An on-demand Resource Allocation Algorithm for a Quantum Network Hub and its Performance Analysis

Gauthier, Scarlett; Vasantam, Thirupathaiah; Vardoyan, Gayane

DOI
10.1109/QCE60285.2024.00204

Publication date
2024

Document Version
Final published version

Published in
2024 IEEE International Conference on Quantum Computing and Engineering (QCE)

An on-demand resource allocation algorithm for a quantum network hub and its performance analysis

Scarlett Gauthier
EEMCS and QuTech
Delft University of Technology
Delft, The Netherlands
s.s.gauthier@tudelft.nl

Thirupathaiah Vasantam
Department of Computer Science
Durham University
Durham, UK
thirupathaiah.vasantam@durham.ac.uk

Gayane Vardoyan
EEMCS and QuTech; CICS
Delft University of Technology, The Netherlands;
University of Massachusetts, Amherst, USA
gvardoyan@cs.umass.edu

***Abstract*—To support the execution of multiple simultaneously-running quantum network applications, 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 modelled 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. 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 their underlying distributions. Our numerical results support our analysis. 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, queueing 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], [4], and several entanglement-based quantum sensing techniques [5], [6], [7], [8]. 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. When multiple applications have simultaneous demand for shared and limited resources, contention can arise, and the network must enact a resource allocation scheme.

GV acknowledges support from NWO QSC grant BGR2 17.269. SG acknowledges the support of the European Union's Horizon Europe research and innovation program under grant agreement No. 101102140.

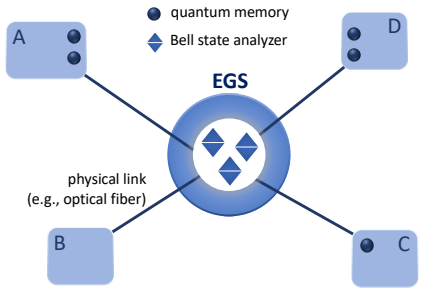

Fig. 1: An EGS with three BSAs servicing four nodes.

We study here a type of quantum network hub previously referred to as an Entanglement Generation Switch (EGS) [9]. 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) [10], [11], 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 successful swap generating end-to-end entanglement. We remark that general quantum network applications (e.g., [3], [4], [5], [6], [7], [8]), require at least one node involved in the application execution to possess a qubit that may act as a quantum memory. These end nodes are able to support such swapping.

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 1 for an example. 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.

With these properties, the EGS is poised to be an excellent candidate for a scalable and straightforwardly implementable

979-8-3315-4137-8/24/$31.00 ©2024 IEEE
DOI 10.1109/QCE60285.2024.00204

metropolitan-area quantum network component, especially in the Noisy Intermediate Scale Quantum (NISQ) era [12]. While the EGS can be used to directly connect end nodes, as shown in Figure 1, 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. Its versatility warrants investigation into its practical operation; we provide a detailed explanation of this in Section III. We then model the hub as an Erlang loss system, with the EGS acting as a server and the nodes attached to it generating entanglement requests. We assume that these entanglement requests arrive 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: the first two specify session termination behavior, and the third mode enforces EGS resource relinquishment when nodes are not actively utilizing them. Throughout, we heed 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. 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 analyze our model of the EGS to obtain (1) the stationary distribution of the number of active requests being served at the switch; (2) request blocking probabilities; and (3) an insensitivity result that highlights the broad applicability of our model to practical systems;
- 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 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 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 telephone exchange that directs and facilitates communication between sets of callers. Traditionally, such 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 lines are occupied upon arrival. The blocking probability is computed using the well-known Erlang formula [13]. 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 [14]. 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 [15]. 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 [16], 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, requests 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 our work is based on that of [16], although it is significantly different due to the presence of new parameters and characteristics of quantum systems.

The EGS architecture was initially introduced in [9], where the authors highlighted its scalable properties. The authors then proposed and studied a Rate Control Protocol whose aim is to modulate user demand rates 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 framework [17]. 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., [18], [19], [20], [21]. In contrast to EDSs, 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 EDSs, 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 system, several repeater architectures (see e.g., [22], [23], [24], [25], [26]) 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 *resources* such as BSAs; (2) a *switch* capable of allocating a resource to any pair of nodes; (3) and a *processor* capable of making scheduling decisions, controlling the operation of the switch, and sending and receiving classical messages. Nodes are connected to an EGS via physical links, such as optical fiber. To gain access to an EGS resource, pairs of nodes send a *request* to the EGS – a demand for the generation of one or more Einstein-Podolsky–Rosen (EPR) pairs.

A node possesses (i) one or more *communication qubits*, each capable of preparing a quantum state and emitting one or more photons; (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 to control quantum states prepared in communication qubits, trigger swaps to memory, trigger measurement of a communication qubit, and send and receive classical messages.

Bipartite Heralded Entanglement (HE) generation [23], [22] and generation of Correlated Information (CI) [1], [2] are two ways in which a pair of nodes can interact via the EGS. Production of entanglement by the method of HE generation has been successfully demonstrated in several experimental platforms, including Color Centers [27], [28], Ion Traps [29], [30] Atomic Ensembles [31], [32] and Neutral Atoms [33].

Applications of HE generation include BQC, teleportation and clock-synchronization [6], and an application of CI generation is Measurement Device Independent QKD [34], [35]. Each of these tasks can be enabled by an EGS where the shareable resource is a BSA. To motivate our service models for the EGS we describe in detail the process of HE generation following a single-click scheme.

The goal of a node pair (n_i, n_j) running the bipartite HE generation protocol is to entangle a communication qubit of node n_i with that of node n_j. For every entanglement generation attempt the nodes make, a success or failure flag is generated and converted into a message that is sent to the nodes. At a high level, a single-click HE 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 encoded photons. Fourth, if the BSM succeeds the communication qubits of the two nodes will have become entangled and a success flag is sent to the nodes. The second, third and fourth stages occurring sequentially constitute a single HE generation attempt. As an example, for the NV center in diamond platform, the calibration operations correspond to a Charge and Resonance (CR) check [36].

Attempts can be repeated in batches that are interleaved with repetition of the first step – calibration of the communication qubit. 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.

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. 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} [28]. 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 [29], [30], 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 earlier (see Theorem 3 in Section V).

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*, i.e., 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 [36]. This effect results from a persistent non-zero coupling to the memory. 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. We first present a number of abstractly-defined terms which bridge the gap between the physical and queueing-theoretic models of the EGS. 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. Service models will be explained in more detail later in this section; we first provide a description of the underlying components of service.

A *call* is the basic service component for two nodes communicating via the EGS. A call involves the active use of EGS components for a period of time, such as the utilization of a resource to attempt entanglement generation between two nodes. For the purpose of this work, we establish two additional service component types, where EGS resources are not in active use for their duration. The first is an *idle period* during which the nodes relinquish all EGS resources, so that a subsequent call in the session would require a new service reservation. In this work, we assume that a node's communication qubit is unavailable during an idle period, so that it cannot be used to initiate a new service request for the node until the current one completes service. We leave relaxations of this assumption for future work.

The second is a *calibration period* which requires node hardware resources, but no active utilization of EGS resources. The duration of such calibration periods typically has a finite mean, albeit it can be randomly distributed. Depending on the service model, nodes may continue to hold onto EGS resources for the duration of a calibration period, precluding other nodes from using them. We assume in this work that a calibration period engages all qubits of the corresponding communication session, 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 sessions may no longer be treated independently.

Note 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.

A request from nodes n_i and n_j to access an EGS resource for entanglement generation, triggers the creation of a *session*, denoted by the tuple (n_i, n_j). This is a sequence of calls, possibly interleaved with idle and/or calibration periods generated by the two nodes. By convention, sessions begin and end with calls, and not idle or calibration periods. Throughout this work, we often use the term "session" and the (queueing theory inspired) term "flow" interchangeably. As in [16], we permit the existence of differently-structured sessions within one system. Physically, these may correspond to different applications, entanglement distribution algorithms, or even application *instantiations* between the same node pair. The set of all possible flows is denoted as $\mathcal{F}$. Let K be the number of nodes connected to the EGS; then the cardinality of $\mathcal{F}$ is given by $F = \binom{K}{2}$. We note that while flows uniquely identify a pair of nodes, sessions need not be unique: n_i and n_j can have multiple concurrent sessions, as long as resources (communication qubits and EGS resources) are available. Similar to the work of Bonald in [16], we assume that sessions are independent, their arrivals form a Poisson process, and that session components (calls, idle and calibration periods) have Coxian distributed durations. Note that Coxian distributions approximate general distributions to arbitrary accuracy [37], [38]. Every active session – one that has been accepted for service by the EGS – uses a single qubit from each participating node.

We study three distinct service models of the EGS:

• *Single EPR Pair Generation with Strict Resource Reservation*: a session consists of entanglement generation attempts interleaved with calibration periods. Once a session is admitted at the EGS, attempts are carried out until one is successful, or until the last attempt is complete. Both events result in session termination. A flow holds onto its EGS resource during a calibration period, even though it is not actively utilized.

• *Multiple EPR Pair Generation with Strict Resource Reservation*: all properties of the previous service model apply, with the exception that a session terminates only when all attempts

are carried out. A session can thus produce multiple EPR pairs.

• *Multiple EPR Pair Generation with Resource Relinquishment*: a session consists of entanglement generation attempts, interleaved with "idle" periods at the beginning of which the EGS resource is given up, and at the end of which the flow attempts to re-obtain a resource. 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. As in the previous service model, successful entanglement generation by itself does not cause session termination.

We model both Single and Multiple EPR Pair Generation with Strict Resource Reservation service models to capture the physical requirement that a limited quantum node may only be equipped to engage in single entanglement generation, whereas a quantum node with advanced capabilities may possibly be able to engage in multiple entanglement generation (see Section III for details). Inclusion of the Multiple EPR Pair Generation with Resource Relinquishment in addition to the two strict reservation service models is valuable as *a priori* it is not clear whether releasing or retaining use of EGS resources between batches of entanglement generation attempts will result in more favorable performance of the EGS.

Irrespective of the service model, the state space of the system can be represented using a vector $\mathbf{x} = [\mathbf{x}^{f_1}, \ldots, \mathbf{x}^{f_F}]$, where F is the number of possible flows, and each $\mathbf{x}^f$ is a vector describing the number of jobs in the queues corresponding to flow $f \in \mathcal{F}$. As per the description above, each component of a flow is modeled using a Coxian distribution. We define the following variables for a flow f:

- $A^f_{i,j}/C^f_{i,j}/I^f_{i,j}$: ith phase of call/calibration/idle period j;
- $N^f_{A/C/I}$: number of phases per call/calibration/idle period;
- $M^f_{A/C/I}$: number of call/calibration/idle periods.
- $L^f \equiv N^f_A \times M^f_A + N^f_C \times M^f_C + N^f_I \times M^f_I$ as the total number of phases in session type f;
- $L \equiv \sum_{f\in\mathcal{F}} L^f$ – the dimension of vector $\mathbf{x}$;
- $x^{f,A/C/I}_{i,j}$ is used to refer to the number of jobs present in the ith phase of the jth call/calibration/idle period;
- $\mathbf{e}^{f,A/C/I}_{i,j}$ are vectors of dimension L, with all entries zero except the one corresponding to the ith phase of call/calibration/idle period j of flow f, which is one;
- $\mathbf{e}^f_i$, $i \in \{1, \ldots, L\}$, is a vector of dimension L with all entries zero except the one corresponding to the ith component of $\mathbf{x}^f$, which is equal to one;
- x_i, $i \in \{1, \ldots, L\}$, refers to the ith element of the vector $\mathbf{x}$ when there is no need to identify a flow or period within it;
- x^f_i, $i \in \{1, \ldots, L^f\}$, refers to the ith element of $\mathbf{x}^f$ when there is no need to identify a specific period or phase.
- For a state $\mathbf{x}$, the number of active sessions of flow f is $q^f(\mathbf{x}^f) \equiv \sum_{i=1}^{L^f} x^f_i$.

Finally, we assume that quantum nodes have a limited number of communication qubits – c_k for node n_k, and that the EGS has a total of C resources to be allocated to flows for distributing entangled states.

V. Analysis

In this section, we analyze the three EGS service models. For each scenario, we derive the stationary distribution of observing the system in a given state. The main quantity of interest is the probability that a request for accessing EGS resources is blocked – an event that occurs when said request sees all C EGS resources engaged. In strict resource reservation modes, blocking can only occur at the beginning of a session. On the other hand, if sessions contain idle periods as in the resource relinquishing mode, then blocking may also occur throughout the session, after departures from idle periods. We prove insensitivity by showing that in all cases, the blocking probability depends only on flows' traffic intensities.

A. Single EPR Pair Generation, Strict Resource Reservation

Recall that in this service model, admitted sessions are processed in their entirety, terminating only if an EPR pair attempt is successful. This means that (i) by definition, there is no notion of an idle period in this service model, and (ii) a session that is being actively serviced by the EGS does not relinquish its BSA even during a calibration period. Figure 2 provides the general form of a session within this service model. In Figure 3 the attempt and calibration periods are depicted in their decomposed form; e.g., exponential phases $A_{1,1}, \ldots, A_{N_A,1}$ comprise the Coxian-distributed period $\mathcal{A}_1$.

Given these specifications, we can define the state space of the associated continuous-time Markov chain. Namely, the EGS must heed the resource (BSA) capacity, and each network node must heed its own communication qubit limit. The admissible state space is thus given by

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

where the notation $n_k \in f$ means that node n_k partakes in flow f. The set $\mathcal{S}$ is coordinate convex, i.e., if $\mathbf{x} \in \mathcal{S}$, then $\mathbf{y} \in \mathcal{S}$ for all $\mathbf{y}$ such that $\mathbf{0} \le \mathbf{y} \le \mathbf{x}$ component-wise. For the queueing network in Figure 3 where each phase corresponds to a queue, under the strict resource reservation model we specify the following properties:

- All external arrival rates (i.e., those originating from outside of the network) $\nu^f_i(\mathbf{x})$ are zero, except for those of queues $A^f_{1,1}$, $f \in \mathcal{F}$. We denote these rates with $\nu^f_1(\mathbf{x})$, $f \in \mathcal{F}$, $\mathbf{x} \in \mathcal{S}$, so that the transition $\mathbf{x} \to \mathbf{x} + \mathbf{e}^{f,A}_{1,1}$ occurs with rate

$$\nu^f_1(\mathbf{x}) = \begin{cases} \nu^f_1, & \text{if } \mathbf{x} + \mathbf{e}^{f,A}_{1,1} \in \mathcal{S}, \\ 0, & \text{else.} \end{cases} \tag{2}$$

- Transition $\mathbf{x} \to \mathbf{x} - \mathbf{e}^f_i + \mathbf{e}^f_{i+1}$ occurs with probability $p^f_{i,i+1}$, for $1 \le i < L^f$.
- A special case of the above is that $p^f_{i,i+1} = 1$, $\forall f \in \mathcal{F}$, if x^f_i corresponds to the last phase of a calibration period.
- Transition $\mathbf{x} \to \mathbf{x} - \mathbf{e}^f_i + \mathbf{e}^f_j$ occurs with probability $p^f_{i,j}$ if j is such that x^f_j corresponds to the initial phase of the call/calibration period that follows the call/calibration period corresponding to x^f_i.

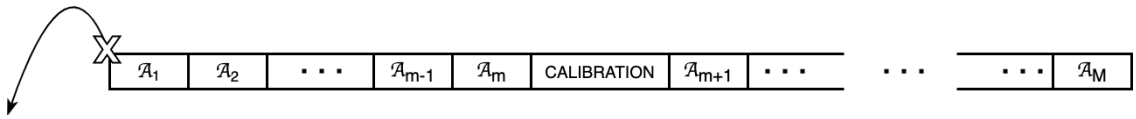

Fig. 2: Strict resource reservation service model. A session consists of multiple EPR pair generation attempts, i.e., calls, denoted by $\mathcal{A}_i$, $i = 1, \ldots, M$. A calibration period is carried out after every m attempts. In the "multiple EPR pair generation" variant of this service model, 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.

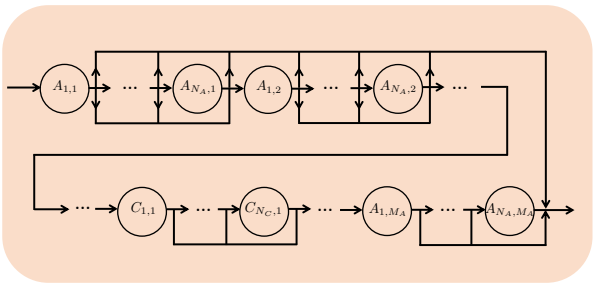

Fig. 3: A session from the single EPR pair generation strict resource reservation service model, shown with the periods of Fig. 2 decomposed into exponentially-distributed phases, so as to result in Coxian-distributed call and calibration periods.

- 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 and the transition represents the event that entanglement generation succeeds after the ith phase of flow f – in this service model, this event causes the session to end. We note that leaving the session from a calibration phase is not possible.

Finally, we define μ_i^f as flow f's job processing rate at queue i of flow f, and $\lambda_i^f(\mathbf{x})$ as the total arrival rate into queue i of flow f while in state $\mathbf{x}$. When $i = 1$, i.e., it corresponds to the first phase of a session, the total arrival rate is $\lambda_1^f(\mathbf{x}) = \nu_1^f(\mathbf{x})$. For all other queues, the arrival rate is given by

$$\lambda_i^f(\mathbf{x}) = \begin{cases} \nu_1^f \tilde{p}_i^f \equiv \lambda_i^f, & \text{if } \mathbf{x} + \mathbf{e}_i^f \in \mathcal{S}, \\ 0, & \text{else.} \end{cases} \tag{3}$$

Above, $\tilde{p}_i^f$, $2 \le i \le L^f$, denotes the probability of reaching the ith phase starting from the first phase of a session belonging to flow f. The following lemma provides an expression for these probabilities; we omit the proof due to space constraints as it follows easily.

Lemma 1. *For a flow f, let p_l^k be the probability of leaving the queueing network after phase l of period k due to a successful BSM, and $p_{i,j}^k$ denote the probability of transitioning from the ith to the jth phase of period k. Then the probability P_k of leaving during a call period k, conditioned on the event that the session has entered period k, is $P_k = \sum_{m=1}^{N^k} \left\{ \prod_{l=1}^{m-1} p_{l,l+1}^k \right\} p_m^k$, where N^k is the number of phases in the period. Further, the probability that the session reaches the ith phase of period j starting from the first phase is $\tilde{p}_i^f = \prod_{k=1}^{j-1} (1 - P_k) \prod_{l=1}^{i-1} p_{l,l+1}^j$. Here we use the convention that $\prod_{m=a}^{b} x_m = 1$ if $b < a$ and $P_k = 0$ for any k that corresponds to a calibration period.*

We next perform an analysis of the system, beginning with the derivation of the stationary distribution.

Theorem 1. *The stationary distribution $\pi(\mathbf{x})$ of the system with single EPR pair generation while in strict resource reservation mode is given by*

$$\pi(\mathbf{x}) = \left(\sum_{\mathbf{y} \in \mathcal{S}} \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{y_i^f}}{y_i^f!} \right)^{-1} \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{x_i^f}}{x_i^f!}, \tag{4}$$

where $\rho_i^f = \frac{\lambda_i^f}{\mu_i^f}$ is the traffic intensity of the ith queue of a session corresponding to flow f.

Proof. It can be verified that this stationary distribution satisfies local balance equations, indicating that the rate of leaving a state $\mathbf{x}$ due to departure from queue i of flow f coincides with the rate of entering $\mathbf{x}$ due to an arrival at queue i of flow f, for any $\mathbf{x}$, i, and f. The proof is standard, see, e.g., [39, Chapter 17] for more details on the theory of Jackson networks. □

Having determined the stationary distribution, we are now ready to derive the blocking probability for an attempt of a given flow. First, we define $\mathcal{Q}(h)$ as the set of $\mathbf{q} = [q_1, \ldots, q_F] \in \mathbb{Z}_+^F$ that satisfy the relations $\sum_{i=1}^F q_i = h$ and $\sum_{i:n_k \in f_i} q_i \le c_k, \forall k \in \{1, \ldots, K\}$. Similarly, we define $\mathcal{Q}'(i)$ as the set of $\mathbf{q} = [q_1, \ldots, q_F] \in \mathbb{Z}_+^F$ that satisfy the relations $\sum_{j:n_k \in f_j} q_j \le c_k, \forall n_k \notin f_i$ and $\sum_{j:n_l \in f_j} q_j < c_l, \forall\, n_l \in f_i$. 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 qubit communication constraints are not violated, with the additional constraint that nodes belonging to flow f_i have at least one unoccupied communication qubit each.

Theorem 2. *For the system with single EPR pair generation operating in strict reservation mode, the blocking probability of an arriving request (session) belonging to a flow f_i, $i \in \{1, \ldots, F\}$ is*

$$\overline{\pi}_i(C) = \left(\sum_{\mathbf{q} \in \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!} \right)^{-1} \sum_{\mathbf{q} \in \mathcal{Q}(C) \bigcap \mathcal{Q}'(i)} \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!}, \tag{5}$$

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

Proof. 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

$$\overline{\pi}_i(h) = \frac{\mathrm{P}(\Omega_1(h) \cap \Omega_2(i))}{\mathrm{P}(\Omega_2(i))} = \left(\sum_{\mathbf{q} \in \mathcal{Q}'(i)} \mathrm{P}(\mathbf{q}) \right)^{-1} \sum_{\mathbf{q} \in \mathcal{Q}(h) \cap \mathcal{Q}'(i)} \mathrm{P}(\mathbf{q}). \tag{6}$$

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

$$\mathrm{P}(\mathbf{q}) = \sum_{\mathbf{x}:\mathbf{q}(\mathbf{x})=\mathbf{q}} \pi(\mathbf{x}) = \sum_{\substack{\mathbf{x}^{f_j}: q^{f_j}(\mathbf{x}^{f_j})=q_j, \\ j=1,\ldots,F}} D \prod_{f \in \mathcal{F}} \prod_{i=1}^{L^f} \frac{(\rho_i^f)^{x_i^f}}{x_i^f!}, \tag{7}$$

where $\mathbf{q}(\mathbf{x})$ is a vector containing $q^f(\mathbf{x}^f)$ values for $f \in \mathcal{F}$ and D is the denominator of $\pi(\mathbf{x})$ in (4). Recursive application of the multinomial theorem on (7) results in

$$P(\mathbf{q}) = D \prod_{j=1}^{F} \frac{\left(\rho^{f_j}\right)^{q_j}}{q_j!}, \tag{8}$$

where $\rho^{f_j} \coloneqq \sum_{i=1}^{L^{f_j}} \rho_i^{f_j}$ and $D = \left(\sum_{h=0}^{C} \sum_{\mathbf{q} \in \mathcal{Q}(h)} \prod_{j=1}^{F} \frac{(\rho^{f_j})^{q_j}}{q_j!} \right)^{-1}$. Using (8) and (6), 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 as in (5). □

Remark 1. *Let us take a closer look at the overall traffic intensity of a flow,* ρ^f*, for a given* $f \in \mathcal{F}$*:*

$$\rho^f = \sum_{i=1}^{L^f} \rho_i^f = \sum_{i=1}^{L^f} \frac{\lambda_i^f}{\mu_i^f} = \sum_{i=1}^{L^f} \frac{\nu_1^f \tilde{p}_i^f}{\mu_i^f} = \nu_1^f \sum_{i=1}^{L^f} \frac{\tilde{p}_i^f}{\mu_i^f}. \tag{9}$$

The sum here represents the mean duration of a type f *session, so that* ρ^f *is the overall traffic intensity of flow* f*, i.e., it is the product of mean arrival rate and mean session duration.*

Using Theorem 2 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*

$$\overline{\pi}(C) = \sum_{f \in \mathcal{F}} \left\{ \frac{\mathrm{P}(\mathcal{Q}'(f))\nu_1^f}{\sum_{g \in \mathcal{F}} \mathrm{P}(\mathcal{Q}'(g))\nu_1^g} \right\} \overline{\pi}_f(C). \tag{10}$$

Proof. For a flow f, entanglement requests are generated according to a Poisson process with rate ν_1^f only when all the associated users 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 2 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 (10). □

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 3. *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 durations.*

Proof. 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 next derive the expressions for the stationary distributions for the case where each attempt and calibration periods are exponentially distributed. If M_A^f and M_C^f are the number of attempt and calibration periods, respectively, for flow f (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

$$\mathbf{Z} = \left[\mathbf{Z}^{f_1}, \ldots, \mathbf{Z}^{f_F}\right] = \left[Z_i^f, 1 \le i \le M^f, f \in \mathcal{F}\right], \tag{11}$$

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 denoted by $\mathcal{S}'$, the set of $\mathbf{Z} \in \mathbb{N}^L$ satisfying

$$\sum_{f \in \mathcal{F}} \sum_{i=1}^{M^f} Z_i^f \le C, \sum_{f \in \mathcal{F}: n_k \in f} \sum_{i=1}^{M^f} Z_i^f \le c_k, \forall k \in \{1, \ldots, K\}.$$

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 jth period of a session belonging to flow f. The stationary distribution for this system is

$$\pi'(\mathbf{Z}) = \left(\sum_{\mathbf{Y} \in \mathcal{S}'} \prod_{f \in \mathcal{F}} \prod_{i=1}^{M^f} \frac{(\eta_i^f)^{Y_i^f}}{Y_i^f!} \right)^{-1} \prod_{f \in \mathcal{F}} \prod_{j=1}^{M^f} \frac{(\eta_j^f)^{Z_j^f}}{Z_j^f!}, \tag{12}$$

where period j's traffic intensity for flow f is $\eta_j^f = \omega_j^f/\theta_j^f$ if j corresponds to an attempt period, and $\eta_j^f = \omega_j^f/\sigma_j^f$ if it corresponds to a calibration period. Analogous to the Coxian case, $\omega_j^f = \nu_j^f$ if $j = 1$, else it is $\omega_j^f = \nu_1^f \zeta_{1,2}^f \cdots \zeta_{j-1,j}^f$, 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 jth

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 jth period, depending on whether it is an attempt or calibration period, respectively,

$$\frac{1}{\theta_j^f} = \sum_{i=1}^{N_A^f} \frac{r_{i,j}^f}{\mu_{i,j}^f}, \quad \text{or} \quad \frac{1}{\sigma_j^f} = \sum_{i=1}^{N_C^f} \frac{r_{i,j}^f}{\mu_{i,j}^f}, \tag{13}$$

where $r_{i,j}^f$ denotes the probability of reaching the ith phase of the jth period starting from its initial phase, and $\mu_{i,j}^f$ denotes the average duration of the ith 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, $\zeta_j^f = P_j^f$ in this case, where P_j^f is the probability of leaving the queueing network during the jth period of a flow-f session in the Coxian scenario, as computed in Lemma 1. 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 in Section III.

In the Coxian scenario, we can rewrite the state representation as $\mathbf{x} = [\mathbf{x}_1^{f_1}, \ldots, \mathbf{x}_{M^{f_1}}^{f_1}, \ldots, \mathbf{x}_1^{f_F}, \ldots, \mathbf{x}_{M^{f_F}}^{f_F}]$, 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 jth period of sessions belonging to flow f. Then for $\mathbf{Z} \in \mathcal{S}'$,

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

where D is the normalizing constant of the Coxian distribution (see (4)), N_j^f is the number of phases in the jth 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 ith phase of this period. Multiple applications of the multinomial theorem yield

$$\sum_{\mathbf{x}: q_m^f(\mathbf{x}_m^f) = Z_m^f, \forall m, f} \pi(\mathbf{x}) = D \prod_{f \in \mathcal{F}} \prod_{j=1}^{M^f} \frac{1}{Z_j^f!} \left(\sum_{i=1}^{N_j^f} \rho_{i,j}^f \right)^{Z_j^f}. \tag{14}$$

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

$$\begin{aligned} \sum_{i=1}^{N_j^f} \rho_{i,j}^f &= \nu_1^f \sum_{i=1}^{N_A^f} \frac{\tilde{p}_{i,j}^f}{\mu_{i,j}^f} = \nu_1^f \prod_{k=1}^{j-1} \left(1 - P_k^f\right) \sum_{i=1}^{N_A^f} \frac{r_{i,j}^f}{\mu_{i,j}^f} \\ &= \frac{\nu_1^f}{\theta_j^f} \prod_{k=1}^{j-1} \left(1 - \zeta_k^f\right) = \frac{\nu_1^f}{\theta_j^f} \prod_{k=1}^{j-1} \zeta_{k,k+1}^f = \eta_j^f. \end{aligned} \tag{15}$$

We can use similar arguments to show that when j corresponds to a calibration period we get $\eta_j^f = \omega_j^f / \sigma_j^f$, and D is the same for both Coxian and exponential distribution cases. By combining these arguments we obtain $\sum_{\mathbf{x}: q_m^f(\mathbf{x}_m^f) = Z_m^f, \forall m, f} \pi(\mathbf{x}) = \pi'(\mathbf{Z})$, confirming insensitivity. □

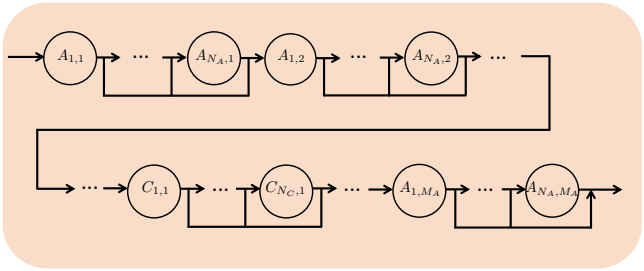

Fig. 4: A session from the multiple EPR pair generation strict resource reservation service model, shown at the level of periods in Figure 2, decomposed into exponentially-distributed phases so as to result in Cox-distributed attempt and calibration periods.

B. Multiple EPR Pair Generation, 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 (albeit, not necessarily all phases). Consequently, as shown in Figure 4, transitions to outside of the queueing network are only permitted from the final attempt period of the session. The overall system is thus very similar to that of Section V-A, and all previous assumptions hold, with the exception that $p_i^f = 0$, $\forall f \in \mathcal{F}$, whenever i belongs to a phase other than that of the last period. This modification merely affects the "overall traffic intensity" ρ, but does not change the form of the stationary distribution or the blocking probability.

C. Multiple EPR Pair Generation, Resource Relinquishment

We next study a type of scenario in which flows relinquish resources, such as during calibration, thereby inducing what we refer to as "idle" periods. There are several other situations where a flow is amenable to giving up its EGS resource module: for instance, depending on the application/protocol, nodes may wish to perform processing in-between entanglement generation attempts. Further, flows may wish to relinquish the resources during such processing periods, but not during calibration periods, or vice-versa. In this service mode, a retrial model is necessary, since a flow that has relinquished its EGS resource must re-obtain it to continue service after an idle period. We opt for the jump-over blocking mechanism, wherein a flow, when blocked, transitions the session to the beginning of the next idle period, or ends the session if no idle periods remain. Figure 5 depicts the general form of a session within this service mode: note that periods labeled $\mathcal{A}_i$ are not necessarily entanglement generation attempts – we now refer to them as "active" periods, which unlike idle periods *do* engage EGS resources.

Figure 6 depicts a session at phase-level detail. For the analysis that follows, we assume that transitions to outside

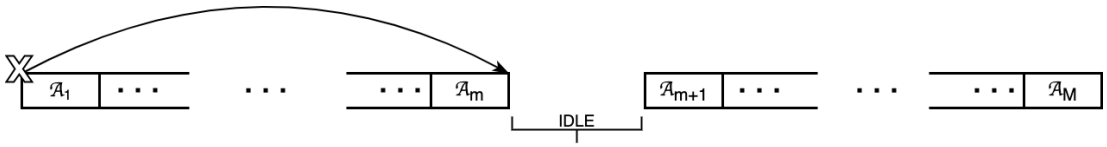

Fig. 5: Service model with resource relinquishment and jump-over blocking. A session consists of multiple "active" (periods engaging an EGS resource module), denoted by $\mathcal{A}_i$, $i = 1, \ldots, M$, interspersed with idle periods. In this model, a blocked session goes to the beginning of the next idle period, if there is one, or terminates if no idle periods remain in the session.

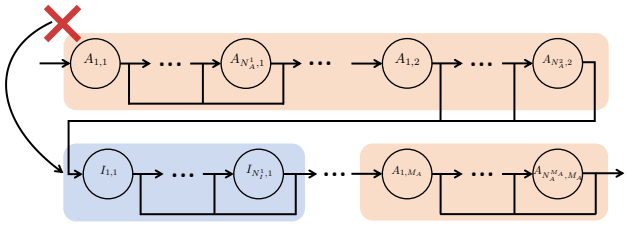

Fig. 6: A session with jump-over blocking, shown at the level of periods in Figure 5, decomposed into exponentially-distributed phases so as to result in Cox-distributed active and idle periods. Transitions to outside the queueing network are permitted only from the last period of the session.

of the queueing network are not permitted from any phases that do not belong to the last period of the session. The state space of this system is given by the set

$$\mathcal{S}'' = \left\{ \mathbf{x} \in \mathbb{N}^L : \sum_{f \in \mathcal{F}} \tilde{q}^f(\mathbf{x}) \leq C, \sum_{f \in \mathcal{F}: n_k \in f} q^f(\mathbf{x}^f) \leq c_k, 1 \leq k \leq K \right\}, \tag{16}$$

where for a given state $\mathbf{x}$, $\tilde{q}^f(\mathbf{x})$ is the number of active sessions of flow f: $\tilde{q}^f(\mathbf{x}) \equiv \sum_{j=1}^{M_A^f} \sum_{i=1}^{N_A^{f,j}} x_{i,j}^{f,A}$. Here, M_A^f is now defined as the number of active periods for a flow f session, and $N_A^{f,j}$ is the number of phases in the jth active period of the session. Since the subscript A now denotes any kind of active period and not only an attempt, we introduce a dependence on the specific period for the number of phases – this enables us to model generally-distributed period durations. Note that in this service mode, communication qubits are reserved for the entire duration of a session, including idle periods. This is reflected in the usage of $q^f(\mathbf{x}^f)$ in the second sum of (16).

The blocking probability for this service mode is of the same form as for the two previously discussed service modes, with the main differences again manifesting through the traffic intensities ρ_i^f. 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.

Theorem 4. *The stationary distribution $\pi(\mathbf{x})$ and the blocking probability of an active period $\overline{\pi}_i(C)$ take the same form as relations* (4) *and* (5)*, respectively.*

Proof. We first derive the stationary distribution $\pi(\mathbf{x})$, $\mathbf{x} \in \mathcal{S}''$ for the service mode with resource relinquishment and jump-over blocking as the retrial mechanism. In this service mode, 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 jth 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 Section V-B apply (i.e., $p_{i,j}^f(\mathbf{x}) = p_{i,j}^f$, $\forall f, i, j, \mathbf{x}$), with the following exceptions:

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

These amendments describe the jump-over blocking dynamics.

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 ith 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

$$\lambda_i^f(\mathbf{x}) = \begin{cases} \nu_1^f \gamma_i^f \equiv \lambda_i^f, & \text{if } \mathbf{x} + \mathbf{e}_i^f \in \mathcal{S}'', \\ 0, & \text{else.} \end{cases} \tag{17}$$

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 $\gamma_i^f = p_{j,j+1}^f \ldots p_{i-1,i}^f$. 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. By applying the local balance approach as in Section V-A, we can show that the stationary distribution $\pi(\mathbf{x})$ has the same form as in (4) (note that the definition of $\lambda_i^f(\mathbf{x})$ is now given by (17) throughout).

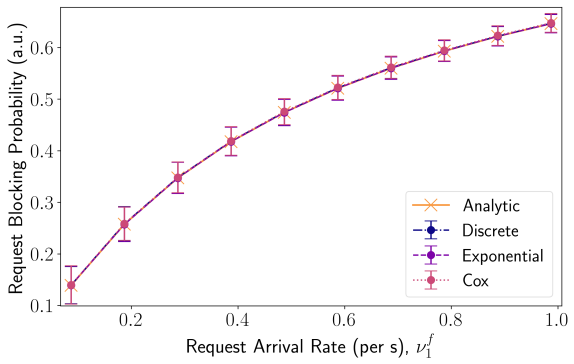

Fig. 7: Comparison of the average blocking probability per flow according to (5) with simulations for an EGS with one resource, connected to eight nodes, and serving $\binom{8}{2} = 28$ flows. Every node is restricted to a single communication qubit. Session traffic is homogeneous. The absolute relative errors are $\delta_{\text{discrete}} = 0.004, \delta_{\text{exponential}} = 0.001, \delta_{\text{Cox}} = 0.003$.

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 PASTA property so long as session arrival process is a Poisson process. For the latter, we utilize the result that "Departures See Time Averages" from [16, 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. The corollary nevertheless applies to our modified system, and the proof is identical.

From [16, Corollary 1], we conclude that for an active period that immediately follows an idle period, the blocking probability is the stationary probability that upon departure from said idle period, all EGS resource modules are engaged. As mentioned earlier, the same applies to the first period of a session. Finally, consider any active period that immediately follows another active period in the session: its blocking probability is equal to that of the first active period in the batch. In other words, if j is the active period under consideration, and i ($i < j$) is the closest idle period that precedes it (such that there are no other idle periods between i and j), then $i+1$ is the first active period in the batch that contains j and j has the same blocking probability as $i + 1$. If there are no idle periods preceding j, then j has the same blocking probability as the first active period in the session. We thus conclude that the blocking probability for an arbitrary active period has the same form as (5), with the only difference being the definitions of ρ_i^f. □

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 [28], 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 $\Delta t_{\text{attempt}}$.

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 of single attempt success, p_{gen}	1e-5 (a.u.)
Duration of a single attempt, T_{attempt}	115.072 μs
Attempt batch size	100
# batches (strict allocation) or re-trials (jump-over)	10
Total calibrations (idle periods) in strict (jump-over) modes	9

TABLE I: Physical parameters used in simulations correspond to an EGS supporting batched single click HE generation for NV colour centers in diamond as quantum nodes [28].

To validate our analysis, we simulate three models of entanglement generation experiments. These correspond to *discrete*, *Coxian*, and *exponential* distributions for the durations of periods. To ensure the simulations are compatible with our analysis two key assumptions are made. First, in all simulation modes and in numeric evaluation of (5) we fix the mean duration of every attempt (calibration period) to some value T_{attempt} (T_{calib}). In discrete simulations the duration of each attempt (calibration period) is set exactly to these values. Settings (number of phases, duration of phases, transition probabilities between phases) for the Coxian distribution, are chosen to ensure (13) is satisfied. 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. Simulation parameters are detailed in Table I.

To quantify agreement between numerical evaluation of (5) 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,

$$\delta_{\text{sim. type}} := \frac{|\max_x(y_{\text{Analytic}}[x] - y_{\text{Sim.}}[x])|}{y_{\text{Analytic}}[\arg\max_x(y_{\text{Analytic}}[x] - y_{\text{Sim.}}[x])]}, \quad (18)$$

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 equivalent to 1150.73 seconds of *simulated real-time*. 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 all independent runs of the standard deviation in blocking probabilities between flows.

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 V. In what follows, these service models are referred to simply as *strict single*, *strict multiple* and *jump-over*, respectively.

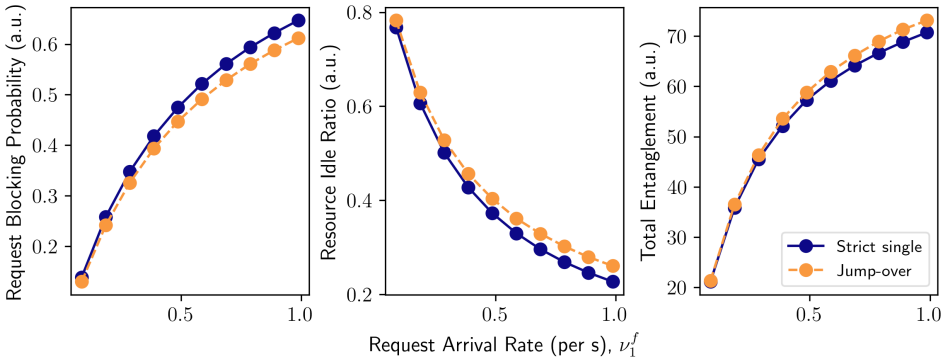

Fig. 8: Comparison of the strict single and jump-over service models, for an EGS with one resource, serving eight nodes via $\binom{8}{2} = 28$ flows. Every node is restricted to a single communication qubit. Session traffic is homogeneous. Left: blocking probability per request. Middle: proportion of time 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.

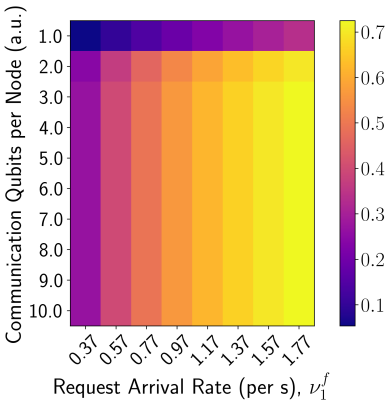

Fig. 9: Per-flow blocking probability heatmap. Data results from numeric evaluation of (5) for an EGS with two resources, serving eight nodes via 28 flows. Session traffic is homogeneous.

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 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 by comparing numeric evaluation of (5) with the simulation results for the cases of discrete, exponential, and Coxian distributions. Figure 7 shows results for an EGS operated in the strict single service model, with control of one resource ($C = 1$), connected to 8 nodes, each with a single communication qubit. We observe close agreement 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 3) 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 (18). These errors are $< 1\%$ for each simulation type. We validate the other service models for the same EGS configuration and obtain error parameters of $(\delta_{\text{discrete}}, \delta_{\text{exponential}}, \delta_{\text{Cox}}) = (0.061, 0.015, 0.006)$ for the strict multiple EPR pair service mode and $(0.001, 0.003, 0.004)$ for the jump-over service mode.

The performance of the strict single service model, 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, is contrasted in Figure 8. To highlight the effect of this resource relinquishment, the mean duration of the calibration and idle periods are set equal for these simulations. Load on the EGS network results from requests for resource access by the flows. Increased load directly leads to increased blocking probability and decreased idle time ratio in each service mode. When there is a relatively low-load on the EGS, the difference between each performance metric of the two service modes is marginal. When there is high load on the EGS, the jump-over service mode results in lower blocking probabilities, indicating better handling of the high load. The lower blocking probability of the jump-over 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. This indicates that for this EGS configuration, the jump-over service mode makes more efficient and productive use of the EGS resources. For the physically motivated simulation parameters used, the performance of the strict multiple service mode is not significantly different from that of the strict single service model, hence it is omitted from Figure 8. With these parameters, the expectation value of successful entanglements per session is 0.01.

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 mode with 2 (Figure 9) or 3 resources as the restriction varies from 1 to 10 communication qubits per node. In each case, we observe that increasing the number of communication qubits per node from 1 to 2 has a large impact on the blocking probability, but further increases have little impact. Numeric evaluation for strict single service mode scenarios where an EGS with 1, 2, or 3 resources is connected to 20 nodes and serves $\binom{20}{2} = 190$ flows confirm that the same effect holds for an EGS serving a large number of flows. We conclude that when an EGS controlled by the strict single service mode serves homogeneous traffic, an increase in the number of communication qubits per node from one to two has a large impact on the blocking probability, but further increases have a limited impact.

VII. Conclusion

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. 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 [12], 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.

References

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

[2] A. K. Ekert, "Quantum cryptography based on Bell's theorem," *Phys. Rev. Lett.*, vol. 67, pp. 661–663, Aug. 1991.

[3] P. Arright and L. Salvail, "Blind quantum computation," *International Journal of Quantum Information*, vol. 04, no. 05, pp. 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*. Atlanta, Georgia, USA: IEEE, Oct. 2009, pp. 517–526.

[5] D. Gottesman, T. Jennewein, and S. Croke, "Longer-Baseline Telescopes Using Quantum Repeaters," *Phys. Rev. Lett.*, vol. 109, p. 070503, Aug 2012.

[6] 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*, vol. 10, no. 8, pp. 582–587, Jun. 2014.

[7] V. Giovannetti, S. Lloyd, and L. Maccone, "Quantum-enhanced positioning and clock synchronization," *Nature*, vol. 412, no. 6845, pp. 417–419, 2001.

[8] 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*, vol. 16, no. 3, pp. 281–284, 2020.

[9] S. Gauthier, G. Vardoyan, and S. Wehner, "An Architecture for Control of Entanglement Generation Switches in Quantum Networks," *IEEE Transactions on Quantum Engineering*, vol. 4, pp. 1–17, 2023.

[10] S. L. Braunstein and A. Mann, "Measurement of the Bell operator and quantum teleportation," *Physical Review A*, vol. 51, no. 3, p. R1727, 1995.

[11] M. Michler, K. Mattle, H. Weinfurter, and A. Zeilinger, "Interferometric Bell-state analysis," *Physical Review A*, vol. 53, no. 3, p. R1209, 1996.

[12] J. Preskill, "Quantum Computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, Aug. 2018.

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

[14] B. A. Sevast'yanov, "An ergodic theorem for Markov processes and its application to telephone systems with refusals," *Theory of Probability & Its Applications*, vol. 2, no. 1, pp. 104–112, 1957.

[15] T. Bonald, "Insensitive Queueing Models for Communication Networks," in *Proceedings of the 1st International Conference on Performance Evaluation Methodolgies and Tools*, ser. valuetools '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 57–es.

[16] ——, "The Erlang model with non-Poisson call arrivals," *ACM SIGMETRICS Performance Evaluation Review*, vol. 34, no. 1, pp. 276–286, 2006.

[17] F. Kelly, "Charging and rate control for elastic traffic," *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

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

[19] ——, "On the exact analysis of an idealized quantum switch," *Performance Evaluation*, vol. 144, p. 102141, 2020.

[20] 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*. IEEE, 2023, pp. 1–10.

[21] 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)*, 2022, pp. 389–399.

[22] L.-M. Duan, M. D. Lukin, J. I. Cirac, and P. Zoller, "Long-distance quantum communication with atomic ensembles and linear optics," *Nature*, vol. 414, no. 6862, pp. 413–418, Nov. 2001.

[23] C. Cabrillo, J. I. Cirac, P. García-Fernández, and P. Zoller, "Creation of entangled states of distant atoms by interference," *Phys. Rev. A*, vol. 59, pp. 1025–1033, Feb. 1999.

[24] 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.*, vol. 98, p. 190503, May 2007.

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

[26] N. Sangouard, R. Dubessy, and C. Simon, "Quantum repeaters based on single trapped ions," *Phys. Rev. A*, vol. 79, p. 042340, Apr 2009.

[27] 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*, vol. 497, no. 7447, pp. 86–90, Apr. 2013.

[28] 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*, vol. 558, no. 7709, pp. 268–273, Jun. 2018.

[29] 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*, vol. 3, no. 8, pp. 538–541, 2007.

[30] 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.*, vol. 130, p. 050803, Feb. 2023.

[31] 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*, vol. 438, no. 7069, pp. 828–832, Dec. 2005.

[32] 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*, vol. 316, no. 5829, pp. 1316–1320, Jun. 2007.

[33] 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*, vol. 607, no. 7917, pp. 69–73, Jul. 2022.

[34] H.-K. Lo, M. Curty, and B. Qi, "Measurement-Device-Independent Quantum Key Distribution," *Phys. Rev. Lett.*, vol. 108, p. 130503, Mar 2012.

[35] S. L. Braunstein and S. Pirandola, "Side-Channel-Free Quantum Key Distribution," *Phys. Rev. Lett.*, vol. 108, p. 130502, Mar 2012.

[36] 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*, vol. 372, no. 6539, pp. 259–264, Apr. 2021.

[37] W. Whitt, "Continuity of Generalized Semi-Markov Processes," *Mathematics of Operations Research - MOR*, vol. 5, pp. 494–501, 11 1980.

[38] M. van der Heijden, "On the Three-Moment Approximation of a General Distribution by a Coxian Distribution," *Probability in the Engineering and Informational Sciences*, vol. 2, pp. 257 – 261, 04 1988.

[39] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, 1st ed. USA: Cambridge University Press, 2013.