubit from each node in flow f. The admissible state space is thus given by

$$\mathcal{S} = \Big\{ \mathbf{x} \in \mathbb{N}^L : \sum_{f \in \mathcal{F}} q^f(\mathbf{x}^f) \le C, \sum_{f \in \mathcal{F}: n_k \in f} q^f(\mathbf{x}^f) \le c_k, \forall k \in \{1, \dots, K\} \Big\}, \tag{20}$$

where we recall that the notation $n_k \in f$ means that node n_k partakes in flow f.

(ii) *Resource Relinquishment Model*:
The following considerations apply to the Multiple EPR Pair Generation with Resource Relinquishment (Section IV-C3) service model. As with the strict resource reservation service models, each session of flow type f occupies one communication qubit from each node in flow f. In this service model however, only sessions in active (non-idle) phases occupy EGS resources. The admissible state space is thus given by

$$\mathcal{S}'' = \Big\{ \mathbf{x} \in \mathbb{N}^L : \sum_{f \in \mathcal{F}} \tilde{q}^f(\mathbf{x}) \le C, \sum_{f \in \mathcal{F}: n_k \in f} q^f(\mathbf{x}^f) \le c_k, \forall k \in \{1, \dots, K\} \Big\}, \tag{21}$$

where $\tilde{q}^f(\mathbf{x})$ is given by (19).

Both state spaces are coordinate convex, i.e., if $\mathbf{x} \in \mathcal{S}$ (or $\mathcal{S}''$), then $\mathbf{y} \in \mathcal{S}$ (or $\mathcal{S}''$) for all $\mathbf{y}$ such that $\mathbf{0} \le \mathbf{y} \le \mathbf{x}$ component-wise.

5) Traffic Intensities

Each phase i of flow type f, $i \in \{1, \cdots, L\}$ has a service rate μ_i^f, and an arrival rate of $\lambda_i^f(\mathbf{x})$, which in general may depend on the state $\mathbf{x}$ of the system. We sometimes write $\lambda_i^f \equiv \lambda_i^f(\mathbf{x})$, when the state vector $\mathbf{x}$ is clear from context. The traffic intensity for phase i of flow type f is

$$\rho_i^f \equiv \frac{\lambda_i^f}{\mu_i^f}. \tag{22}$$

6) Transitions in the Single EPR Pair Generation with Strict Resource Reservation Service Model

In the Single EPR Pair Generation with Strict Resource Reservation service model, the following properties apply:

- All external arrival rates (*i.e.*, those originating from outside of the network) $\nu_i^f(\mathbf{x})$ are zero, except for those to the first phase of the first attempt period of a session $A_{1,1}^f$, $f \in \mathcal{F}$. We denote these rates with $\nu_1^f(\mathbf{x})$, $f \in \mathcal{F}$,

$\mathbf{x} \in \mathcal{S}$, so that the transition from $\mathbf{x}$ to $\mathbf{x} + \mathbf{e}_{1,1}^{f,A}$ occurs with rate

$$\nu_1^f(\mathbf{x}) = \begin{cases} \nu_1^f, & \text{if } \mathbf{x} + \mathbf{e}_{1,1}^{f,A} \in \mathcal{S}, \\ 0, & \text{else.} \end{cases} \tag{23}$$

- Transition $\mathbf{x} \to \mathbf{x} - \mathbf{e}_i^f + \mathbf{e}_{i+1}^f$ occurs with probability $p_{i,i+1}^f$, for $1 \leq i < L^f$.
- A special case of the above is that $p_{i,i+1}^f = 1, \forall f \in \mathcal{F}$, if x_i^f corresponds to the last phase of a calibration period.
- Transition $\mathbf{x} \to \mathbf{x} - \mathbf{e}_i^f + \mathbf{e}_j^f$ occurs with probability $p_{i,j}^f$ if j is such that x_j^f corresponds to the initial phase of the attempt/calibration period that follows the attempt/calibration period corresponding to x_i^f.
- Transition $\mathbf{x} \to \mathbf{x} - \mathbf{e}_i^f$ occurs with probability p_i^f if i is such that x_i^f corresponds to an attempt phase. This transition represents the event that entanglement generation succeeds after the ith phase of flow f – in this service model, this causes the session to end. We note that leaving the session from a calibration phase is not possible.

When $i = 1$, i.e., for the first phase of a session, the total arrival rate is $\lambda_1^f(\mathbf{x}) = \nu_1^f(\mathbf{x})$. For all other phases, the arrival rate is given by

$$\lambda_i^f(\mathbf{x}) = \begin{cases} \nu_1^f \tilde{p}_i^f \equiv \lambda_i^f, & \text{if } \mathbf{x} + \mathbf{e}_i^f \in \mathcal{S}, \\ 0, & \text{else.} \end{cases} \tag{24}$$

Above, $\tilde{p}_i^f$, $2 \leq i \leq L^f$, denotes the probability of reaching the ith phase starting from the first phase of a session belonging to flow f. Appendix B provides a derivation of these probabilities. In the main text the quantities $\mathcal{P}_{A/C/I,j}^f$ refer to the probability of reaching the j^{th} attempt/calibration/idle period. This probability is equivalent to the probability of reaching the first phase of the j^{th} attempt/calibration/idle period.

A session from the Single EPR Pair Generation with Strict Resource Reservation service model is illustrated with phase level detail in Figure 14. In this service model, a session may terminate (transition out of the system) following any phase of an attempt period. In the figure, these transitions are illustrated by upward arrows linked to system exit.

7) Transitions in the Multiple EPR Pair Generation with Strict Resource Reservation Service Model

In the Multiple EPR Pair Generation with Strict Resource Reservation service model, each session, if admitted for service, traverses all periods (albeit, not necessarily all phases). Consequently, as shown in Figure 15, transitions to outside of the set of active sessions are only permitted from the final attempt period of the session. The overall system is thus very similar to the service model of Appendix A6, and all previous assumptions hold, with the exception that the probability p_i^f of exiting the set of active sessions in the network following a phase i is $p_i^f = 0$, $\forall f \in \mathcal{F}$, whenever i belongs to a phase other than that of the last attempt period. Recall from Section IV that by convention a session always ends with an attempt period.

8) Transitions in the Multiple EPR Pair Generation with Resource Relinquishment Service Model

In the Multiple EPR Pair Generation with Resource Relinquishment service model, a session that had initially been admitted for service by the EGS may get blocked later on, depending on the state of the system $\mathbf{x}$ at the moment the session leaves an idle period. Thus, to define the traffic characteristics within a session, we require transition probabilities that are functions of the state. Let $p_{i,j}^f(\mathbf{x})$ be the probability of transitioning from the ith phase of a type f session to its j^{th} phase. Since in this service mode, a session can only end during its last period, the corresponding model is most similar to the "multiple EPR pair generation with strict resource reservation" scenario. Thus, all traffic characteristics of Appendix A7 apply (i.e., $p_{i,j}^f(\mathbf{x}) = p_{i,j}^f$, $\forall f, i, j, \mathbf{x}$), with the following exceptions:

- If j is the starting phase of an active period (excluding the first period of a session), i is any phase of the preceding idle period, and $\mathbf{x} - \mathbf{e}_i + \mathbf{e}_j \notin \mathcal{S}''$, then $p_{i,j}^f(\mathbf{x}) = 0$;
- If moreover k is the starting phase of the next idle period, then $p_{i,k}^f(\mathbf{x}) = p_{i,j}^f$.

These amendments describe the jump-over blocking dynamics and are illustrated in Figure 17.

The external arrival rates into the system are zero for all phases, except for the first of every session; these are given by $\nu_1^f(\mathbf{x}) = \nu_1^f$ if $\mathbf{x} + \mathbf{e}_1^f \in \mathcal{S}''$, and zero otherwise. Letting γ_i^f be the probability that a session belonging to flow f reaches its i^{th} phase starting from its first phase, we have that the total arrival rate into phase i for flow f while in state $\mathbf{x}$ is given by

$$\lambda_i^f(\mathbf{x}) = \begin{cases} \nu_1^f \gamma_i^f \equiv \lambda_i^f, & \text{if } \mathbf{x} + \mathbf{e}_i^f \in \mathcal{S}'', \\ 0, & \text{else.} \end{cases} \tag{25}$$

Let us examine why λ_i^f has no dependence on the state $\mathbf{x} \in \mathcal{S}''$. For the following, suppose $\mathbf{x} + \mathbf{e}_i^f \in \mathcal{S}''$. First, consider i to be any phase of an active period, and j to be the first phase of the same period. Then $\gamma_i^f = p_{j,j+1}^f \cdots p_{i-1,i}^f$. Next, let i be the first phase of an idle period. The arrival rate into this phase is ν_1^f, regardless of whether the transition is happening from the preceding active period, or from the previous idle period. The latter would happen if the system was at capacity (the EGS did not have enough resource modules) at the time of the transition, thereby causing the next active period to be skipped (along with any other active periods that immediately follow it). The new routing rules introduced above ensure that the arrival rate into the idle period is equal to that of the period(s) being jumped over. The arrival rate into any other phase of an idle period is then computed similar to that of an active period's non-initial phase.

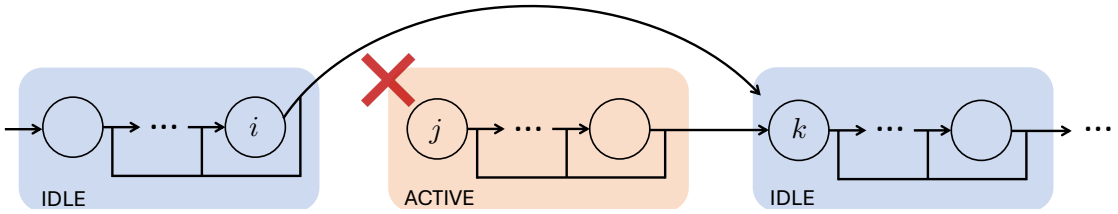


Figure 17: Illustration of jump-over blocking: if a transition to an active period is not possible due to a state space restriction, the active period is skipped, and a transition occurs to the first phase of the next idle period.

B. PROBABILITY OF REACHING A PHASE IN SINGLE EPR PAIR GENERATION MODE

In this appendix, we derive an expression for $\tilde{p}_i^f$ as seen in (24), the probability of reaching the i^{th} phase of a type f session, starting from its first phase, while in strict resource reservation, single EPR pair generation service mode. Suppose that the i^{th} phase is located within the j^{th} period of the session. To have reached this phase, the session must not have ended during any of the previous attempt phases (recall that in this service mode, sessions retain EGS resources during calibration periods). In the following, let $p_{i,j}^k$ denote the probability of transitioning from the i^{th} phase of period k to the j^{th} phase of the same period; we omit f below for cleanliness. Suppose that $p_{\text{gen}}^{\text{cox}}$ denotes the probability with which a BSM is successful under Coxian distributions. Then the probability of leaving during an attempt period k due to success in generating entanglement, conditioned on the event that the session has entered period k, is given by

$$P_k \equiv \mathrm{P}\left(\text{leave during attempt period } k \middle| \text{session entered period k}\right) \tag{26}$$

$$= p_1^k + p_{1,2}^k p_2^k + \cdots + (p_{1,2}^k \cdots p_{N^k-1,N^k}^k p_N^k) \tag{27}$$

$$= p_{\text{gen}}^{\text{cox}}, \tag{28}$$

where $p_l^k = (1 - p_{l,l+1}^k) p_{\text{gen}}^{\text{cox}}$ is the probability of the session leaving the system after phase l of period k due to a successful BSM at the EGS, and $p_{l,l+1}^k$ is the probability of transitioning to the next phase of the period. Then, the probability of the session reaching the i^{th} phase of the j^{th} period is given by

$$\mathrm{P}\big(\text{session reaches } i^{\text{th}} \text{ phase of period } j, \text{ starting from the } 1^{\text{st}} \text{ phase of the session}\big) \tag{29}$$

$$= \prod_{l=1}^{i-1} p_{l,l+1}^j \prod_{k=1}^{j-1} (1 - P_k)\,, \tag{30}$$

where for any k that corresponds to a calibration period $P_k = 0$.

C. DERIVATION OF THE STATIONARY DISTRIBUTION FOR EACH SERVICE MODEL

1) Single EPR Pair Generation with Strict Resource Reservation

Theorem 4. *The stationary distribution $\pi(\mathbf{x})$ of the system with single EPR pair generation while in strict resource reservation mode is given by*

$$\pi(\mathbf{x}) = \left\{ \left(\sum_{\mathbf{y} \in \mathcal{S}} \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{y_i^f}}{y_i^f!} \right)^{-1} \right\} \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{x_i^f}}{x_i^f!}, \tag{31}$$

where ρ_i^f is the traffic intensity of the i^{th} phase of a session corresponding to flow f.

Proof. To obtain the stationary probability $\pi(\mathbf{x})$ of the system with state space $\mathcal{S}$ as defined in (20), we apply the local balance approach. Henceforth, we adopt the convention that $\pi(\mathbf{x}) = 0$ for any $\mathbf{x} \notin \mathcal{S}$. Consider any state $\mathbf{x} \in \mathcal{S}$; we have that

- The rate of leaving $\mathbf{x}$ due to outside arrival coming into the system is given by
$$A = \pi(\mathbf{x}) \sum_{f \in \mathcal{F}} \nu_1^f(\mathbf{x}); \tag{32}$$
- The rate of entering $\mathbf{x}$ due to job departure to outside of the system is given by
$$A' = \sum_{f \in \mathcal{F}} \sum_{i=1}^{L^f} \pi(\mathbf{x} + \mathbf{e}_i^f)(x_i^f + 1) p_i^f \mu_i^f, \tag{33}$$
where we take advantage of our convention that $\pi(\mathbf{x} + \mathbf{e}_i^f) = 0$ if $\mathbf{x} + \mathbf{e}_i^f \notin \mathcal{S}$, as well as use the fact that $p_i^f = 0$ if i corresponds to a calibration phase of a session belonging to flow f.
- The rate of leaving $\mathbf{x}$ due to departure from phase i of flow f is given by
$$B_i^f = \pi(\mathbf{x}) x_i^f \mu_i^f; \tag{34}$$
- The rate of entering $\mathbf{x}$ due to an arrival at phase i of flow f is given by
$$B_i'^f = \pi(\mathbf{x} - \mathbf{e}_i^f)\nu_i^f(\mathbf{x} - \mathbf{e}_i^f) + \sum_{j \neq i} \pi(\mathbf{x} + \mathbf{e}_j^f - \mathbf{e}_i^f)(x_j^f + 1) p_{j,i}^f \mu_j^f. \tag{35}$$

To obtain the stationary distribution of the system, we solve the following equations and then show that this solution is in fact the stationary distribution of the system,

$$\pi(\mathbf{x} + \mathbf{e}_i^f) = \frac{\lambda_i^f(\mathbf{x})}{\mu_i^f(x_i^f + 1)} \pi(\mathbf{x}) \tag{36}$$

$$= \begin{cases} \frac{\rho_i^f}{x_i^f+1}\pi(\mathbf{x}), & \text{if } \mathbf{x}+\mathbf{e}_i^f \in \mathcal{S}, \\ 0, & \text{else} \end{cases} \tag{37}$$

where $\rho_i^f \equiv \lambda_i^f/\mu_i^f$. Substituting this expression into $A = A'$ and simplifying, we obtain

$$\sum_{f\in\mathcal{F}} \nu_1^f(\mathbf{x}) = \sum_{f\in\mathcal{F}} \sum_{i=1}^{L^f} \lambda_i^f(\mathbf{x}) p_i^f. \tag{38}$$

The expression above is the traffic conservation equation: the aggregate arrival rate into the system while in state $\mathbf{x}$ equals the overall departure rate from it. Continuing the application of local balance, we require that for all i and f, $B_i^f = B_i'^f$. Note that if $\mathbf{x} = \mathbf{0}$, then both B_i^f and $B_i'^f$ are zero, since we cannot leave this state due to a departure from any phase, and we cannot enter this state due to an arrival to any phase, respectively, since there are no active sessions while in state $\mathbf{0}$. For any other $\mathbf{x} \in \mathcal{S}$, we require

$$\begin{aligned} \pi(\mathbf{x}) x_i^f \mu_i^f =& \pi(\mathbf{x}-\mathbf{e}_i^f)\nu_i^f(\mathbf{x}-\mathbf{e}_i^f) \\ &+ \sum_{j\neq i} \pi(\mathbf{x}+\mathbf{e}_j^f-\mathbf{e}_i^f)(x_j^f+1)p_{j,i}^f\mu_j^f. \end{aligned} \tag{39}$$

Using (37) with (39) and simplifying, we obtain

$$\begin{aligned} \pi(\mathbf{x}-\mathbf{e}_i^f)\lambda_i^f(\mathbf{x}-\mathbf{e}_i^f) =& \pi(\mathbf{x}-\mathbf{e}_i^f)\nu_i^f(\mathbf{x}-\mathbf{e}_i^f) \\ &+ \sum_{j\neq i} \pi(\mathbf{x}-\mathbf{e}_i^f)\lambda_j^f(\mathbf{x}-\mathbf{e}_i^f)p_{j,i}^f. \end{aligned} \tag{40}$$

If $\mathbf{x}-\mathbf{e}_i^f \notin \mathcal{S}$, then $\pi(\mathbf{x}-\mathbf{e}_i^f) = 0$ and the equation above holds; else we obtain

$$\lambda_i^f(\mathbf{x}-\mathbf{e}_i^f) = \nu_i^f(\mathbf{x}-\mathbf{e}_i^f) + \sum_{j\neq i} \lambda_j^f(\mathbf{x}-\mathbf{e}_i^f)p_{j,i}^f. \tag{41}$$

We have thus recovered the traffic equations for each individual phase of the sessions in the network: the total arrival rate into phase i of flow f equals the sum of the exogenous arrival rates and the arrival rates from other phases. Thus our solution satisfying (37) represents the stationary distribution.

Next, from (37), it follows that

$$\pi(\mathbf{x}) = D \prod_{f\in\mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{x_i^f}}{x_i^f!}, \tag{42}$$

where $D = \pi(\mathbf{0})$. Using the fact that $\sum_{\mathbf{x}\in\mathcal{S}} \pi(\mathbf{x}) = 1$, we can solve for D, which results in the stationary distribution (31), where the term inside the curly brackets is $\pi(\mathbf{0})$. Note from (31) that if a subset of flows $\mathcal{F}' \subseteq \mathcal{F}$ never wish to generate requests – i.e., $\nu_1^f(\mathbf{x}) = 0, \forall \mathbf{x} \in \mathcal{S}, f \in \mathcal{F}'$ – then these flows' traffic intensities are equal to 0. If we take the convention that $0^0 \equiv 1$, then (31) supports this setting by considering only flows from $\mathcal{F} \setminus \mathcal{F}'$ when $\mathbf{x} \in \mathcal{S}$, and causing $\pi(\mathbf{x})$ to equal zero whenever $x_i^f \neq 0$ for $f \in \mathcal{F}'$. □

2) Multiple EPR Pair Generation with Strict Resource Reservation

In this service model the stationary distribution is given by (31). The same derivation as in Appendix C1 applies, with the application of the considerations in Appendix A7, which only affect the traffic intensities.

3) Multiple EPR Pair Generation with Resource Relinquishment

We derive the stationary distribution $\pi(\mathbf{x})$, $\mathbf{x} \in \mathcal{S}''$ for the service model with resource relinquishment and jump-over blocking as the retrial mechanism and show that the blocking probability is given by (13). The admissible state space $\mathcal{S}''$ is defined in Appendix A4 and the transitions possible in this service model are detailed in Appendix A8.

To obtain the stationary distribution of this system, we apply the local balance approach as in Appendix C1. The rate of leaving state $\mathbf{x} \in \mathcal{S}''$ due to an outside arrival is

$$A = \pi(\mathbf{x}) \sum_{f\in\mathcal{F}} \nu_1^f(\mathbf{x}), \tag{43}$$

and the rate of entering $\mathbf{x}$ due to termination of a session is

$$A' = \sum_{f\in\mathcal{F}} \sum_{i=1}^{L^f} \pi(\mathbf{x}+\mathbf{e}_i^f)(x_i^f+1)\mu_i^f p_i^f(\mathbf{x}+\mathbf{e}_i^f). \tag{44}$$

In (44), we can replace $p_i^f(\mathbf{x}+\mathbf{e}_i^f)$ with p_i^f since if $\mathbf{x}+\mathbf{e}_i^f \notin \mathcal{S}''$, then by our convention $\pi(\mathbf{x}+\mathbf{e}_i^f) = 0$. Recall that if $\mathbf{x}+\mathbf{e}_i^f \in \mathcal{S}''$ but i corresponds to any phase that does not belong to the last period of a session, then $p_i^f = 0$. Then, $A = A'$ yields a traffic conservation equation of the same form as (38), if we let $\pi(\mathbf{x})$ take the same form as (37) (note that the definition of $\lambda_i^f(\mathbf{x})$ is now given by (25) throughout). Local balance for each phase i of a session belonging to flow f yields the expected traffic equations

$$\lambda_i^f(\mathbf{x}) = \nu_i^f(\mathbf{x}) + \sum_{j\neq i} \lambda_j^f(\mathbf{x}) p_{j,i}^f(\mathbf{x}+\mathbf{e}_j^f-\mathbf{e}_i^f). \tag{45}$$

D. PROOF OF THE BLOCKING PROBABILITY THEOREM FOR THE SINGLE EPR PAIR GENERATION WITH STRICT RESOURCE RESERVATION SERVICE MODEL

In this appendix we present a proof of Theorem 1, stated in Section V-A.

Proof of Theorem 1. To begin, let $\mathrm{P}(\mathbf{q})$ denote the probability that EGS resources are occupied according to $\mathbf{q}$. Consider the following two events: $\Omega_1(h)$ is the event that h EGS resources are occupied, and $\Omega_2(i)$ is the event that flow f_i has available communication qubits. By the PASTA (Poisson Arrivals See Time Averages) property, we can write the probability that an arriving request of flow f_i sees h occupied resources (conditioned on the flow having enough qubits to generate a request), i.e., $\mathrm{P}(\Omega_1(h)|\Omega_2(i))$, as

$$\overline{\pi}_i(h) = \frac{\mathrm{P}(\Omega_1(h) \cap \Omega_2(i))}{\mathrm{P}(\Omega_2(i))}$$

$$= \left(\sum_{\mathbf{q} \in \mathcal{Q}'(i)} \mathrm{P}(\mathbf{q}) \right)^{-1} \sum_{\mathbf{q} \in \mathcal{Q}(h) \cap \mathcal{Q}'(i)} \mathrm{P}(\mathbf{q}). \quad (46)$$

Here we use the fact that flow f_i would not be able to initiate a session if any node associated with f_i does not have an unoccupied communication qubit.

The probability that flows occupy EGS resources according to $\mathbf{q}$ is

$$\mathrm{P}(\mathbf{q}) = \sum_{\mathbf{x}:\mathbf{q}(\mathbf{x})=\mathbf{q}} \pi(\mathbf{x}) = \sum_{\mathbf{x}^{f_1}:q^{f_1}(\mathbf{x}^{f_1})=q_1} \cdots \sum_{\mathbf{x}^{f_F}:q^{f_F}(\mathbf{x}^{f_F})=q_F} D \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{x_i^f}}{x_i^f!}, \quad (47)$$

where $\mathbf{q}(\mathbf{x})$ is a vector containing $q^f(\mathbf{x}^f)$ values for $f \in \mathcal{F}$.

Recursive application of the multinomial theorem on (47) results in

$$P(\mathbf{q}) = D \prod_{j=1}^{F} \frac{1}{q_j!} \left(\sum_{i=1}^{L^{f_j}} \rho_i^{f_j} \right)^{q_j} \equiv D \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!}, \quad (48)$$

where $\rho^{f_j} := \sum_{i=1}^{L^{f_j}} \rho_i^{f_j}$. The constant D can be simplified similarly, after the additional step of rewriting the sum over $\mathbf{y} \in \mathcal{S}$ in terms of the total number of occupied EGS resources. Although D cancels out within the blocking probability, we present it in simplified form ($\tilde{D}$) for completeness:

$$\tilde{D}^{-1} = \sum_{\mathbf{y} \in \mathcal{S}} \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{y_i^f}}{y_i^f!} = \sum_{h=0}^{C} \sum_{\mathbf{q} \in \mathcal{Q}(h)} \prod_{j=1}^{F} \frac{(\rho^{f_j})^{q_j}}{q_j!}. \quad (49)$$

Using (48) and (46), we can obtain the blocking probabilities for the system – that is, the probability that an arriving request belonging to flow f_i sees C resources occupied is given by

$$\overline{\pi}_i(C) = \left(\sum_{\mathbf{q} \in \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!} \right)^{-1} \sum_{\mathbf{q} \in \mathcal{Q}(C) \bigcap \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!}, \quad (50)$$

In the above equation the denominator term represents the probability that user-pairs of flow f_i have unoccupied communication qubits. Note that flow f_i triggers sessions only when both of its users have unoccupied communication qubits. □

Remark 1. *For our model the blocking probability for a given flow f is expressed as a conditional probability, as entanglement requests are triggered only when all the associated nodes have sufficient communication qubits. In [29], conditioning is not necessary since flows always trigger requests according to an unrestricted Poisson process. Further, in our model, blocking probabilities can be different for different flows as users may have different number of communication qubits whereas all the flows have the same blocking probabilities in [29].*

Remark 2. *Let us take a closer look at the form of ρ^f, for a given $f \in \mathcal{F}$ and validate that it is consistent with* (3)*, the definition of the overall traffic intensity of flow f:*

$$\rho^f = \sum_{i=1}^{L^f} \rho_i^f = \sum_{i=1}^{L^f} \frac{\lambda_i^f}{\mu_i^f} = \sum_{i=1}^{L^f} \frac{\nu_1^f \tilde{p}_i^f}{\mu_i^f} = \nu_1^f \sum_{i=1}^{L^f} \frac{\tilde{p}_i^f}{\mu_i^f}. \quad (51)$$

Noting that the sum represents the mean duration of a type f session, we see that we recover (3)*, $\rho^f = \nu_1^f \times \mathbb{E}[\text{session duration}]$.*

E. PROOF OF INSENSITIVITY THEOREM FOR SINGLE EPR PAIR GENERATION WITH STRICT RESOURCE RESERVATION

In this appendix, we present a proof of Theorem 2, stated in Section V-A.

Proof of Theorem 2. To show insensitivity to the distributions of periods we will prove that the stationary distribution for the total number of ongoing sessions of flows remains the same when attempts and calibration periods have either Coxian or exponential distributions with the same mean.

We first derive the expressions for the stationary distribution for the case where each attempt and calibration period are exponentially distributed. If M_A^f and M_C^f are the number of attempt and calibration periods, respectively, for flow f (and there are no idle periods in this system), then $M^f \equiv M_A^f + M_C^f$ is the total number of periods in each session of type f. The state of the system can then be described using the vector

$$\mathbf{Z} = \left[\mathbf{Z}^{f_1}, \ldots, \mathbf{Z}^{f_F}\right] = \left[Z_i^f, 1 \le i \le M^f, f \in \mathcal{F}\right], \quad (52)$$

where Z_i^f indicates the total number of ongoing sessions of type f in period i. Let L be the state dimension, i.e., $L \equiv \sum_{f \in \mathcal{F}} M^f$; then the admissible state space for this system is given by

$$\mathcal{S}' = \left\{ \mathbf{Z} \in \mathbb{N}^L : \sum_{f \in \mathcal{F}} \sum_{i=1}^{M^f} Z_i^f \le C, \sum_{f \in \mathcal{F}: n_k \in f} \sum_{i=1}^{M^f} Z_i^f \le c_k, \forall k \in \{1, \ldots, K\} \right\}. \quad (53)$$

Let $1/\theta_j^f$ and $1/\sigma_j^f$ be the average duration of attempt and calibration periods j, respectively, for $f \in \mathcal{F}$. Further, let ω_j^f be the arrival rate into the j^{th} period of a session belonging to flow f. By the results of the previous subsection, we know that the stationary distribution for this system is

$$\pi'(\mathbf{Z}) = \left(\sum_{\mathbf{Y} \in \mathcal{S}'} \prod_{f \in \mathcal{F}} \prod_{i=1}^{M^f} \frac{(\xi_i^f)^{Y_i^f}}{Y_i^f!} \right)^{-1} \prod_{f \in \mathcal{F}} \prod_{j=1}^{M^f} \frac{(\xi_j^f)^{Z_j^f}}{Z_j^f!}, \quad (54)$$

where period j's traffic intensity for flow f is $\xi_j^f = \omega_j^f/\theta_j^f$ if j corresponds to an attempt period, and $\xi_j^f = \omega_j^f/\sigma_j^f$ if it corresponds to a calibration period. Analogous to the Coxian case, $\omega_j^f = \nu_j^f$ if $j = 1$, else it is

$$\omega_j^f = \nu_1^f \zeta_{1,2}^f \cdots \zeta_{j-1,j}^f, \tag{55}$$

where $\zeta_{l-1,l}^f$ is the probability of transitioning from period $l-1$ to period l of a session belonging to flow f. Let ζ_j^f be the probability of leaving the queueing network after the j^{th} queue due to a successful BSM at the EGS for creating an entanglement, and note that in this service mode $\zeta_j^f = 0$ if j corresponds to a calibration period.

To prove insensitivity, we assume that the average duration of a period in the exponential scenario is equal to the average duration of the corresponding period in the Coxian scenario. In other words, we have that for the j^{th} period, depending on whether it is an attempt or calibration period, respectively,

$$\frac{1}{\theta_j^f} = \sum_{i=1}^{N_A^f} \frac{r_{i,j}^f}{\mu_{i,j}^f}, \quad \text{or} \quad \frac{1}{\sigma_j^f} = \sum_{i=1}^{N_C^f} \frac{r_{i,j}^f}{\mu_{i,j}^f}, \tag{56}$$

where $r_{i,j}^f$ denotes the probability of reaching the i^{th} phase of the j^{th} period starting from its initial phase, and $1/\mu_{i,j}^f$ denotes the average duration of the i^{th} phase of period j within a flow-f session.

We further assume that for each attempt period of the exponential scenario, the entanglement success probability equals that of the corresponding attempt period in the Coxian scenario. That is, $\zeta_j^f = P_j^f = p_{\text{gen}}$ in this case, where P_j^f is the probability of leaving the queueing network during the j^{th} period of a flow-f session in the Coxian scenario, as computed in (28). This assumption is physically motivated for scenarios where the mean duration of an attempt is significantly shorter than the timescale of parameter drift affecting a quantum node. For a detailed justification, see the discussion on success probability, Section III-A2.

In the Coxian scenario, we can rewrite the state representation in (17) as follows:

$$\mathbf{x} = [\mathbf{x}_1^{f_1}, \ldots, \mathbf{x}_{M^{f_1}}^{f_1}, \ldots, \mathbf{x}_1^{f_F}, \ldots, \mathbf{x}_{M^{f_F}}^{f_F}], \tag{57}$$

where M^{f_i} is the number of periods in sessions of type f_i. For any $\mathbf{x} \in \mathcal{S}$, let

$$q_j^f(\mathbf{x}_j^f) = \sum_i x_{i,j}^f,$$

i.e., this is the number of sessions in the j^{th} period of sessions belonging to flow f. Then for a given $\mathbf{Z} \in \mathcal{S}'$,

$$\sum_{\mathbf{x}: q_m^f(\mathbf{x}_m^f) = Z_m^f, \forall m, f} \pi(\mathbf{x}) = D \cdot \Bigg(\sum_{\mathbf{x}_1^{f_1}: q_1^{f_1}(\mathbf{x}_1^{f_1}) = Z_1^{f_1}} \cdots \sum_{\mathbf{x}_{M^{f_F}}^{f_F}: q_{M^{f_F}}^{f_F}(\mathbf{x}_{M^{f_F}}^{f_F}) = Z_{M^{f_F}}^{f_F}} \prod_{f \in \mathcal{F}} \prod_{j=1}^{M^f} \prod_{i=1}^{N_j^f} \frac{\left(\rho_{i,j}^f\right)^{x_{i,j}^f}}{x_{i,j}^f!} \Bigg), \tag{58}$$

where D is the normalizing constant of the Coxian distribution (see (31)), N_j^f is the number of phases in the j^{th} period of a session belonging to flow f, and $\rho_{i,j}^f = \lambda_{i,j}^f/\mu_{i,j}^f$ is the traffic intensity in the i^{th} phase of this period. Multiple applications of the multinomial theorem yield

$$\sum_{\mathbf{x}: q_m^f(\mathbf{x}_m^f) = Z_m^f, \forall m, f} \pi(\mathbf{x}) = D \prod_{f \in \mathcal{F}} \prod_{j=1}^{M^f} \frac{1}{Z_j^f!} \left(\sum_{i=1}^{N_j^f} \rho_{i,j}^f \right)^{Z_j^f}. \tag{59}$$

When j corresponds to an attempt period, $N_j^f = N_A^f$, and we have that

$$\sum_{i=1}^{N_j^f} \rho_{i,j}^f = \sum_{i=1}^{N_A^f} \rho_{i,j}^f = \sum_{i=1}^{N_A^f} \frac{\lambda_{i,j}^f}{\mu_{i,j}^f} \tag{60}$$

$$= \nu_1^f \sum_{i=1}^{N_A^f} \frac{\tilde{p}_{i,j}^f}{\mu_{i,j}^f} \tag{61}$$

$$= \nu_1^f \sum_{i=1}^{N_A^f} \frac{\prod_{l=1}^{i-1} p_{l,l+1}^{j,f} \prod_{k=1}^{j-1} \left(1 - P_k^f\right)}{\mu_{i,j}^f} \tag{62}$$

$$= \nu_1^f \prod_{k=1}^{j-1} \left(1 - P_k^f\right) \sum_{i=1}^{N_A^f} \frac{r_{i,j}^f}{\mu_{i,j}^f} \tag{63}$$

$$= \frac{\nu_1^f}{\theta_j^f} \prod_{k=1}^{j-1} \left(1 - \zeta_k^f\right) \tag{64}$$

$$= \frac{\nu_1^f}{\theta_j^f} \prod_{k=1}^{j-1} \zeta_{k,k+1}^f = \xi_j^f, \tag{65}$$

where in the last equality of (61) we use (24), in (62) we use (30); for (63) we use the fact that $\prod_{l=1}^{i-1} p_{l,l+1}^{j,f}$ simply represents the probability of going from the first phase of the period to the i^{th}; (64) follows from our two assumptions on the entanglement generation success probability and mean period duration; and the last equality of (65) follows from (55) and the definition of traffic intensity. We can use a similar argument to show that when j corresponds to a calibration period, the sum within (59) equals $\xi_j^f = \omega_j^f/\sigma_j^f$.

We next address the normalization constant D within (59):

$$D = \sum_{\mathbf{y} \in \mathcal{S}} \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{y_i^f}}{y_i^f!}$$

$$= \sum_{\mathbf{Z} \in \mathcal{S}'} \sum_{\mathbf{x}: q_m^{f'}(\mathbf{x}_m^{f'}) = Z_m^{f'}, \forall m, f'} \prod_{f \in \mathcal{F}} \prod_{j=1}^{M^f} \prod_{i=1}^{N_j^f} \frac{(\rho_{i,j}^f)^{y_{i,j}^f}}{y_{i,j}^f!}, \tag{66}$$

where we have partitioned the state space $\mathcal{S}$ based on the number of jobs in each period of a session. Applications of the multinomial theorem as in (58), followed by utilization of (56), results in D's equivalence to the normalization constant

of the exponential-scenario stationary distribution, see (54). This result, combined with (59), means that

$$\sum_{\mathbf{x}:q_m^f(\mathbf{x}_m^f)=Z_m^f,\forall m,f} \pi(\mathbf{x}) = \pi'(\mathbf{Z}). \tag{67}$$

The above equation shows that the stationary distribution for the total number of ongoing sessions of flows remains the same when attempts and calibration periods have either Coxian or exponential distributions with the same mean. From this, we conclude that the insensitivity property holds. □

F. PROOF OF THE BLOCKING PROBABILITY THEOREM FOR THE MULTIPLE EPR PAIR GENERATION WITH RESOURCE RELINQUISHMENT SERVICE MODEL

In this appendix we provide a proof of Theorem 3. The proof relies on the phase-level system model introduced in Appendix A and the derivation of the stationary distribution of the system in this service model given in Appendix C3.

Proof of Theorem 3. In order to obtain the blocking probability, we must consider not only the external arrival process, but also the internal jump-over blocking mechanism. For the former, we can apply the Poisson-Arrivals-See-Time-Averages (PASTA) property as sessions arrive according to a Poisson process. For the latter, we give the corresponding result in Corollary 2 below that extends the result of "Departures See Time Averages" stated in [29, Corollary 1], which was proven for a slightly different model presented in that manuscript – namely, there each active period is followed by an idle period and a flow generates sessions without any restrictions from users. In our case, both the users of a flow must have one unoccupied communication qubit to submit their request for EGS resource access. Corollary 2 enables us to show that the blocking probability of the first attempt of the session is identical to the blocking probability of its retrials.

Corollary 2. *For the service model introduced in Section V-C, let $\mathbf{x}+\mathbf{e}_i^f \in \mathcal{S}''$, with i corresponding to an idle or calibration phase of the flow f, where $\mathcal{S}''$ is defined in Appendix A4. The probability that a session of flow f departing phase i from state $\mathbf{x}+\mathbf{e}_i^f \in \mathcal{S}''$ sees the system in state $\mathbf{x} \in \mathcal{S}''$, equals the stationary conditional probability that the system is in state $\mathbf{x}$ given that both nodes of the flow f have at least one unoccupied communication qubit.*

Proof. Define $\pi^{i,f}(\mathbf{x})$ as the probability that sessions belonging to flow f leaving an idle phase i see the network in state $\mathbf{x}$ immediately after their departure in the stationary regime. This probability is given by

$$\pi^{i,f}(\mathbf{x}) = \frac{\pi(\mathbf{x}+\mathbf{e}_i^f)\mu_i^f(x_i^f+1)}{\sum_{\mathbf{y}:\mathbf{y}+\mathbf{e}_i^f\in\mathcal{S}''}\pi(\mathbf{y}+\mathbf{e}_i^f)\mu_i^f(y_i^f+1)}, \tag{68}$$

where the numerator represents the transition rate out of state $\mathbf{x}+\mathbf{e}_i^f$ and the denominator represents the transition rate out of all possible states associated with termination of sessions.

For $\mathbf{x}+\mathbf{e}_i^f \in \mathcal{S}''$ (and therefore, by the convexity of the state space, $\mathbf{x} \in \mathcal{S}''$), we know from the stationary distribution analysis that

$$\pi(\mathbf{x}+\mathbf{e}_i^f) = \frac{\rho_i^f}{x_i^f+1}\pi(\mathbf{x}). \tag{69}$$

Using this result with (68) and (25), upon simplifying we obtain

$$\pi^{i,f}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{\sum_{\mathbf{y}:\mathbf{y}+\mathbf{e}_i^f\in\mathcal{S}''}\pi(\mathbf{y})}. \tag{70}$$

□

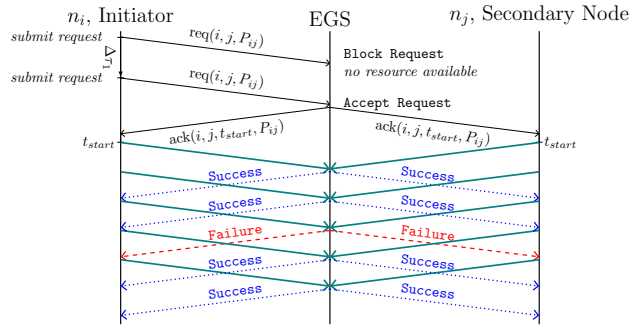
From Corollary 2 and (70), the probability that a job of flow f leaving an idle period finds all EGS rource modules to be occupied is equal to $\frac{\sum_{\mathbf{x}:\mathbf{x}\in\mathcal{Q}(C)\cap\mathcal{Q}'(f)}\pi(\mathbf{x})}{\sum_{\mathbf{y}:\mathbf{y}\in\mathcal{Q}'(f)}\pi(\mathbf{y})}$, which coincides with $\overline{\pi}_f(C)$. Therefore we conclude that for an active period that immediately follows an idle period, the blocking probability is the same as that for the first period of a session (recall that by our convention, the first period is always an active period). □

G. EXTENDED INFORMATION ON PHYSICAL EGS OPERATION PROTOCOLS

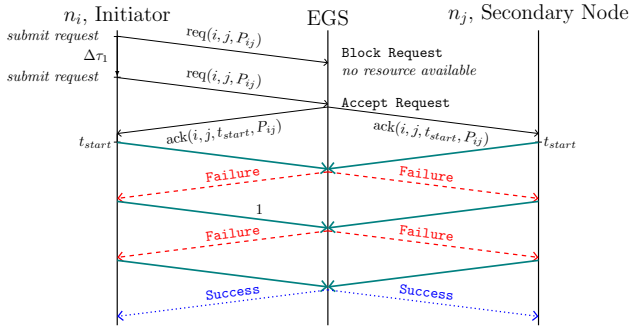
1) CI Protocol

The goal of a node pair (n_i, n_j) running a CI protocol is to generate a data record of a qubit string at each of nodes n_i and n_j such that the data records of n_i and n_j have entanglement-like correlations. These data records can serve as raw key, and the CI protocol is a form of QKD [1], [2] that uses a BSA as a measurement device at a central midpoint, known as MDI-QKD [41], [49], [50]. The advantage of MDI-QKD over traditional QKD protocols is that it is more secure, as it removes the possibility of detector side-channel attacks. Detector side-channel attacks have been shown to be easy to implement and challenging to defend against [54]–[56], hence this is an important advantage. To support running the CI generation protocol, nodes n_i and n_j require a photon source. This source can be a multi-photon emitter attenuated to the single-photon level, such as an attenuated laser pulse, or a true single photon source like a quantum communication qubit. The CI generation protocol's compatibility with multi-photon sources is advantageous. In the early stages of quantum network development, multi-photon sources are more cost-effective and easier to build/purchase and maintain compared to single photon sources [41].

In broad terms, the protocol consists of four main stages. First, nodes exchange calibration information to ensure that the photons emitted by their sources will be spectrally indistinguishable. Second, each node encodes a qubit state into an emission of the photon source. Third, the photons are sent to a BSA, at which a BSM is attempted between the two received photons, which if successful, projects the photons into a maximally entangled state. Fourth, if the BSM



(a) CI protocol



(b) HE protocol

Figure 18: Communication sequences for the (a) CI and (b) HE protocols. The request field P_{ij} represents a packet of parameters, such as the number of attempts requested. $\Delta\tau_1$ is an exponentially distributed inter-arrival time between subsequent requests. Black arrows indicate classical messages. Teal arrows indicate pulses of light at the single photon level sent over optical fiber. Blue, dotted (red, dashed) arrows indicate a success (failure) flag communicated over a direct connection such as optical fiber.

succeeds, the records of the qubits prepared by (n_i, n_j) share entanglement-like correlations and a success flag is communicated to the nodes. In case of success the nodes store the qubit record. If the measurement is not successful, a failure flag is passively communicated to the nodes as the absence of sending a success flag; after the RTT time between each node and the EGS passes, plus some buffer time accounting for measurement and jitter, nodes discard qubit records from failed attempts. It is not necessary to wait for the outcome flag before commencing subsequent attempts. However, if the nodes use communication qubits to generate the single photons, the communication qubits must be measured in between subsequent attempts in order to obtain the data record. If a simple photon source is used this measurement step is not necessary, as the data record corresponds to the preparation record of the photon. Sequential execution of the second and third stages constitute a single CI generation attempt and may be repeated in batches. For any attempt, the fourth stage must eventually occur – the outcome flag must eventually be received by the nodes. Otherwise, a stop of the protocol may be triggered.

A communication sequence between the node pair (n_i, n_j) allocated use of an EGS resource to perform CI generation is illustrated in Figure 18a.

2) Comparison of the HE and CI protocols

The main difference between these two protocols is that in the HE protocol a successful BSM at the EGS means that nodes become entangled whereas in the CI protocol a successful BSM at the EGS means that the data records corresponding to the photons sent to the EGS will display entanglement like correlations. Other important differences are the repetition rate and the probability that single attempts succeed. See Figures 18a and 18b for visualization of how the communication sequences differ between the protocols. The frequency at which CI generation can be attempted is inherently faster than HE generation because subsequent attempts do not require waiting for the outcome flag to be received (step four in the HE and CI communication sequences). The rate at which CI generation can be attempted with simple photon sources is limited by the emission rate of the photon source, the channel capacity of the fiber over which photons are sent, or the rate at which the BSA detectors become responsive again following detection of a photon (recovery from dead-time), whichever of these factors is most restrictive. If quantum communication qubits are used to enact the CI generation protocol, the other potential rate limiting factor is how quickly the qubits can be measured following photon emission. To summarize, subsequent attempts of the CI generation protocol can begin before the completion of previous attempts, and the rate limiting factors in an implementation may allow for significantly high attempt repetition rates (MHz or GHz). In contrast, the HE generation protocol requires nodes to receive the heralding flag of any previous attempt before commencing any subsequent attempt, i.e., attempts must be non-overlapping. In a local quantum network with optical fiber links between nodes and the EGS ranging from 5 to 50 km, the RTTs vary from approximately 25 μs to 250 μs. Such RTTs dominate the time duration of a single HE generation attempt, limiting the maximum repetition rate of attempts in the protocol to the level of kHz.

A real world implementation of CI generation with simple photon sources [41] displayed a high total probability of a single attempt succeeding. In this implementation the main limitations against an attempt succeeding were loss of a photon traveling in fiber or a detection pattern incompatible with projection into a maximally entangled state. In contrast, in real world implementations of single slick bipartite heralded entanglement generation the probability of an attempt succeeding is limited by the probability of single photon emission from the communication qubit, and further suppressed by the probability of losing a photon traveling in fiber and the success probability of a BSM.

H. SIMULATION IMPLEMENTATION

Simulations of the discrete, Cox and exponential models allow a method of comparing the predictions of Theorems 1 and 2 to the blocking probabilities we would expect to observe in a real implementation where the distribution governing the duration of events coincides with the simulated model. The discrete simulation mode is implemented as a discrete time simulation where the time step is set to the duration of a single attempt, which is the shortest timescale in the system. The Poisson process by which flows submit requests is implemented by scaling the exponentially distributed inter-arrival times between requests to a number of time steps and rounding up to the nearest time step. The exponential and Cox simulation modes are implemented as discrete event simulations.

The expected number of requests placed over the duration of each simulation depends on the request arrival rate from each flow, ν_1^f. Each simulation corresponded to 1150.73 seconds of simulated real-time. The data in Figure 7 is representative for a discussion of the expected number of requests placed over that time. For the minimum (maximum) request arrival rate simulated, the expected number of requests from each flow is 100 (1135). Since there are 28 flows in total, the total expected number of requests placed during a simulation is then 2800 (31780) for the minimum (maximum) request rate simulated. Note that we in general observe very close agreement between the blocking probabilities resulting from discrete/exponential/Cox simulation types as well as with our numeric evaluation of (7). The magnitude of disagreement between simulation types may be referred to as the error between simulation types. To investigate whether these errors change systematically when simulation parameters are altered we have executed simulations with various time-step durations pertaining to the discrete simulation as well as with various numbers of attempts per batch of entanglement generation attempts. Under changes to these parameters, the errors between simulation types remain very small ($< 1\%$) and do not change in any systematic way.

In the implementation of the multiple EPR pair generation with resource relinquishment (jump-over) service mode, the mean duration of the idle periods, T_{idle}, is set to be equivalent to the mean duration of the calibration periods of the single/multiple EPR pair generation with strict resource reservation service modes, i.e., $T_{\text{idle}} = T_{\text{calib}}$. The motivation for this implementation choice is to enable direct comparison between the different service modes, as in Figure 9.

In the non-homogeneous traffic scenario considered, for every flow, if one or both of the users involved is connected by a longer link, then the attempt time for the flow is lengthened to accommodate the greater RTT between the more distant node and the EGS. Thus flows in S_2 require longer resource reservation times to achieve the same number of attempts. As mentioned above, in the discrete simulations of homogeneous traffic, the time step is set to the duration of a single attempt. In the implementation of our non-homogeneous traffic model, we set the time step in discrete simulations to the duration of an attempt in S_1, which is the shortest time scale in the system. We then fix the mean attempt time in S_2 to double the mean attempt time of flows in S_1. This time is longer than the total of the RTT required for flows in S_2 plus the buffer included to capture measurement times for flows in S_1. Hence we have modeled adding some additional buffer time to each attempt for flows in S_2. As a consequence, in discrete simulations the duration of each attempt for flows in S_2 is exactly two-time steps. We assume all flows submit requests at the same rate, thus isolating the source of non-homogeneity to the different attempt times in sets S_1 and S_2, which are largely due to the different link-lengths.

Parameter type	Value
Number of phases	4
Mean duration of phase 1	$0.41\overline{6} \cdot T_{\text{attempt}}$
Mean duration of phase 2	$0.521 \cdot T_{\text{attempt}}$
Mean duration of phase 3	$0.651 \cdot T_{\text{attempt}}$
Mean duration of phase 4	$0.814 \cdot T_{\text{attempt}}$
Transition probabilities	(0.6, 0.48,0.384, 0.307)

Table 5: Cox distribution parameters for individual entanglement generation attempts in the homogeneous traffic scenario.

I. EXTENDED DATA: HOMOGENOUS TRAFFIC IN THE SINGLE ENTANGLEMENT GENERATION WITH STRICT RESOURCE RESERVATION SERVICE MODE

ACKNOWLEDGMENT

SG thanks Arian J. Stolk and Kian L. van der Enden for helpful conversations about the system description. SG thanks her PhD supervisor Stephanie Wehner for her support in conducting this research.

References

[1] B. H. Charles and G. Brassard. Quantum cryptography: Public key distribution and coin tossing. Theoretical Computer Science, 560:7–11, December 2014.

[2] A. K. Ekert. Quantum cryptography based on bell's theorem. Phys. Rev. Lett., 67:661–663, August 1991.

[3] P. Arright and L. Salvail. Blind quantum computation. International Journal of Quantum Information, 04(05):883–898, 2006.

[4] A. Broadbent, J. Fitzsimons, and E. Kashefi. Universal blind quantum computation. In 2009 50th Annual IEEE Symposium on Foundations of Computer Science, pages 517–526, Atlanta, Georgia, USA, October 2009. IEEE.

[5] D. Leichtle, L. Music, E. Kashefi, and H. Ollivier. Verifying bqp computations on noisy devices with minimal overhead. PRX Quantum, 2:040302, Oct 2021.

[6] D. Gottesman, T. Jennewein, and S. Croke. Longer-baseline telescopes using quantum repeaters. Phys. Rev. Lett., 109:070503, Aug 2012.

[7] P. Kómár, E. M. Kessler, M. Bishof, L. Jiang, A. S. Sørensen, J. Ye, and M. D. Lukin. A quantum network of clocks. Nature Physics, 10(8):582–587, June 2014.

[8] V. Giovannetti, S. Lloyd, and L. Maccone. Quantum-enhanced positioning and clock synchronization. Nature, 412(6845):417–419, 2001.

[9] X. Guo, C. R Breum, J. Borregaard, S. Izumi, M. V Larsen, T. Gehring, M. Christandl, J. S Neergaard-Nielsen, and U. L Andersen. Distributed

quantum sensing in a continuous-variable entangled network. Nature Physics, 16(3):281–284, 2020.

[10] W. J. Munro, K. Azuma, K. Tamaki, and K. Nemoto. Inside quantum repeaters. IEEE Journal of Selected Topics in Quantum Electronics, 21(3):78–90, 2015.

[11] 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. Physical review letters, 70(13):1895, 1993.

[12] D. Bouwmeester, J-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger. Experimental quantum teleportation. Nature, 390(6660):575–579, 1997.

[13] S. Gauthier, G. Vardoyan, and S. Wehner. An architecture for control of entanglement generation switches in quantum networks. IEEE Transactions on Quantum Engineering, 4:1–17, 2023.

[14] S. L. Braunstein and A. Mann. Measurement of the Bell operator and quantum teleportation. Physical Review A, 51(3):R1727, 1995.

[15] M. Michler, K. Mattle, H. Weinfurter, and A. Zeilinger. Interferometric Bell-state analysis. Physical Review A, 53(3):R1209, 1996.

[16] P. Walther and A. Zeilinger. Experimental realization of a photonic Bell-state analyzer. Physical Review A, 72(1):010302, 2005.

[17] M. Koyama, C. Yun, A. Taherkhani, N. Benchasattabuse, B. O. Sane, M. Hajdušek, S. Nagayama, and R. Van Meter. Optimal switching networks for paired-egress bell state analyzer pools. In 2024 IEEE International Conference on Quantum Computing and Engineering (QCE), volume 01, pages 1897–1907, 2024.

[18] R. A. Thompson. Traffic capabilities of two rearrangeably nonblocking photonic switching modules. AT&T Technical Journal, 64(10):2331–2373, 1985.

[19] J. Preskill. Quantum computing in the nisq era and beyond. Quantum, 2:79, August 2018.

[20] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. Post Office Electrical Engineer's Journal, 10:189–197, 1917.

[21] M. Harchol-Balter. Performance modeling and design of computer systems: queueing theory in action. Cambridge University Press, 2013.

[22] A. Abane, M. Cubeddu, V. S. Mai, and A. Battou. Entanglement routing in quantum networks: A comprehensive survey. IEEE Transactions on Quantum Engineering, 6:1–39, 2025.

[23] L. Bacciottini, M. G. De Andrade, S. Pouryousef, E. A. Van Milligen, A. Chandra, N. K. Panigrahy, N. S. V. Rao, G. Vardoyan, and D. Towsley. Leveraging internet principles to build a quantum network. IEEE Network, pages 1–1, 2025.

[24] S. Pouryousef, H. Shapourian, and D. Towsley. Analysis of asynchronous protocols for entanglement distribution in quantum networks, 2024.

[25] B. A. Sevast'yanov. An ergodic theorem for Markov processes and its application to telephone systems with refusals. Theory of Probability & Its Applications, 2(1):104–112, 1957.

[26] T. Bonald. Insensitive queueing models for communication networks. In Proceedings of the 1st International Conference on Performance Evaluation Methodolgies and Tools, valuetools '06, page 57–es, New York, NY, USA, 2006. Association for Computing Machinery.

[27] R. Serfozo. Introduction to Stochastic Networks. Springer, New York, NY, USA, 01 1999.

[28] W. Whitt. Continuity of generalized semi-markov processes. Mathematics of Operations Research - MOR, 5:494–501, 11 1980.

[29] T. Bonald. The Erlang model with non-Poisson call arrivals. ACM SIGMETRICS Performance Evaluation Review, 34(1):276–286, 2006.

[30] F. Kelly. Charging and rate control for elastic traffic. European transactions on Telecommunications, 8(1):33–37, 1997.

[31] G. Vardoyan, S. Guha, P. Nain, and D. Towsley. On the Stochastic Analysis of a Quantum Entanglement Distribution Switch. IEEE Transactions on Quantum Engineering, 2:1–16, 2021.

[32] G. Vardoyan, S. Guha, P. Nain, and D. Towsley. On the exact analysis of an idealized quantum switch. Performance Evaluation, 144:102141, 2020.

[33] N. K. Panigrahy, T. Vasantam, D. Towsley, and L. Tassiulas. On the capacity region of a quantum switch with entanglement purification. In IEEE INFOCOM 2023-IEEE Conference on Computer Communications, pages 1–10, New York, NY, USA, 2023. IEEE, IEEE.

[34] W. Dai, A. Rinaldi, and D. Towsley. The capacity region of entanglement switching: Stability and zero latency. In 2022 IEEE International Conference on Quantum Computing and Engineering (QCE), pages 389–399, Broomfield, CO, USA, 2022. IEEE.

[35] T. Vasantam and D. Towsley. A throughput optimal scheduling policy for a quantum switch. In Philip R. Hemmer and Alan L. Migdall, editors, Quantum Computing, Communication, and Simulation II, volume 12015, page 1201505, San Francisco, CA, USA, 2022. International Society for Optics and Photonics, SPIE.

[36] L.-M. Duan, M. D. Lukin, J. I. Cirac, and P. Zoller. Long-distance quantum communication with atomic ensembles and linear optics. Nature, 414(6862):413–418, November 2001.

[37] C. Cabrillo, J. I. Cirac, P. García-Fernández, and P. Zoller. Creation of entangled states of distant atoms by interference. Phys. Rev. A, 59:1025–1033, February 1999.

[38] C. Simon, H. de Riedmatten, M Afzelius, N. Sangouard, H. Zbinden, and N. Gisin. Quantum Repeaters with Photon Pair Sources and Multimode Memories. Phys. Rev. Lett., 98:190503, May 2007.

[39] N. Sangouard, C. Simon, H. de Riedmatten, and N. Gisin. Quantum repeaters based on atomic ensembles and linear optics. Rev. Mod. Phys., 83:33–80, Mar 2011.

[40] N. Sangouard, R. Dubessy, and C. Simon. Quantum repeaters based on single trapped ions. Phys. Rev. A, 79:042340, Apr 2009.

[41] R. C. Berrevoets, T. Middelburg, R. F. L. Vermeulen, L. D. Chiesa, F. Broggi, S. Piciaccia, R. Pluis, P. Umesh, J. F. Marques, W. Tittel, and J. A. Slater. Deployed measurement-device independent quantum key distribution and bell-state measurements coexisting with standard internet data and networking equipment. Communications Physics, 5(1):186, 2022.

[42] P. C. Humphreys, N. Kalb, J. P. J. Morits, R. N. Schouten, R. F. L. Vermeulen, D. J. Twitchen, M. Markham, and R. Hanson. Deterministic delivery of remote entanglement on a quantum network. Nature, 558(7709):268–273, June 2018.

[43] H. Bernien, B. Hensen, W. Pfaff, G. Koolstra, M. S. Blok, L. Robledo, T. H. Taminiau, M. Markham, D. J. Twitchen, L. Childress, and R. Hanson. Heralded entanglement between solid-state qubits separated by three metres. Nature, 497(7447):86–90, April 2013.

[44] P. Maunz, D. L. Moehring, S. Olmschenk, K. C. Younge, D. N. Matsukevich, and C. Monroe. Quantum interference of photon pairs from two remote trapped atomic ions. Nature Physics, 3(8):538–541, 2007.

[45] V. Krutyanskiy, M. Galli, V. Krcmarsky, S. Baier, D. A. Fioretto, Y. Pu, A. Mazloom, P. Sekatski, M. Canteri, M. Teller, J. Schupp, J. Bate, M. Meraner, N. Sangouard, B. P. Lanyon, and T. E. Northup. Entanglement of trapped-ion qubits separated by 230 meters. Phys. Rev. Lett., 130:050803, February 2023.

[46] C. W. Chou, H. de Riedmatten, D. Felinto, S. V. Polyakov, S. J. van Enk, and H. J. Kimble. Measurement-induced entanglement for excitation stored in remote atomic ensembles. Nature, 438(7069):828–832, December 2005.

[47] C. W. Chou, J. Laurat, H. Deng, K. S. Choi, H. de Riedmatten, D. Felinto, and H. J. Kimble. Functional quantum nodes for entanglement distribution over scalable quantum networks. Science, 316(5829):1316–1320, June 2007.

[48] T. van Leent, M. Bock, F. Fertig, R. Garthoff, S. Eppelt, Y. Zhou, P. Malik, M. Seubert, T. Bauer, W. Rosenfeld, W. Zhang, C. Becher, and H. Weinfurter. Entangling single atoms over 33 km telecom fibre. Nature, 607(7917):69–73, July 2022.

[49] H-K. Lo, M. Curty, and B. Qi. Measurement-device-independent quantum key distribution. Phys. Rev. Lett., 108:130503, Mar 2012.

[50] S. L. Braunstein and S. Pirandola. Side-channel-free quantum key distribution. Phys. Rev. Lett., 108:130502, Mar 2012.

[51] M. Pompili, S. L. N. Hermans, S. Baier, H. K. C. Beukers, P. C. Humphreys, R. N. Schouten, R. F. L. Vermeulen, M. J. Tiggelman, L. dos Santos Martins, B. Dirkse, S. Wehner, and R. Hanson. Realization of a multinode quantum network of remote solid-state qubits. Science, 372(6539):259–264, April 2021.

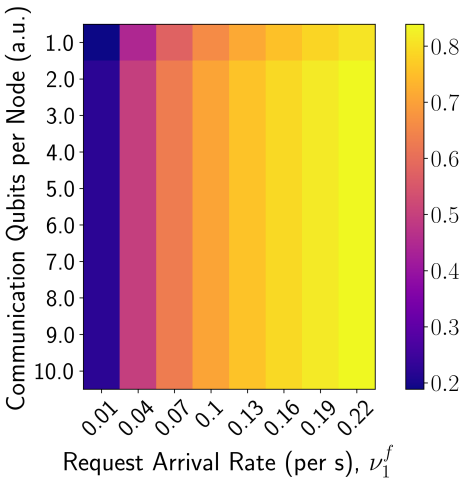
[52] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters. Purification of noisy entanglement and faithful teleportation via noisy channels. Physical review letters, 76(5):722, 1996.

[53] D. Deutsch, A. Ekert, R. Jozsa, C. Macchiavello, S. Popescu, and A. Sanpera. Quantum privacy amplification and the security of quantum cryptography over noisy channels. Physical review letters, 77(13):2818, 1996.

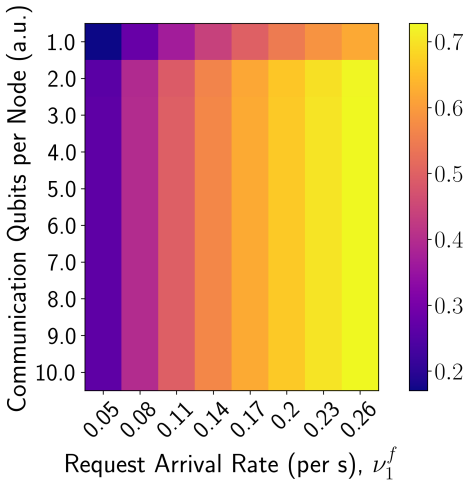
[54] H. Lu, C-H. F. Fung, and Q-Y. Cai. Two-way deterministic quantum key distribution against detector-side-channel attacks. Phys. Rev. A, 88:044302, Oct 2013.

[55] V. Makarov, A. Anisimov, and J. Skaar. Effects of detector efficiency mismatch on security of quantum cryptosystems. Phys. Rev. A, 74:022313, Aug 2006.

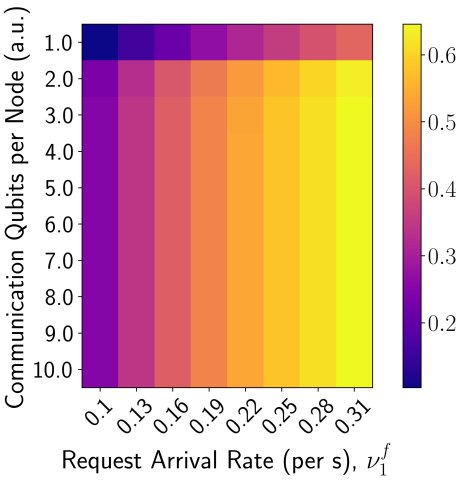
[56] Y. Zhao, C-H. F. Fung, B. Qi, C. Chen, and H-K. Lo. Quantum hacking: Experimental demonstration of time-shift attack against practical quantum-key-distribution systems. Phys. Rev. A, 78:042333, Oct 2008.



(a) one EGS Resource ($C = 1$)



(b) two EGS Resources ($C = 2$)



(c) three EGS Resources ($C = 3$)

Figure 19: Heatmaps of the average blocking probability per flow when the number of communication qubits per node and the request arrival rates are varied. Data results from numeric evaluation of (7) for an EGS with 20 nodes and serving $\binom{20}{2} = 190$ flows, one for each possible node pairing. Session traffic is homogeneous.