Service-the-Longest-Queue Among d Choices Policy for Quantum Entanglement Switching

Guo Xian Yau*, Thirupathaiah Vasantam†, and Gayane Vardoyan‡
*Faculty of Electrical Engineering, Mathematics and Computer Science and QuTech, TU Delft, The Netherlands
†Department of Computer Science, Durham University, UK
‡Manning College of Information and Computer Sciences, University of Massachusetts, Amherst, USA

***Abstract*—An Entanglement Generation Switch (EGS) is a quantum network hub that provides entangled states to a set of connected nodes by enabling them to share a limited number of hub resources. As entanglement requests arrive, they join dedicated queues corresponding to the nodes from which they originate. We propose a load-balancing policy wherein the EGS queries nodes for entanglement requests by randomly sampling d of all available request queues and choosing the longest of these to service. This policy is an instance of the well-known power-of-d-choices paradigm previously introduced for classical systems such as data-centers. In contrast to previous models, however, we place queues at nodes instead of directly at the EGS, which offers some practical advantages. Additionally, we incorporate a tunable back-off mechanism into our load-balancing scheme to reduce the classical communication load in the network. To study the policy, we consider a homogeneous star network topology that has the EGS at its center, and model it as a queueing system with requests that arrive according to a Poisson process and whose service times are exponentially distributed. We provide an asymptotic analysis of the system by deriving a set of differential equations that describe the dynamics of the mean-field limit and provide expressions for the corresponding unique equilibrium state. Consistent with analogous results from randomized load-balancing for classical systems, we observe a significant decrease in the average request processing time when the number of choices d increases from one to two during the sampling process, with diminishing returns for a higher number of choices. We also observe that our mean-field model provides a good approximation to study even moderately-sized systems.**

***Index Terms*—entanglement generation switch, mean-field analysis, load-balancing**

I. Introduction

Quantum networks connect quantum-equipped devices to enable distributed applications that are not attainable via classical means alone. Examples include quantum computing in the cloud [1]–[3]; quantum-enhanced sensing [4], [5]; and quantum key distribution and conference key agreement [6]–[8]. While some of these applications are inherently entanglement-based, others can also consume entangled states for tasks such as remote state preparation [9] and quantum state and gate teleportation [10], [11]. Entanglement is thus an essential resource whose generation and distribution constitute much of the efforts undertaken within a quantum network.

This work is supported by QuTech NWO funding 2020–2024 Part I 'Fundamental Research', Project Number 601.QT.001-1, financed by the Dutch Research Council (NWO). We further acknowledge support from NWO QSC grant BGR2 17.269. TV was supported by the EPSRC funded INFORMED-AI project EP/Y028732/1.

In fiber optic-based quantum networks, photonic losses increase exponentially with distance [12], [13], rendering communication between quantum nodes infeasible without error correcting codes [14], [15] and assistive devices such as quantum repeaters or switches [16]–[20]. In this work, we study one such device – an Entanglement Generation Switch (EGS) [20], [21] – that can accommodate a number of connected nodes with bipartite entanglement generation. The EGS assists with this process by facilitating nodes' access to a limited number of shared hardware components (*e.g.*, Bell state analyzers (BSAs), as depicted in [21, Figure 2], capable of performing optical Bell state measurements on incoming photons), which we refer to as switch "resources". Figure 1 illustrates an EGS serving N nodes with m resources (the architecture is discussed in detail in Section II-A). The EGS resources in principle do not necessitate sophisticated technology such as quantum memories, easing requirements both on cost and fabrication in comparison to memory-equipped counterparts. These properties make the EGS highly relevant to near-term metropolitan-area quantum networks, thus motivating the architecture choice for this study.

Thus far, the EGS has been analyzed in settings where connected nodes are responsible for requesting service. Here, we propose a novel service mode wherein the onus of request solicitation falls on the switch: the EGS thus queries a fixed number of nodes and assigns a resource module to the node with the largest outstanding number of entanglement requests. This operation mode of the EGS is an instance of the so-called power-of-d-random-choices paradigm which has seen a wide variety of applications ranging from hashing to virtual circuit routing [22]. For brevity, we refer to the act of querying d system components (*e.g.*, queues) and selecting one for service as a "d-choices policy". Such policies have proven advantageous as load-balancing techniques in settings like data-centers or computer clusters. Here, an arriving request would ideally be assigned to the least loaded server/compute node, but access to full and up-to-date information about workloads might be unavailable or costly to obtain. Assigning the new task to the least loaded of $d \geq 2$ randomly chosen servers achieves a lower communication cost (relative to querying all servers) while considerably reducing the maximum server load (and therefore the average request response time) even with $d = 2$.

A factor contributing to incomplete information at the EGS during decision-making is node status, namely their readiness

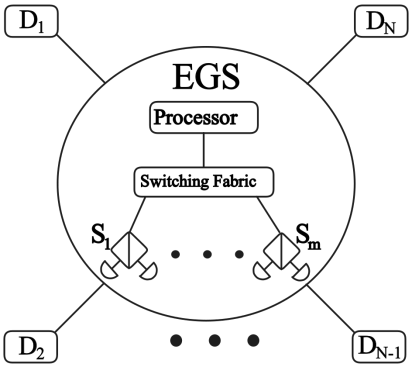

Fig. 1. Architecture of an EGS with m resources – here, BSAs (S_j). N nodes (D_i) are connected to the EGS via classical and quantum channels.

to commence entanglement generation. Each entanglement generation attempt requires nodes' communication qubits to be available for the coordinated emission of photons whose synchronized arrival is expected at an EGS resource. If queues are maintained by the EGS, then a possibility arises that a chosen request cannot be immediately serviced due to the respective nodes' unpreparedness to engage in entanglement generation: *e.g.*, a node may already be using its communication qubits to generate entanglement with other nodes, or these qubits may be involved in processing tasks that require their participation (*e.g.*, two-qubit gates for hardware platforms such as color centers in diamond [23]). To learn this information, the EGS must communicate with the node-pair that issued the request, and multiple communication messages/rounds may be necessary until the EGS finds a serviceable request. We thus have a practical reason to situate request queues at nodes: when the EGS queries d node-pairs for queue sizes, it receives most up-to-date information, including knowledge of which pairs are primed for entanglement generation. This strategy has the potential to amortize some of the communication delays associated with node status querying.

Yet another motivation for placing queues at nodes is that of fairness: the data-center model permits any node to flood the network with its entanglement requests, over time causing the bulk share of the resources to be dedicated to servicing its own demands. Making the network responsible for offering its services to the nodes removes the latter's ability to directly influence resource allocation within it. We envision that an EGS deploying a d-choices policy will provide the aforementioned benefits while at the same time adequately supporting a variety of applications such as distributed quantum computation or entanglement distribution in a high traffic regime – such an environment ensures that with a high probability, there exists a request within the set of queues sampled by the EGS.

We use mean-field analysis to obtain tight approximations of the average response time – the time gap between the entry and exit time of an entanglement request. Such analysis has been widely used to develop efficient algorithms for various computer and communication networks [24]. Due to our design choice of placing queues at nodes, the mean-field limit of the data-center model studied in [25], which has queues at servers, differs from our mean-field limit. Our task in this work is therefore to carry out a performance analysis of the proposed system, which we refer to as the d-choices EGS model. Since this study is the first to consider this operation setting of the EGS, we restrict its scope to a single, isolated instance of the device serving nodes in a star topology as depicted in Figure 1. Each node connected to the EGS is assumed to have a number of communication qubits equal to the number of switch resources, enabling it to concurrently participate in entanglement generation with all of them. Because of this assumption, we are able to analyze the model using mean-field techniques. Scenarios where nodes have memory restrictions are reserved for follow-up study. However, we cannot use mean-field techniques for such models; we elaborate more on this in Section III. Our contributions include the following:

- We derive a set of differential equations that, in the limit of a large number of request queues, accurately describe the system state evolution of an EGS deploying the d-choices policy. This modelling approach enables an asymptotic analysis of the system in the mean-field limit;
- We prove the existence and uniqueness of the equilibrium state that satisfies the mean-field equations;
- We provide analytical expressions for request queue size distributions as well as the average response time for an entanglement request. These performance measures are generally challenging to obtain for scenarios where multiple entities (*e.g.*, EGS resources) concurrently process requests.

Our numerical results support our findings. Namely, they provide supporting evidence that the equilibrium state of the mean-field approximates well the stationary distribution of the system state when the number of request queues is large. Further, we find that the average request response time sees a substantial decrease when d increases from one to two; a further increase in d provides diminishing advantages. We observe approximation errors of less than 5% for $d = 2$ even for a small system with ten resources and 200 entanglement request queues. We also observe that the approximation error of our model increases as d increases. These results demonstrate the potential of mean-field techniques to study near-term quantum systems where a small number of quantum devices serve a large number of applications.

The remainder of this manuscript is structured as follows. In Section II, we provide relevant background. In Section III, we construct the system model and use it to conduct a mean-field analysis of the system. In Section IV, we present numerical results. We conclude our findings and discuss future directions in Section V.

II. Background and Related Work

In this section, we describe the EGS architecture and the d-choices policy in detail. We then provide an overview of relevant literature.

A. Entanglement Generation Switch (EGS)

An EGS is a type of quantum entanglement switch (QES) that facilitates entanglement generation between multiple par-

ties. It consists of a central processor tasked with resource scheduling and classical communication with neighboring nodes; resources for entanglement generation – henceforth we assume for concreteness that these are BSAs although our model is applicable to other types of switch resources; and an optical switching fabric that serves as a bridge between switch interfaces and BSAs (see Figure 1). When two nodes wish to share a bipartite entangled state (*i.e.*, a Bell state/EPR pair [26]), they submit a request to the EGS. In this work, we consider the following request handling procedure: when a BSA is available to process a new request, the central processor first informs the nodes that a BSA has been allocated for entanglement generation. Meanwhile, the processor notifies the optical switch that the BSA should only be accessible to said node pair. The optical switch then configures the quantum channel accordingly, and the nodes can proceed with entanglement generation attempts. The result of each attempt is communicated to the nodes classically.

In near-term quantum networks, BSAs in the EGS are likely to be few in number due to device fabrication cost and complexity; hence we assume that they comprise a shared pool of resources as in [20] and [21]. A key challenge for the EGS is thus resource management. While the authors of [20] and [21] managed these resources via request rate control and blocking mechanisms, respectively, our contribution is to study the EGS using the d-choices load-balancing policy and to understand queue behavior and the average request processing time. For more details on prior work for the EGS see Section II-C.

B. The Traditional d-Choices Policy

The d-choices policy is a load-balancing scheme commonly seen in classical data-center models [25]. It offers both the benefits of the Join-a-Random-Queue (JRQ) and the Join-the-Shortest-Queue (JSQ) policies; JSQ often has better performance in terms of average request waiting time, whereas JRQ incurs a lower classical communication cost. In a classical data-center with m servers, the d-choices policy is referred to as the $\mathrm{JSQ}(d)$ policy and is implemented as follows: every server has a queue to store assigned jobs while a job dispatcher assigns an incoming job to the shortest queue size of d randomly chosen servers. One can easily see that $\mathrm{JSQ}(1)$ and $\mathrm{JSQ}(m)$ correspond to the JRQ and JSQ, respectively. One way to analyse $\mathrm{JSQ}(d)$ is through a mean-field analysis which we introduce in Section III-B. For a more comprehensive summary of techniques developed to study the d-choices policy in the data-center setting, we refer the reader to [22].

C. Related Work

Quantum memory-equipped QESes, sometimes referred to as entanglement distribution switches (EDSes) in the literature, have been studied extensively, *e.g.*, as in [19], [27]–[32]. Architectural differences between EDSes and EGSes give rise to different modes of operation. For instance, the presence of quantum memories in an EDS enables more powerful functionality than that of an EGS, *e.g.*, deterministic entanglement swapping [33]–[35], entanglement distillation [36]–[38], and overall more opportunities for dynamic or adaptive decision-making by virtue of being able to store entanglement at the link level (*i.e.*, switch-to-node entanglement). These operational differences result in EDS models that are not directly applicable to our proposed EGS scheme.

Since the limiting resource for an EGS is the BSA, a resource management algorithm is required to facilitate its sharing. Gauthier *et al.* proposed and analyzed a rate-modulation mechanism in [20], as well as a request-blocking mechanism in [21]. Specifically, the scheme in [20] allows node-pairs to submit target entanglement generation rates to the EGS, which are then updated using a rate control protocol to achieve optimal performance according to a metric such as throughput. Our proposed scheme contrasts from this setup in that we assume no control over entanglement demand rates: these are fixed and predetermined in our model. In [21], each node-pair wishing to communicate submits an entanglement request to the EGS. This request is blocked if all BSAs are busy servicing other requests, and the devices will need to resubmit their requests at a later time. This model contrasts from ours in that the former does not allow request queueing.

In this work, we introduce the d-choices load-balancing policy as a novel alternative for resource management within an EGS. We use a mean-field limit approach to analyze the system's asymptotic behaviour as the number of queues grows while the server-to-queue ratio stays constant. Previously, a Service-the-Longest-Queue among d-choices policy was studied in the context of wireless networks [39]. Their model involves one transceiver and k mobile stations, each with a dedicated wireless transmission channel that is available for data transmission only probabilistically due to fluctuations in channel conditions. Each mobile station has a queue to store its packets that are awaiting transmission. The transceiver transmits a packet from one mobile station at a time; the latter must have access to an available wireless channel. Whenever the transceiver becomes idle, it samples d stations at random from the set of all mobile stations with available wireless channels, and chooses the longest of these d queues for processing. The performance of this system was analyzed through mean-field techniques. Our model is different from that of [39] as we have multiple servers. In addition, the job service rate and the rate with which an idle server re-samples queues can be different in our model, while they were assumed to be the same in [39]. In [39], the authors also studied the policy where the server always processes a job from the longest of all queues with available wireless channels, when the number of queues becomes large. We do not study such a policy as it has a high communication cost.

III. Model and Mean-field Analysis

In this section, we introduce our d-choices EGS model and state our assumptions. We then perform a mean-field analysis by deriving the system's mean-field equations and expressions describing the corresponding equilibrium state.

A. Model and Assumptions

Since we model the EGS as a queueing system, in the following discussion we introduce terminology that will make our subsequent comparison to the data-center model (described in Section II-B) straightforward. Recall that m denotes the number of BSAs within the EGS and that N is the number of nodes connected to the EGS, each with m communication qubits. The BSAs effectively function as *servers*, and we refer to them as such throughout the analysis. We define a *flow* f to be a node-pair, $f = (i, j)$, $1 \leq i < j \leq N$, that desires entanglement. We denote the set of all flows with $\mathcal{F}$, *i.e.*, $\mathcal{F} \subseteq \{(i, j) : 1 \leq i < j \leq N\}$, and $n = |\mathcal{F}|$ the total number of flows. Clearly $1 \leq n \leq \binom{N}{2}$. A *job* is an entanglement request from a flow $f \in \mathcal{F}$. Each job represents the creation of exactly one entanglement between the two nodes of f. We assume all jobs of a flow have identical service time distributions – a reasonable assumption in settings where parameters that drive the entanglement generation process are constant and involuntary parameter drift is insignificant.

We assume that flow f's jobs arrive according to a Poisson process with parameter λ_f and that jobs from different flows arrive independently from each other. Each flow has a queue where jobs await service on a First-Come-First-Served (FCFS) basis. The term *queue size* refers to the number of existing jobs in a queue. We assume successful entanglement generation for flow f has an exponential service time distribution with rate μ_f, which reflects the more realistic view of entanglement generation as a succession of Bernoulli trials with low success probability. Since BSAs can operate in parallel and do not affect each other, we assume servers process jobs independently. Finally, we consider a homogeneous system that satisfies $\lambda_f = \lambda$ and $\mu_f = \mu$, $\forall f \in \mathcal{F}$.

Upon job completion, a (newly idle) server immediately samples d queues at random to acquire a new job to process. Since in our model queues are situated at flows (*i.e.*, both nodes of a flow's node-pair keep track of enqueued jobs, but the EGS does not do any active tracking), all d selected queues may be empty. In this event, the server is said to have carried out a *failed sampling*, and remains idle for a period that is exponentially distributed with parameter γ, before re-sampling d (potentially different) queues. This mechanism exerts a lighter classical communication strain within the system compared to immediate re-sampling. We thus refer to γ as the back-off rate. In a near-term EGS system, we expect that entanglement requests will experience a large service time due to photon losses in fiber and failed optical entanglement swaps at BSAs; the latter has a 0.5 success probability without ancilla qubits [40]. It is thus desirable to have $\gamma > \mu$ so that BSAs have shorter idle periods.

Since the EGS serves the longest among d randomly selected queues, we say it follows the $\mathrm{SLQ}(d)$ policy. Then $\mathrm{SLQ}(1)$ and $\mathrm{SLQ}(n)$ are the Service-a-Random-Queue (SRQ) and the Service-the-Longest-Queue (SLQ) policies, respectively. We study the effects of queue placement by comparing the performance of JSQ(d) and SLQ(d) in Section IV.

B. Derivation of Mean-field Equations

In the analysis that follows, $\mathbb{N}_0$ represents the non-negative integers and $\mathbb{E}[\cdot]$ is the expectation operator. With our homogeneity assumption, we denote all arrival rates with λ and all service rates as μ. We define $r \equiv \frac{m}{n}$ as the server-to-queue ratio. We next define the following random variables:

- $\tilde{X}_i(t)$, $i \in \mathbb{N}_0$, is the number of flows with at least i jobs at time t;
- $X(t) = (X_i(t))_{i \in \mathbb{N}_0}$ where $X_i(t) = \frac{1}{n}\tilde{X}_i(t)$ is the fraction of flows with at least i jobs at time t;
- $\tilde{Y}(t)$ is the number of servers servicing a job at time t;
- $Y(t) = \frac{1}{m}\tilde{Y}(t)$ is the fraction of servers servicing a job at time t.

By convention, we write $\overline{\alpha} = 1 - \alpha$ and $\overline{\alpha}^d = (1 - \alpha)^d$ for $\alpha \in [0, 1]$. For example, $X_i^d(t) = (X_i(t))^d$, $\overline{Y}(t) = 1 - Y(t)$, and $\overline{X}_i^d(t) = (1 - X_i(t))^d$. We note that the empirical process $\{(X(t), Y(t))\}_{t \geq 0}$ is a Markov process.

Remark 1. *Since each node has m communication qubits, it may participate in entanglement generation at all BSAs simultaneously. Hence, a Markovian representation need not track the identity of each flow currently in service. In the scenario where nodes have fewer than m communication qubits, a Markovian representation must include identities of flows currently being serviced, since in this case $\{(X(t), Y(t))\}_{t \geq 0}$ is not a Markov process. As a result, mean-field analysis that requires the empirical process to be a Markov process is not applicable. The average response time of models with nodes having less than m communication qubits will be lower bounded by our model (nodes have m communication qubits).*

Next we derive mean-field equations without giving a formal proof of the existence of the mean-field limit due to space constraints. This proof follows easily from the theory of the convergence of Markov processes as in [25], [41].

Theorem 1. *The* mean-field equations (MFEs) *of the* $\mathrm{SLQ}(d)$ *policy applied to the EGS are given by*

$$\frac{d}{dt}x_i(t) = \lambda(x_{i-1}(t) - x_i(t)) - r(\gamma\overline{y}(t) + \mu y(t))(\overline{x}_{i+1}^d(t) - \overline{x}_i^d(t)), \quad (1)$$

$$\frac{d}{dt}y(t) = \gamma\overline{y}(t)(1 - \overline{x}_1^d(t)) - \mu y(t)\overline{x}_1^d(t), \quad (2)$$

where $i \geq 1$ and $x_0(t) = 1$ for all $t \in [0, \infty)$. We refer to (1) *as the flow equations and to* (2) *as the server equation. The process $(x(t), y(t))_{t \geq 0}$ is called the mean-field limit that represents the limit of $\{(X(t), Y(t))\}_{t \geq 0}$ as $n \to \infty$, where $x(t) = (x_i(t), i \geq 0)$.*

An intuitive explanation for the MFEs is as follows: consider first the flow equations (1). Fix an arbitrary $i \in \mathbb{N}_0$, a sufficiently small $\Delta t > 0$, and let $t \in [0, \infty)$. The following three random variables contribute to change in $\tilde{X}_i(t)$:

- U_i is the change in $\tilde{X}_i(t)$ due to job arrivals during the period $[t, t + \Delta t]$;
- V_i is the change in $\tilde{X}_i(t)$ due to a successful re-sampling by an idle server during the period $[t, t + \Delta t]$;

- W_i is the change in $\tilde{X}_i(t)$ due to a successful sampling by a server upon job completion during $[t, t+\Delta t]$.

The expected number of new job arrivals during $[t, t+\Delta t]$ is $n\lambda\Delta t$. Moreover, the probability that a new job arrives at a queue with $i-1$ jobs is $X_{i-1}(t) - X_i(t)$. Thus,

$$\mathbb{E}[U_i] = n\lambda\Delta t(X_{i-1}(t) - X_i(t)). \tag{3}$$

Next, consider V_i and W_i. When a server selects d queues at random, $(\overline{X}_{i+1}^d(t) - \overline{X}_i^d(t))$ is the probability that they all have at most i jobs, with at least one queue having exactly i jobs. We assume that queues are sampled with replacement which is a valid assumption when n is large. Further, the expected number of re-sampling and sampling operations in a Δt-sized time interval is given by $m\overline{Y}(t)\gamma\Delta t$ and $mY(t)\mu\Delta t$, respectively. We thus obtain

$$\mathbb{E}[V_i] = m\overline{Y}(t)\gamma\Delta t(\overline{X}_{i+1}^d(t) - \overline{X}_i^d(t)), \tag{4}$$

$$\mathbb{E}[W_i] = mY(t)\mu\Delta t(\overline{X}_{i+1}^d(t) - \overline{X}_i^d(t)). \tag{5}$$

The expected change in $\tilde{X}_i(t)$ during $[t, t+\Delta t]$ is therefore

$$\begin{aligned}\mathbb{E}[U_i - V_i - W_i] = \Delta t\big[&n\lambda(X_{i-1}(t) - X_i(t)) \\ &- m(\gamma\overline{Y}(t) + \mu Y(t))(\overline{X}_{i+1}^d(t) - \overline{X}_i^d(t))\big].\end{aligned} \tag{6}$$

Dividing by n and Δt, we obtain

$$\begin{aligned}&\frac{1}{\Delta t}\mathbb{E}[X_i(t+\Delta t) - X_i(t) |\ (X(t), Y(t))] = \\ &\lambda(X_{i-1}(t) - X_i(t)) - r(\gamma\overline{Y}(t) + \mu Y(t))(\overline{X}_{i+1}^d(t) - \overline{X}_i^d(t)).\end{aligned} \tag{7}$$

To derive (2), we observe that a change occurs in $Y(t)$ in the interval $[t, t+\Delta t]$ when a busy server completes a job and fails to obtain a new job via d-choices sampling, or when an idle server becomes busy via successful d-choices sampling. The probability of a failed sampling is given by the probability that all sampled d queues are empty (again assuming sampling with replacement), *i.e.*, $\overline{X}_1^d(t)$. Finally, to obtain MFEs we replace $(X(t), Y(t))$ with the mean-field $(x(t), y(t))$ (which is a deterministic process) in the flow and server drift equations, and let $\Delta t \to 0$.

As a comparison, the MFEs for the JSQ(d) model of [25], which assumes $r = 1$,

$$\frac{d}{dt}x_i(t) = \frac{\lambda}{r}(x_{i-1}^d(t) - x_i^d(t)) - \mu(x_i(t) - x_{i+1}(t)). \tag{8}$$

Due to the queue placement in this model, the set of equations above captures the dynamics of the server queues, with the equilibrium point given by $\pi_i = (\frac{\lambda}{r\mu})^{\frac{d^i-1}{d-1}}$ for $d \geq 2$.

C. Equilibrium State of the Mean-Field

Given the MFEs, we can characterize the system's performance using the equilibrium state, or the fixed point of the mean-field limit. An equilibrium state of the mean-field limit is the solution satisfying $\frac{dy(t)}{dt} = 0$ and $\frac{dx_i(t)}{dt} = 0$ for all $i \geq 1$. Let us denote by (π, ε) the fixed point satisfying the MFEs, where $\pi = (\pi_i)_{i\geq 0}$ is an infinite sequence with $\pi_i \in [0,1]$ for all $i \geq 0$ satisfying the flow equations (1) and $\varepsilon \in [0,1]$ satisfying the server equation (2).

Next, we characterize the equilibrium state of the mean-field. In the following theorem (Theorem 2), the equilibrium state of the mean-field exists as long as $\varepsilon = \frac{\lambda}{r\mu} < 1$, which leads to $\frac{\lambda}{r(\gamma\overline{\varepsilon}+\mu\varepsilon)} < 1$. Note that $\frac{\lambda}{r\mu} < 1$ is the necessary condition for the stability of the system. Here, the probability with which a server is busy equals $\frac{\lambda}{r\mu}$ and $r(\gamma\overline{\varepsilon} + \mu\varepsilon)$ is the rate at which a queue is selected for processing when $d = 1$. It is of interest to find equilibrium states (π, ε) that satisfy $\sum_{i\geq 1}\pi_i < \infty$ (indicating finite average queue size under π) as we want to approximate the stationary distribution of a stable system with a finite average queue size.

Theorem 2. *Among the class of equilibrium states (π, ε) with $\sum_{i\geq 1}\pi_i < \infty$, if $\frac{\lambda}{r\mu} < 1$ there exists a unique equilibrium state of the mean-field. Furthermore, this equilibrium state satisfies the following recursive equations*

$$\pi_{i+1} = 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\pi_i\right)^{\frac{1}{d}}, \tag{9}$$

where $i \geq 1$, $\varepsilon = \frac{\lambda}{r\mu}$, $\pi_0 = 1$, and $\pi_1 = 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon}+\mu\varepsilon)}\right)^{\frac{1}{d}}$.

Proof. Since (π, ε) is the equilibrium state of the MFEs, it follows that $\frac{d}{dt}\pi_i = 0$ for all $i \in \mathbb{N}_0$ and $\frac{d}{dt}\varepsilon = 0$. This property, along with (1) and (2), means that (π, ε) satisfies

$$\lambda(\pi_{i-1} - \pi_i) - r(\gamma\overline{\varepsilon} + \mu\varepsilon)(\overline{\pi}_{i+1}^d - \overline{\pi}_i^d) = 0, \tag{10}$$

$$\gamma\overline{\varepsilon}(1 - \overline{\pi}_1^d) - \mu\varepsilon\overline{\pi}_1^d = 0. \tag{11}$$

Next, for $i \in \mathbb{N}_0$, let $q_i \coloneqq \pi_i - \pi_{i+1}$; then using (10),

$$\begin{aligned}\sum_{i=0}^{j} q_i &= \sum_{i=0}^{j}\frac{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}{\lambda}(\overline{\pi}_{i+2}^d - \overline{\pi}_{i+1}^d) \\ &= \frac{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}{\lambda}(\overline{\pi}_{j+2}^d - \overline{\pi}_1^d).\end{aligned} \tag{12}$$

On the other hand, $\sum_{i=0}^{j} q_i = \pi_0 - \pi_{j+1} = 1 - \pi_{j+1}$. Using this with (12) yields

$$1 - \pi_{j+1} = \frac{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}{\lambda}(\overline{\pi}_{j+2}^d - \overline{\pi}_1^d). \tag{13}$$

Since we are interested in equilibrium states that satisfy $\sum_{j\geq 1}\pi_j < \infty$, we use the condition that $\lim_{j\to\infty}\pi_j = 0$ and $\lim_{j\to\infty}\overline{\pi}_j = 1$. Applying the limit to both sides of (13) results in

$$1 = \lim_{j\to\infty}\frac{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}{\lambda}(\overline{\pi}_{j+2}^d - \overline{\pi}_1^d) = \frac{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}{\lambda}(1 - \overline{\pi}_1^d). \tag{14}$$

Rearranging the above equation, we get

$$\pi_1 = 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\right)^{\frac{1}{d}}. \tag{15}$$

By substituting (15) into (11), we obtain $\varepsilon = \frac{\lambda}{r\mu}$. Next, we show that (9) is valid. By rearranging (10), we obtain

$$\pi_{i+1} = 1 - \left(\overline{\pi}_i^d + \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}(\pi_{i-1} - \pi_i)\right)^{\frac{1}{d}}. \tag{16}$$

In (16), by choosing $i = 1$, we obtain

$$\pi_2 = 1 - \left(\overline{\pi}_1^d + \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}(1 - \pi_1)\right)^{\frac{1}{d}}. \tag{17}$$

By substituting (15) into $\overline{\pi}_1^d$ of (17) we get

$$\pi_2 = 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\pi_1\right)^{\frac{1}{d}}. \tag{18}$$

Similarly, for $i \geq 2$, by expanding the expression for $\overline{\pi}_i^d$ in (16) we obtain (9).

Next, we show that the equilibrium state that we found satisfies $\sum_{i\geq 1} \pi_i < \infty$. It suffices to show that $\pi_i \leq \pi_i^*$ by induction on $i \geq 1$, where (π^*, ε^*) is the equilibrium state of the mean-field when $d = 1$, and $\pi^* = (\pi_i^*, i \geq 0)$. Furthermore, it can be checked that $\sum_{i\geq 1} \pi_i^* = \frac{\rho}{1-\rho}$ where $\rho = \frac{\lambda}{r(\gamma\overline{\varepsilon}+\mu\varepsilon)}$. Observe that $1 - p^{\frac{1}{d}} \leq 1 - p$ for any $p \in [0, 1]$ and $d \in \mathbb{N}$. For the base case, we get

$$\pi_1 = 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\right)^{\frac{1}{d}} \tag{19}$$

$$\leq 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\right) \tag{20}$$

$$= \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)} = \pi_1^*. \tag{21}$$

For the inductive step, we can rewrite (9) and get

$$\pi_{i+1} = 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\pi_i\right)^{\frac{1}{d}} \tag{22}$$

$$\leq 1 - \left(1 - \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\pi_i\right) \tag{23}$$

$$\leq \frac{\lambda}{r(\gamma\overline{\varepsilon} + \mu\varepsilon)}\pi_i^* = \pi_{i+1}^*. \tag{24}$$

Thus, $\pi_{i+1} \leq \pi_{i+1}^*$ whenever $\pi_i \leq \pi_i^*$ as desired. □

It is important to note that we still need to prove that the equilibrium state of the mean-field yields the stationary probability distribution of a queue as π and the stationary probability that a server is busy as ε when $n \to \infty$. A sufficient condition to establish this result is to show the global stability of the mean-field, which we leave for follow-up work. However, as we show in Section IV, our numerical results provide supporting evidence that the equilibrium state of the mean-field approximates the stationary distribution of the system when n is large.

IV. Numerical Results

In this section, we simulate an EGS deploying the d-choices policy and make comparisons to analytical results. The simulation parameters are configured with $N_{\text{arrivals}} = 10^9$ job arrivals per simulation run, service rate $\mu = 1$, back-off rate $\gamma = 1$, $N = 21$, $m = 10$, $n = 200 < \binom{N}{2}$ flows/queues for a server-to-queue ratio of $r = \frac{m}{n} = 0.05$, $\frac{\lambda}{r\mu} = 0.90$, and time units are given in seconds, unless otherwise specified. For our simulations, we choose $m \ll n$ as near-term EGSes will have a limited number of resources.

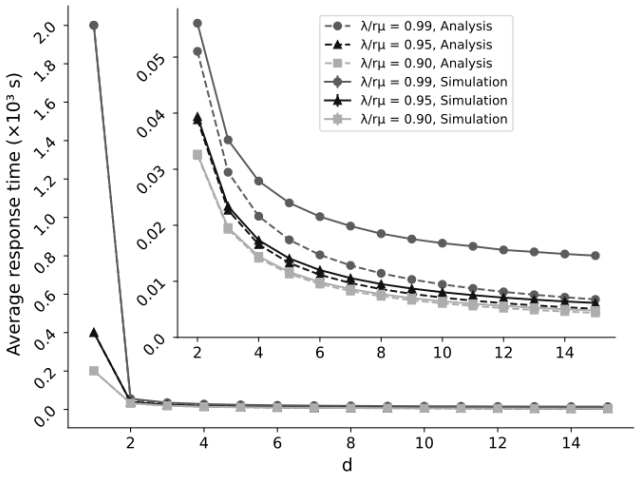

Fig. 2. Average response time of an EGS deploying the d-choices policy, as a function of the number of choices d.

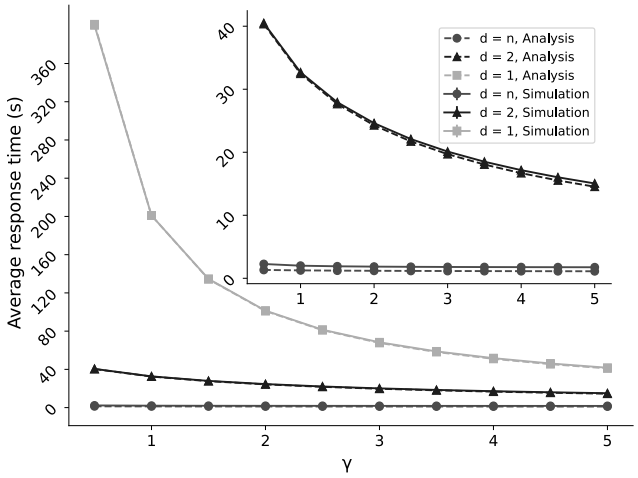

Fig. 3. Average response time of an EGS implementing d-choices policy as a function of back-off rate γ.

First, we explain how we compute the average response time using simulations. We denote by t_i and $u_i > t_i$ the arrival and departure time of job i, respectively. The *response time* of job i is $w(t_i) := u_i - t_i$. The number of jobs departed by time $t > 0$ is denoted by $N_{\text{departures}}(t)$. The *average response time at time* $t > 0$ is given by $\overline{w}(t) = \frac{1}{N_{\text{departures}}(t)} \sum_{t_i<t} w(t_i)$. The *average response time* over an entire simulation run is $\overline{w} = \overline{w}(t_{N_{\text{arrivals}}})$, where $t_{N_{\text{arrivals}}}$ is the arrival time of the last job in the simulation. Let us denote by $\tilde{w}_\pi$ the analytically derived average waiting time at equilibrium. By Little's Law [42, Section 13.7], we have $\tilde{q}(\pi) = \lambda\tilde{w}_\pi$, where $\tilde{q}(\pi) = \sum_{i\geq 1} \pi_i$ is the average queue size at equilibrium, and λ is the request arrival rate. Thus, $\tilde{w}_\pi = \frac{\tilde{q}(\pi)}{\lambda}$. Given the simulation-based average response time $\overline{w}$ and its analytical approximation $\tilde{w}_\pi + 1/\mu$, the percentage error between the two values is computed as $100 \times \left|1 - \frac{\tilde{w}_\pi + \frac{1}{\mu}}{\overline{w}}\right|$ %.

From Figure 2 we observe that for a fixed $\lambda/r\mu$, increasing the number of choices d reduces the average response time. The effect is most prominent when increasing d from one to two; a further increment in d exhibits a diminishing gain

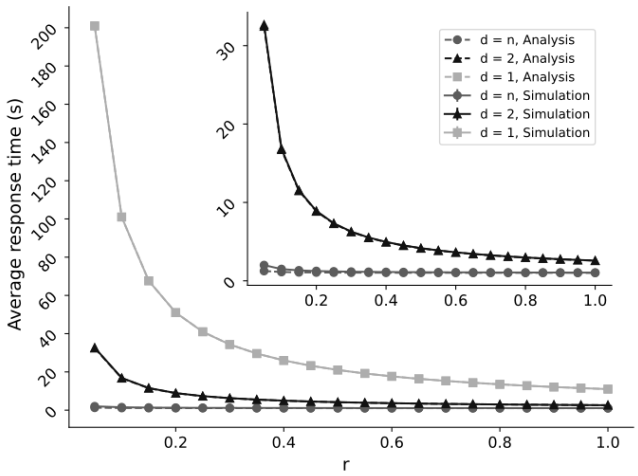

Fig. 4. Average response time as a function of server-to-queue ratio $r = \frac{m}{n}$, with varying m and fixed $n = 200$.

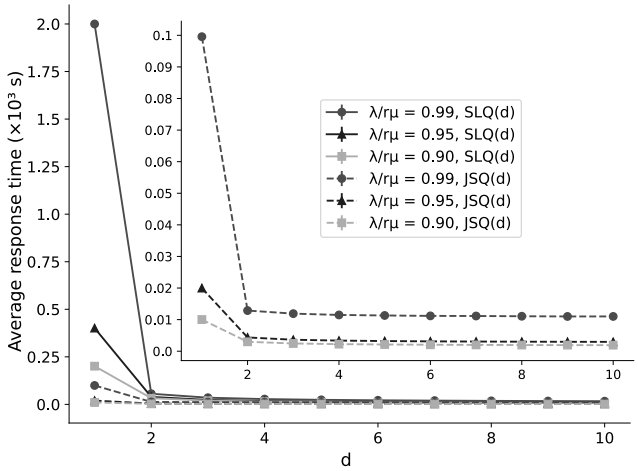

Fig. 5. Average response time of $\mathrm{JSQ}(d)$ and $\mathrm{SLQ}(d)$ as a function of the number of choices d.

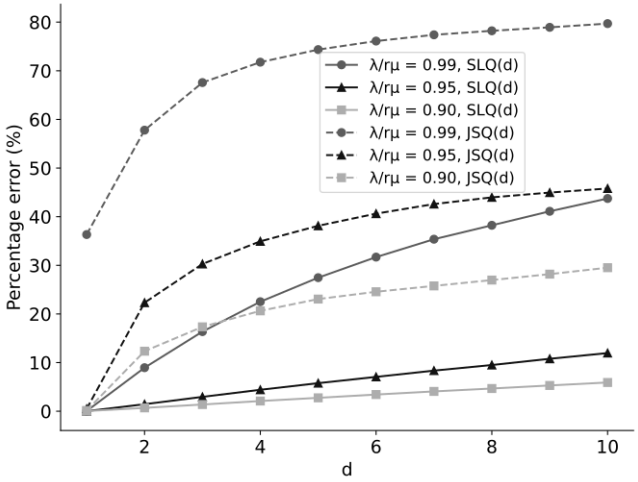

Fig. 6. Percentage error of the average response time of $\mathrm{JSQ}(d)$ and $\mathrm{SLQ}(d)$ as a function of the number of choices d.

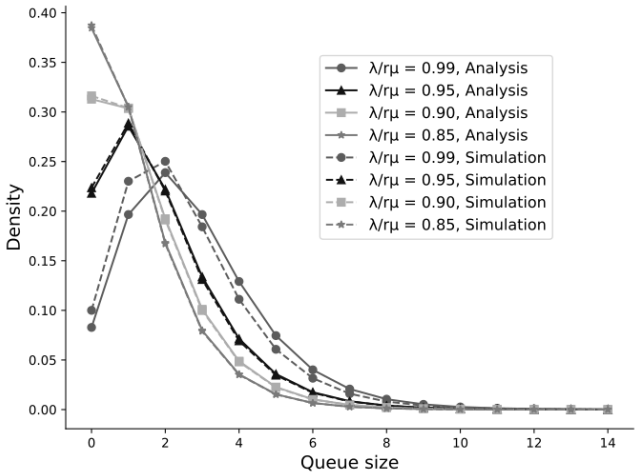

Fig. 7. Comparison of analytical and simulated queue state density functions for an EGS deploying the $\mathrm{SLQ}(2)$ policy.

in performance. The decrease in $\overline{w}$ for higher values of d is expected since the EGS is more likely to select a non-empty queue with large queue sizes and perform a successful sampling. In Figure 3, we observe that the average response time decreases as the back-off rate γ increases. A higher back-off rate allows an idle server to re-sample more frequently after a failed sampling, thus decreasing the server's overall idle time. In Figure 4, we compare the average response time of $\mathrm{SLQ}(d)$ as a function of server-to-queue ratio $r = \frac{m}{n}$ with a fixed number of flows (*i.e.*, $n = 200$). As servers process the jobs independently, increasing the number of servers m increases the overall service capacity of the EGS. Hence, the overall average response time decreases.

We now compare the performance of $\mathrm{SLQ}(d)$ with that of $\mathrm{JSQ}(d)$. From Figure 5 we observe that $\mathrm{JSQ}(d)$ has a lower average response time for a fixed $\frac{\lambda}{r\mu}$ ratio. The reason for this is that in the $\mathrm{SLQ}(d)$ model, BSAs may experience additional idle time due to failed (re-)sampling. In contrast, in the $\mathrm{JSQ}(d)$ model new jobs immediately join the least loaded amongst d randomly chosen queues, and there is never a need to re-sample. Figure 6 illustrates that $\mathrm{SLQ}(d)$ has a lower percentage error for the same $\frac{\lambda}{r\mu}$ ratio. Mean-field analysis necessitates $\mathrm{SLQ}(d)$ (resp. $\mathrm{JSQ}(d)$) to have a large n (resp. m); our system parameters do not meet the large m requirement. Figure 7 portrays queue state probability density functions, obtained via simulation as well as using the equilibrium state of the mean-field. While the mean-field approximation becomes less accurate for larger $\frac{\lambda}{r\mu}$ values (*e.g.*, at $\frac{\lambda}{r\mu} = 0.99$), overall we observe that our model exhibits close correspondence to the simulation.

V. Conclusion

Motivated by the importance of algorithm design and performance analysis of quantum switches with multiple entanglement swapping devices, we proposed the d-choices policy for an EGS and used mean-field techniques to study its performance. To this end, we developed a model for a homogeneous star network topology to study the role of this device in near-term quantum networks. For practical reasons, we placed queues at service-requesting nodes, instead of directly at the

EGS. In this way, our model contrasts from traditional d-choices policies applied within classical data-center models, necessitating new analysis. We then derived a set of differential equations describing the system's evolution in the mean-field limit, and proved the existence and uniqueness of the equilibrium state. Our numerical results show that the mean-field limit is useful to study an EGS even when it has a small number of resources (*e.g.*, 10) provided there is a large number of request queues. Particularly, for $\mathrm{SLQ}(d)$ with $d = 2$, mean-field approximations are tight except when $\frac{\lambda}{r\mu}$ is very close to one. Our analysis also applies to classical systems in which servers sample queues of flows to obtain a job for processing.

Our work serves as a preliminary study of an EGS deploying the d-choices policy. A valuable follow-up contribution would be to show global stability of the mean-field, and prove that its equilibrium state approximates the system's stationary distribution as the number of queues increases. The homogeneity assumption can be relaxed by introducing different classes of flows with identical arrival and service rates. Mean-field analysis requires each class to be adequately populated, thus the EGS service model may require alterations to ensure stability. Our study concentrated on a single, isolated EGS within a star topology. A valuable extension would be to explore multiple EGSes within more complex topologies.

REFERENCES

[1] A. Broadbent, J. Fitzsimons, and E. Kashefi, "Universal blind quantum computation," in *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, 2009.

[2] D. Leichtle, L. Music, E. Kashefi, and H. Ollivier, "Verifying BQP computations on noisy devices with minimal overhead," *PRX Quantum*, 2021.

[3] L. Jiang, J. M. Taylor, A. S. Sørensen, and M. D. Lukin, "Distributed quantum computation based on small quantum registers," *Phys. Rev. A*, 2007.

[4] V. Giovannetti, S. Lloyd, and L. Maccone, "Quantum-enhanced measurements: Beating the standard quantum limit," *Science*, 2004.

[5] D. Gottesman, T. Jennewein, and S. Croke, "Longer-baseline telescopes using quantum repeaters," *Phys. Rev. Lett.*, 2012.

[6] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theoretical Computer Science*, 2014.

[7] C. H. Bennett, G. Brassard, and N. D. Mermin, "Quantum cryptography without Bell's theorem," *Phys. Rev. Lett.*, 1992.

[8] F. Hahn, J. de Jong, and A. Pappa, "Anonymous quantum conference key agreement," *PRX Quantum*, 2020.

[9] C. Bennett, P. Hayden, D. Leung, P. Shor, and A. Winter, "Remote preparation of quantum states," *IEEE Transactions on Information Theory*, 2005.

[10] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels," *Phys. Rev. Lett.*, 1993.

[11] D. Gottesman and I. L. Chuang, "Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations," *Nature*, 1999.

[12] W. J. Munro, K. Azuma, K. Tamaki, and K. Nemoto, "Inside quantum repeaters," *IEEE Journal of Selected Topics in Quant. Electronics*, 2015.

[13] K. Azuma, S. E. Economou, D. Elkouss, P. Hilaire, L. Jiang, H.-K. Lo, and I. Tzitrin, "Quantum repeaters: From quantum networks to the quantum internet," *Reviews of Modern Physics*, 2023.

[14] E. Knill and R. Laflamme, "Theory of quantum error-correcting codes," *Phys. Rev. A*, 1997.

[15] M. Varnava, D. E. Browne, and T. Rudolph, "Loss tolerance in one-way quantum computation via counterfactual error correction," *Phys. Rev. Lett.*, 2006.

[16] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, "Quantum repeaters: the role of imperfect local operations in quantum communication," *Phys. Rev. Lett.*, 1998.

[17] W. Dür, H.-J. Briegel, J. I. Cirac, and P. Zoller, "Quantum repeaters based on entanglement purification," *Phys. Rev. A*, 1999.

[18] S. Muralidharan, C.-L. Zou, L. Li, and L. Jiang, "One-way quantum repeaters with quantum Reed-Solomon codes," *Phys. Rev. A*, 2018.

[19] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the stochastic analysis of a quantum entanglement distribution switch," *IEEE Transactions on Quantum Engineering*, 2021.

[20] S. Gauthier, G. Vardoyan, and S. Wehner, "A control architecture for entanglement generation switches in quantum networks," *IEEE Transactions on Quantum Engineering*, 2023.

[21] S. Gauthier, T. Vasantam, and G. Vardoyan, "An on-demand resource allocation algorithm for a quantum network hub and its performance analysis," in *IEEE QCE*, 2024.

[22] M. Mitzenmacher, A. W. Richa, and R. Sitaraman, "The power of two random choices: A survey of techniques and results," *Handbook of Randomized Computing*, 2001.

[23] L. Childress, M. Gurudev Dutt, J. Taylor, A. Zibrov, F. Jelezko, J. Wrachtrup, P. Hemmer, and M. Lukin, "Coherent dynamics of coupled electron and nuclear spin qubits in diamond," *Science*, 2006.

[24] M. Benaïm and J.-Y. L. Boudec, "A class of mean field interaction models for computer and communication systems," *2008 6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops*, 2008.

[25] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE Transac. on Parallel and Distributed Systems*, 2001.

[26] A. Einstein, B. Podolsky, and N. Rosen, "Can quantum-mechanical description of physical reality be considered complete?" *Physical review*, 1935.

[27] 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.

[28] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the exact analysis of an idealized quantum switch," *Performance Evaluation*, 2020.

[29] G. Vardoyan, P. Nain, S. Guha, and D. Towsley, "On the capacity region of bipartite and tripartite entanglement switching," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 2023.

[30] W. Dai, A. Rinaldi, and D. Towsley, "The Capacity Region of Entanglement Switching: Stability and Zero Latency," in *IEEE QCE*, 2022.

[31] P. Nain, G. Vardoyan, S. Guha, and D. Towsley, "On the analysis of a multipartite entanglement distribution switch," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2020.

[32] G. Avis, F. Rozpędek, and S. Wehner, "Analysis of multipartite entanglement distribution using a central quantum-network node," *Phys. Rev. A*, 2023.

[33] W. Pfaff, T. H. Taminiau, L. Robledo, H. Bernien, M. Markham, D. J. Twitchen, and R. Hanson, "Demonstration of entanglement-by-measurement of solid-state qubits," *Nature Physics*, 2013.

[34] M. Riebe, T. Monz, K. Kim, A. S. Villar, P. Schindler, M. Chwalla, M. Hennrich, and R. Blatt, "Deterministic entanglement swapping with an ion-trap quantum computer," *Nature Physics*, 2008.

[35] M. K. Bhaskar, R. Riedinger, B. Machielse, D. S. Levonian, C. T. Nguyen, E. N. Knall, H. Park, D. Englund, M. Lončar, D. D. Sukachev *et al.*, "Experimental demonstration of memory-enhanced quantum communication," *Nature*, 2020.

[36] 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," *Phys. Rev. Lett.*, 1996.

[37] 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," *Phys. Rev. Lett.*, 1996.

[38] S. Krastanov, V. V. Albert, and L. Jiang, "Optimized entanglement purification," *Quantum*, 2019.

[39] M. Alanyali and M. Dashouk, "Occupancy distributions of homogeneous queueing systems under opportunistic scheduling," *IEEE Transactions on Information Theory*, 2011.

[40] W. P. Grice, "Arbitrarily complete Bell-state measurement using only linear optical elements," *Phys. Rev. A*, 2011.

[41] S. R. Turner, "The Effect of Increasing Routing Choice on Resource Pooling," *Probability in the Engineering and Info. Sciences*, 1998.

[42] P. van Mieghem, *Performance Analysis of Complex Networks and Systems*. Cambridge University Press, 2014.