Mahdi Chehimi, Bernd Simon, Walid Saad, Anja Klein, Don Towsley and Mérouane Debbah "Matching Game for Optimized Association in Quantum Communication Networks", in *Proceedings of the IEEE Global Communications Conference - (IEEE GLOBECOM 2023)*, December 2023.

Matching Game for Optimized Association in Quantum Communication Networks

Mahdi Chehimi[1], Bernd Simon[3], Walid Saad[1,2], Anja Klein[3], Don Towsley[4], Mérouane Debbah[5]
[1]Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA USA
[2]Cyber Security Systems and Applied AI Research Center, Lebanese American University, Lebanon
[3]Communication Engineering Lab, Technische Universität Darmstadt, Darmstadt, Germany
[4]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA USA
[5]Electrical Engineering and Computer Science Department, Khalifa University of Science and Technology, Abu Dhabi, UAE
Emails: *{mahdic,walids}@vt.edu*, *{b.simon, a.klein}@nt.tu-darmstadt.de*, *towsley@cs.umass.edu*, *merouane.debbah@ku.ac.ae*

***Abstract*—Enabling quantum switches (QSs) to serve requests submitted by quantum end nodes in quantum communication networks (QCNs) is a challenging problem due to the heterogeneous fidelity requirements of the submitted requests and the limited resources of the QCN. Effectively determining which requests are served by a given QS is fundamental to foster developments in practical QCN applications, like quantum data centers. However, the state-of-the-art on QS operation has overlooked this association problem, and it mainly focused on QCNs with a single QS. In this paper, the request-QS association problem in QCNs is formulated as a matching game that captures the limited QCN resources, heterogeneous application-specific fidelity requirements, and scheduling of the different QS operations. To solve this game, a *swap-stable* request-QS association (RQSA) algorithm is proposed while considering partial QCN information availability. Extensive simulations are conducted to validate the effectiveness of the proposed RQSA algorithm. Simulation results show that the proposed RQSA algorithm achieves a near-optimal (within 5%) performance in terms of the percentage of served requests and overall achieved fidelity, while outperforming benchmark greedy solutions by over 13%. Moreover, the proposed RQSA algorithm is shown to be scalable and maintain its near-optimal performance even when the size of the QCN increases.**

I. Introduction

Quantum communication networks (QCNs) are seen as a pillar of future communication technologies due to their advantages in terms of security, sensing capabilities, and computational powers. QCNs rely on the creation and distribution of Einstein-Podolsky-Rosen (EPR) pairs of *entangled quantum states* between distant QCN nodes [1]. Each EPR pair consists of two inherently-correlated photons, each of which is transferred to a QCN node to establish an end-to-end (e2e) entangled connection. However, the fragile nature of entangled photons results in exponential losses that increase with the travelled distance over quantum channels, e.g., optical fiber. As such, intermediate quantum repeater nodes are needed to split long distances into shorter segments by performing *entanglement swapping* on entangled photons to connect distant QCN nodes [2]. When such repeaters share multiple EPR pairs with several QCN nodes to create e2e connections, they are called *quantum switches (QSs)*.

This work was supported in part by Zaiku Group, the National Science Foundation (NSF) under Grant CNS-2201641, the German Research Foundation (DFG) within the Collaborative Research Center (CRC) 1053 MAKI, the BMBF project Open6GHub (Nr. 16KISK014), NSF grant CNS-1955834, NSF-ERC Center for Quantum Networks grant EEC-1941583, and the MURI ARO Grant W911NF2110325

In practice, a QS has a limited-capacity quantum memory for photon storage. A heralding station is responsible for generating EPR pairs and distributing each pair between the QS and other QCN nodes to create link-level connections (LLCs). The *fidelity*, or quality, of an LLC can be enhanced by performing *entanglement distillation* before swapping two LLCs to establish an e2e connection [3]. Practical applications, like quantum data centers and quantum cloud networks, encompass QCN setups with multiple QSs connecting several end-node quantum devices. The design of such multiple-QS QCNs requires overcoming many challenges such as the limited storage capacity of QSs, imperfections associated with EPR generation and transmission, the need to schedule the different QS operations (i.e., entanglement swapping and distillation), and the presence of heterogeneous application-specific minimum fidelity requirements.

Multiple prior works [4]–[10] attempted to address some of the aforementioned challenges, and those works can be divided into three main types. First, some works, like [4], considered a QS-based multi-hop QCN and performed entanglement provisioning and *path selection* to maximize throughput. Second, prior works, such as [5], considered *routing* EPR pairs over several QCN paths to create e2e connections. The last type, which is the most relevant to our work, considered star-shaped QCNs, where several nodes are connected to a single QS through EPR pairs [6]–[10]. For instance, the work in [6] was the first to consider aggregate QS capacity and analytically analyze its stability. However, almost-perfect conditions were assumed in [6]. Additionally, the work in [7] considered a QCN with a QS serving requests having minimum fidelity constraints. However, entanglement distillation was not considered in [7]. Meanwhile, the work in [8] studied QS stability and swap scheduling. However, the authors in [8] did not include entanglement distillation and assumed an infinite lifetime of EPR pairs. Moreover, the work in [9] analyzed the capacity regions and stability of a single QS and scheduled swapping/distillation operations to satisfy minimum fidelity requirements while considering noisy gates and measurements. However, [9] considered a homogeneous fidelity for all link-level EPR pairs. Finally, the authors in [10] proposed a memory allocation policy for a constrained QS operation in a star-shaped QCN. However, the model proposed in [10] did

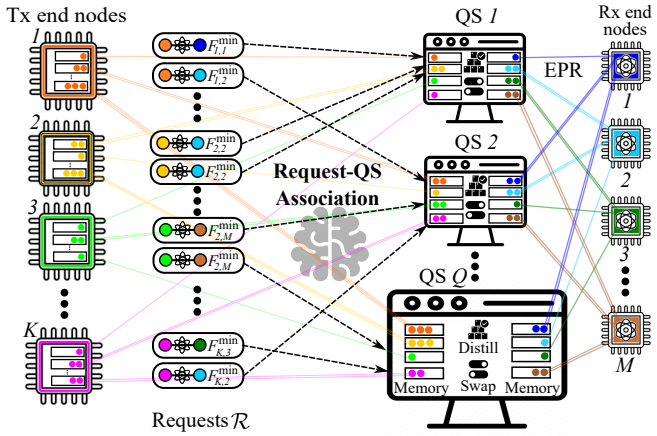

Fig. 1. Studied QCN model for the requests-QSs association problem.

not account for fidelity requirements of both link-level and e2e connections and did not schedule distillation operations.

Furthermore, these prior works [6]–[10] focused on a single QS handling all e2e requests and did not consider multiple QSs connected to several end nodes with heterogeneous resources and fidelity constraints. In such a QCN setup (see Fig. 1), it is essential to *associate* each request with the QS that optimizes its fidelity. This *request-QS association problem*, which is essential for designing quantum data centers, has been overlooked in prior works. Accordingly, there is a need for a thorough investigation of the request-QS association problem in QCNs with multiple QSs while taking into consideration the scheduling of QS entanglement swapping/distillation operations, memory limitations, and performance requirements.

The main contribution of this work is a novel matching-based framework for optimizing request-QS association in QCNs with multiple QSs, possessing heterogeneous resources, that satisfy QCN users' performance requirements while considering practical constraints of QCN elements. To the best of our knowledge, this is the first work to explore this research area, and therefore, we make the following key contributions:

- We propose the first holistic analysis of the request-QS association problem in QCNs under limited resource constraints and heterogeneous fidelity requirements.
- We formulate the request-QS association problem as a *matching game* [11] where both requests and QSs rank each other based on fidelity-maximizing preferences. This novel matching approach enables us to solve the considered association problem without requiring full knowledge of QCN information.
- We propose a novel *request-QS association (RQSA)* algorithm based on *swap-matching* [12] to solve the formulated matching game while guaranteeing convergence under partial QCN information availability.
- Simulation results show that our RQSA algorithm is scalable and achieves a near-optimal performance within 5% of the optimal solution in terms of served requests and overall served e2e fidelity.

II. System Model

Consider a QCN composed of a set $\mathcal{Q}$ of Q QSs connected to a set of end nodes through link-level EPR pairs. The end nodes are split into transmitting (Tx) and receiving (Rx) nodes, where requests for e2e connections are sent from Tx nodes to the QSs (see Fig. 1). Moreover, $\mathcal{K}$ denotes the set of K Tx nodes, and $\mathcal{M}$ the set of M Rx nodes.

The operation of the QCN occurs in a time-slotted manner. Prior to each time slot, heralding stations installed between QSs and end nodes attempt to create n link-level EPR pairs to connect every QS $q \in \mathcal{Q}$ to every Tx (and Rx) node, $k \in \mathcal{K}$ (and $m \in \mathcal{M}$), respectively, with a probability of success $p_{k,q}$ (and $p_{q,m}$) for each pair that depends on the corresponding link length. Accordingly, the link-level EPR generation process between a QS and a Tx (or Rx) node follows a *binomial distribution* with parameters n and $p_{k,q}$ (or $p_{q,m}$) [9]. Thus, each QS is connected to all Tx nodes through $N^{\text{Tx}}_{k,q}$ successfully-generated link-level EPR pairs each having a *fidelity* of $F^{\text{Tx}}_{k,q}$. Similarly, every QS is linked to each Rx node through $N^{\text{Rx}}_{q,m}$ successfully-generated link-level EPR pairs each of fidelity $F^{\text{Rx}}_{q,m}$. The EPR pairs are then stored in quantum memories at both the QSs and end nodes. Those pairs are assumed to remain coherent and maintain their fidelities for one time slot, before being discarded.

At the beginning of each time slot, Tx nodes submit a set $\mathcal{R}$ of R requests to the QSs. Each request is represented as a tuple, $r^{i,j}_l = (i, j, F^{\min}_{i,j})$, where $l \in \{1, 2, ..., R\}$, $i \in \mathcal{K}$, and $j \in \mathcal{M}$. Here, $r^{k,m}_l$ represents a request by Tx node $k \in \mathcal{K}$ to establish a single e2e EPR pair with Rx node $m \in \mathcal{M}$ with a minimum fidelity of $F^{\min}_{k,m}$. In addition, each submitted request must be served, if feasible, during its submission time slot, or be discarded. We assume that, during each time slot, every Tx node may submit multiple repeated requests that are identical and have exactly the same required minimum fidelity, since they intend to serve the same application.

In our model, we consider that only partial QCN information is available to the Tx nodes when submitting their requests. In particular, each Tx node has access to only the information related to its link-level EPR pairs with every QS. Moreover, the QSs publicly announce information about their link-level EPR pairs with every Rx node to the Tx nodes.

Each QS $q \in \mathcal{Q}$ can perform two distinct quantum operations: 1) *entanglement swapping*, to connect a Tx node to an Rx node, and 2) *entanglement distillation* to enhance the fidelity of link-level EPR pairs. Every link-level EPR pair is represented by a Werner state $\rho = W\left|\psi_{00}\right\rangle\left\langle\psi_{00}\right| + \frac{1-W}{4}\Pi$, where W is the Werner parameter that directly affects the fidelity of those pairs, which is given as: $F = \frac{3W+1}{4}$ [13].

When a QS $q \in \mathcal{Q}$ swaps two link-level EPR pairs, one with Tx node $k \in \mathcal{K}$ having fidelity $F^{\text{Tx}}_{k,q}$, and the other with Rx node $m \in \mathcal{M}$ having fidelity $F^{\text{Rx}}_{q,m}$, the resulting e2e EPR pair has a fidelity given by [2]:

$$S(F^{\text{Tx}}_{k,q}, F^{\text{Rx}}_{q,m}) = \frac{1}{4} + \frac{3}{4}\left(\frac{4F^{\text{Tx}}_{k,q} - 1}{3}\right)\left(\frac{4F^{\text{Rx}}_{q,m} - 1}{3}\right). \quad (1)$$

We adopt the Oxford entanglement distillation protocol [3] for performing entanglement distillation of two link-level EPR pairs. According to this protocol, two identical EPR pairs with initial fidelity F_{initial} can be distilled into one EPR pair having

a higher fidelity given by [3]:

$$D(F_{\text{initial}}) = \frac{(F_{\text{initial}})^2 + (\frac{1-F_{\text{initial}}}{3})^2}{(F_{\text{initial}})^2 + 2F_{\text{initial}}(\frac{1-F_{\text{initial}}}{3}) + 5(\frac{1-F_{\text{initial}}}{3})^2}. \quad (2)$$

To simplify the analysis, a QS is assumed to perform at most one distillation operation for each link-level EPR pair. Also, if performed, distillation is considered to always precede entanglement swapping [9]. Accordingly, there are four possible actions regarding the scheduling of the entanglement swapping/distillation operations to handle each submitted request that every QS can take.[1] The action choice directly affects the fidelities of the resulting e2e EPR pairs and the number of available link-level EPR pairs in quantum memories. Here, we introduce α_j^{Tx} and α_j^{Rx} to denote the number of *utilized link-level EPR pairs* from both Tx and Rx nodes' memories, respectively, as a result of each possible QS action $j \in \{1, 2, 3, 4\}$. The four considered actions and their corresponding impacts are:

1) Direct entanglement swapping: Swap one link-level EPR pair connected to Tx node $k \in \mathcal{K}$ with one link-level EPR pair connected to Rx node $m \in \mathcal{M}$. When QS $q \in \mathcal{Q}$ performs this action to serve request $r_l^{k,m}$, the fidelity of the resulting e2e EPR pair will be $F^{\text{e2e}}_{q,k,m,1} = S(F^{\text{Tx}}_{k,q}, F^{\text{Rx}}_{q,m})$. Consequently, the number of link-level EPR pairs between QS $q \in \mathcal{Q}$ and Tx node $k \in \mathcal{K}$ and Rx node $m \in \mathcal{M}$, i.e., $N^{\text{Tx}}_{k,q}$ and $N^{\text{Rx}}_{q,m}$, respectively, are both reduced by 1. The number of utilized link-level EPR pairs associated with the *direct entanglement swapping* action are given by $\alpha_1^{\text{Tx}} = \alpha_1^{\text{Rx}} = 1$.

2) Tx distillation, then entanglement swapping: Distill two link-level EPR pairs connected to the Tx node $k \in \mathcal{K}$, then swap the distilled pair with an EPR pair connected to the Rx node $m \in \mathcal{M}$. When QS $q \in \mathcal{Q}$ performs this action to serve a request $r_l^{k,m}$, the fidelity of the resulting e2e EPR pair is $F^{\text{e2e}}_{q,k,m,2} = S(D(F^{\text{Tx}}_{k,q}), F^{\text{Rx}}_{q,m})$. Consequently, the number of link-level EPR pairs between QS $q \in \mathcal{Q}$ and Tx node $k \in \mathcal{K}$ is reduced by 2, while the number of link-level EPR pairs between QS $q \in \mathcal{Q}$ and Rx node $m \in \mathcal{M}$ is reduced by 1, as the entanglement distillation utilizes two link-level EPR pairs. The number of utilized link-level EPR pairs associated with the *Tx distillation, then entanglement swapping* action are $\alpha_2^{\text{Tx}} = 2$, and $\alpha_2^{\text{Rx}} = 1$.

3) Rx distillation, then entanglement swapping: Distill two link-level EPR pairs connected to Rx node $m \in \mathcal{M}$, then swap the distilled pair with an EPR pair connected to Tx node $k \in \mathcal{K}$. When QS $q \in \mathcal{Q}$ performs this action to serve request $r_l^{k,m}$, the fidelity of the resulting e2e EPR pair is $F^{\text{e2e}}_{q,k,m,3} = S(F^{\text{Tx}}_{k,q}, D(F^{\text{Rx}}_{q,m}))$. Consequently, the number of link-level EPR pairs between QS q and Tx node k is reduced by 1, while the number of link-level EPR pairs between QS $q \in \mathcal{Q}$ and Rx node m is reduced by 2. The number of utilized link-level EPR pairs associated with the *Rx distillation, then entanglement swapping* action are $\alpha_3^{\text{Tx}} = 1$, and $\alpha_3^{\text{Rx}} = 2$.

4) Tx & Rx distillation, then entanglement swapping: Distill two link-level EPR pairs connected to Tx node $k \in \mathcal{K}$, and simultaneously distill two EPR pairs connected to Rx node $m \in \mathcal{M}$, then swap the two distilled pairs. When QS $q \in \mathcal{Q}$ performs this action to serve request $r_l^{k,m}$, the fidelity of the resulting e2e EPR pair is $F^{\text{e2e}}_{q,k,m,4} = S(D(F^{\text{Tx}}_{k,q}), D(F^{\text{Rx}}_{q,m}))$. Consequently, the number of link-level EPR pairs between QS q and Tx node k and Rx node m are both reduced by 2. The numbers of utilized link-level EPR pairs associated with the *Tx & Rx distillation, then entanglement swapping* action are $\alpha_4^{\text{Tx}} = \alpha_4^{\text{Rx}} = 2$.

To simplify notation, we introduce the vectors $\boldsymbol{\alpha}^{\text{Tx}} = [1, 2, 1, 2]^{\text{T}}$ and $\boldsymbol{\alpha}^{\text{Rx}} = [1, 1, 2, 2]^{\text{T}}$ of *utilized link-level EPR pairs* that result from the four possible QS actions. Next, we formulate the request-QS association problem and propose a matching game formulation [14].

III. Request-QS Association as a Matching Game

A. Request-QS Association Problem

In the request-QS association problem, a submitted request $r_l^{k,m} \in \mathcal{R}$ must be associated, if feasible, with at most one QS $q \in \mathcal{Q}$, or be discarded. This QS performs one of the four aforementioned actions to serve the request during a time slot. We define *matching* η as an association between QSs and requests. The association between a submitted request $r_l^{k,m}$ and a QS q is denoted as $(r_l^{k,m}, q) \in \eta$. Each QS $q \in \mathcal{Q}$ can serve multiple requests. We define $\mathcal{R}_q^\eta \subseteq \mathcal{R}$ as the set of requests associated with QS q in matching η. As multiple requests are associated with each QS, we have a *many-to-one* matching problem.

Each submitted request must be served with the highest fidelity possible. Therefore, we define the utility of a submitted request $r_l^{k,m} \in \mathcal{R}$ when associated with QS $q \in \mathcal{Q}$ as the *fidelity* of its generated e2e EPR pair:

$$U_l(q) = F^{\text{e2e}}_{q,k,m,1} = S(F^{\text{Tx}}_{k,q}, F^{\text{Rx}}_{q,m}), \quad (3)$$

where $k \in \mathcal{K}$ and $m \in \mathcal{M}$ are the corresponding Tx/Rx nodes, respectively, in $r_l^{k,m} = (k, m, F^{\text{min}}_{k,m})$. (3) considers the worst case for QS q, which corresponds to taking the *direct entanglement swap* action without any distillation, since that action yields the lowest fidelity of the resulting e2e EPR pair. This worst-case assumption stems from the fact that the request (i.e., the end node) does not know which action will be taken by its prospective QS q.

Similarly, each QS aims to serve each request with the highest fidelity possible. In matching η, for each individual request $r_l^{k,m} \in \mathcal{R}$ served by QS $q \in \mathcal{Q}$, the respective QS utility for that request is the resulting e2e EPR pair's fidelity:

$$\tilde{U}_q(r_l^{k,m}) = \begin{cases} F^{\text{e2e}}_{q,k,m,j_{q,k,m}(\eta)}, & \text{if } F^{\text{e2e}}_{q,k,m,j_{q,k,m}(\eta)} \geq F^{\text{min}}_{k,m} \\ -\infty, & \text{else.} \end{cases} \quad (4)$$

In (4), $j_{q,k,m}(\eta)$ captures the fact that the fidelity of the resulting e2e EPR pair depends on the action taken by the QS. For instance, $j_{q,k,m}(\eta) \in \{1, 2, 3, 4\}$ represents the action taken by the QS to serve request $r_l^{k,m}$ based on matching η. The second case in the above expression corresponds to the situation when the QS cannot serve the request because it cannot provide the request's minimum fidelity requirement.

Each QS $q \in \mathcal{Q}$ must decide on the actions that maximize the fidelity for its associated requests, i.e., maximize (4) for

[1] A higher number of possible actions can be easily integrated into our model by allowing QSs to perform more distillation operations.

each request. After the optimal actions are identified, the overall utility of QS $q \in \mathcal{Q}$ for its associated set of requests $\mathcal{R}_q^\eta$ in matching η is the sum of the individual request utilities:

$$U_q(\mathcal{R}_q^\eta) = \sum_{r_l^{k,m} \in \mathcal{R}_q^\eta} \tilde{U}_q(r_l^{k,m}), \tag{5}$$

which captures the fact that the goal of each QS $q \in \mathcal{Q}$ is to maximize the overall delivered e2e fidelities for the set of associated requests $\mathcal{R}_q^\eta$.

The process of selecting the actions to serve the associated requests in $\mathcal{R}_q^\eta$ by QS $q \in \mathcal{Q}$ can be formulated as an optimization problem. To do so, we define $\boldsymbol{A}_q$ as the actions matrix for QS q, which includes all possible actions for all its associated requests $r_l^{k,m} \in \mathcal{R}_q^\eta$. In particular, $\boldsymbol{A}_q = [\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4]$, where each vector $\boldsymbol{a}_j$ is of dimension $\left|\mathcal{R}_q^\eta\right| \times 1$, and each entry $a_{l,j}$ of $\boldsymbol{a}_j$, given $l \in \{1, 2, ..., \left|\mathcal{R}_q^\eta\right|\}$ and $j \in \{1, 2, 3, 4\}$, corresponds to a request $r_l^{k,m} \in \mathcal{R}_q^\eta$. Each element $a_{l,j}$ is binary, where it takes a value of one when action j is performed to serve request $r_l^{k,m}$. The dimension of $\boldsymbol{A}_q$ is $\left|\mathcal{R}_q^\eta\right| \times 4$. Accordingly, the action-selection optimization problem for QS $q \in \mathcal{Q}$ is:

$$\mathcal{P}1: \max_{\boldsymbol{A}_q} \sum_{r_l^{k,m} \in \mathcal{R}_q^\eta} U_q(r_l^{k,m}) \tag{6a}$$

$$s.t. \sum_{i: r_i = r_i^{k,m}, \forall m \in \mathcal{M}_q} \boldsymbol{A}_q \cdot \boldsymbol{\alpha}^{\mathrm{Tx}} \leq N_{k,q}^{\mathrm{Tx}}, \quad \forall k \in \mathcal{K}_q, \tag{6b}$$

$$\sum_{i: r_i = r_i^{k,m}, \forall k \in \mathcal{K}_q} \boldsymbol{A}_q \cdot \boldsymbol{\alpha}^{\mathrm{Rx}} \leq N_{q,m}^{\mathrm{Rx}}, \quad \forall m \in \mathcal{M}_q, \tag{6c}$$

where the objective function corresponds to the overall utility achieved by QS $q \in \mathcal{Q}$ from all its associated requests $\mathcal{R}_q^\eta$. Constraint (6b) ensures that the number of used link-level EPR pairs between the QS and Tx node $k \in \mathcal{K}_q$ does not exceed the number of available link-level EPR pairs between them, $N_{k,q}^{\mathrm{Tx}}, \forall k, q \in \mathcal{K}_q, \mathcal{Q}$. Similarly, constraint (6c) ensures that the number of consumed link-level EPR pairs between the the QS and Rx node $m \in \mathcal{M}_q$ does not exceed the number of available link-level EPR pairs between them, $N_{q,m}^{\mathrm{Rx}}, \forall q, m \in \mathcal{Q}, \mathcal{M}_q$.

Solving the request-QS association problem is challenging, because it must factor in the limited number of available link-level EPR pairs of the Tx and Rx nodes and the QSs. Also, each QS must schedule its actions such that the maximum number of submitted requests in the QCN is served during each time step. Solving the request-QS association problem using classical optimization techniques is impractical because the number of possible combinations of associated requests per QS is 2^R, i.e., the complexity grows exponentially with R. Accordingly, we propose a computationally efficient, decentralized approach that accounts for the partial QCN information availability.

B. Matching Game Formulation

Matching theory [14] is a powerful tool that has been adopted to solve several complex communication network problems [15]. Here, we leverage matching theory to formulate the request-QS association problem as a matching game so as to overcome its exponentially growing complexity. Note that our formulation differs from prior works on matching games for classical wireless systems [16] in the fact that we have to consider quantum-specific constraints regarding the fidelity of EPR pairs, limited quantum memory, and heterogeneous minimum fidelity requirements. Formally, the proposed matching game is defined as follows.

Definition 1 (Matching game)**.** *A matching game is defined by two sets of matching parties $(\mathcal{R}, \mathcal{Q})$ and two preference relations $\succ_r^{\mathrm{Req}}$, $\succ_q^{\mathrm{QS}}$ allowing each submitted request $r_l^{k,m} \in \mathcal{R}$ to rank the QSs and each QS $q \in \mathcal{Q}$ to rank sets of associated requests.*

For any request $r_l^{k,m}$, a *preference relation* $\succ_r^{\mathrm{Req}}$ is defined over the set of QSs $\mathcal{Q}$ such that, for any two QSs, $q, q' \in \mathcal{Q}$, we have:

$$q \succ_r^{\mathrm{Req}} q' \Leftrightarrow U_l(q) > U_l(q'), \tag{7}$$

which means that request $r_l^{k,m}$ prefers QS $q \in \mathcal{Q}$ over QS $q' \in \mathcal{Q}$ whenever the utility (3) associated with $q \in \mathcal{Q}$ is higher than the utility associated with q'.

Similar to the case of requests, for any QS $q \in \mathcal{Q}$, we define a *preference relation* $\succ_q^{\mathrm{QS}}$ over the set of associated requests $\mathcal{R}_q^\eta$. For any two matchings η, η', the QS ranks the corresponding sets of associated requests $\mathcal{R}_q^\eta$ in matching η and $\mathcal{R}_q^{\eta'}$ in matching η' as follows:

$$\mathcal{R}_q^\eta \succ_q^{\mathrm{QS}} \mathcal{R}_q^{\eta'} \Leftrightarrow U_q(\mathcal{R}_q^\eta) > U_q(\mathcal{R}_q^{\eta'}), \tag{8}$$

which means that the QS $q \in \mathcal{Q}$ prefers the set $\mathcal{R}_q^\eta$ of associated requests in matching η over the set $\mathcal{R}_q^{\eta'}$ in matching η' whenever the overall utility (5) associated with $\mathcal{R}_q^\eta$ is higher than the overall utility associated with $\mathcal{R}_q^{\eta'}$.

C. Proposed Solution and Algorithm

In this section we propose an algorithm to find a *stable* matching η. Classical definitions of stability [11], [14] in matching games, which rely on preferences of individual matching parties, cannot be applied to our proposed matching formulation. This is because the preference relations (8) require QSs to rank sets of associated requests instead of individual requests. To overcome this challenge, we adopt the definition of *swap stability* [12], which means that no submitted request or QS can increase its utility by swapping its current matching partner.[2] The foundation for the analysis of swap stability is a *swap matching*, which simply results from two requests r and r' exchanging their respective associated QSs q and q' in η. Formally, given a matching η, two submitted requests $r, r' \in \mathcal{R}$ and two QSs $q, q' \in \mathcal{Q}$ with $(r, q), (r', q') \in \eta$, a swap matching is defined as $\eta_{r,r'}^q = \eta \setminus \{(r, q), (r', q')\} \cup \{(r', q), (r, q')\}$. Accordingly, swap stability is defined as:

Definition 2 (Swap stability)**.** *A matching η is said to be swap stable if no swap matching $\eta' = \eta_{r,r'}^q$ exists such that:*
(i) Request r and r' prefer the QSs associated in swap matching η' over the QSs associated in η. Formally, both prefer to swap their respective QSs $q' \succ_r^{\mathrm{Req}} q$ and $q \succ_{r'}^{\mathrm{Req}} q'$, and

[2]Note that swap stability is not to be confused with the entanglement swap operation that a QS can perform on two link-level EPR pairs.

Algorithm 1 Request-QS Association (RQSA)

Require: Set of requests $\mathcal{R}$, set of QSs $\mathcal{Q}$.
Phase 1: Initialization Phase
1: Each request $r_l^{k,m}$ determines its worst-case fidelities from (7)
2: Match each request $r_l^{k,m}$ to the QS q with the highest worst-case fidelity as long as constraints (6b) and (6c) are satisfied.
Phase 2: Swap Matching Phase
3: **repeat**
4: **for all** $r \in \mathcal{R}$ **do**
5: Select a QS q' that yields a higher utility then the currently matched QS q
6: **for all** Requests r' matched to QS q', i.e., $r' \in \mathcal{R}_{q'}^{\eta}$ **do**
7: QS q' identifies a request r' that shares the same Tx or Rx node with request r which is not matched to q'
8: **if** $q \succ_{r'}^{\text{Req}} q'$ **and** $q' \succ_{r}^{\text{Req}} q$ **then**
9: Construct the swap matching $\eta' \leftarrow \eta_{r,r'}^{q}$
10: **if** $\mathcal{R}_q^{\eta'} \succ_q^{\text{QS}} \mathcal{R}_q^{\eta}$**and** $\mathcal{R}_{q'}^{\eta'} \succ_{q'}^{\text{QS}} \mathcal{R}_{q'}^{\eta}$ **and** (6b),(6c) are satisfied **then**
11: The swap of r' and r is approved, i.e., $\eta \leftarrow \eta'$
12: **else**
13: The swap of r' and r is denied.
14: **end if**
15: **end if**
16: **end for**
17: **end for**
18: **until** no more pairs of requests r, r' to swap are found
19: All QSs $q \in \mathcal{Q}$ solve $\mathcal{P}1$ from (6a) to determine the action for each request
Stage 3: e2e EPR Pair Generation Phase
20: Each QS performs its respective actions to create e2e EPR pairs according to η

(ii) QS $q \in \mathcal{Q}$ and q' prefer the requests associated in swap matching η' over the requests associated in η. Formally, $\mathcal{R}_q^{\eta'} \succ_q^{\text{QS}} \mathcal{R}_q^{\eta}$ and $\mathcal{R}_{q'}^{\eta'} \succ_{q'}^{\text{QS}} \mathcal{R}_{q'}^{\eta}$.

To solve the proposed matching game, i.e. finding a swap stable matching η, a key challenge is that the preferences (8) of the QSs do not rank individual requests, but rather sets of requests. Accordingly, a QS $q \in \mathcal{Q}$ cannot decide whether to accept or defer an individual request. Instead, the QS has to consider all its other associated requests in $\mathcal{R}_q^{\eta}$. Furthermore, instead of having a fixed quota at each QS, we must consider the limited quantum memory of each QS given by constraints (6b) and (6c). Therefore, the well-known deferred acceptance algorithm [14] cannot be applied to this game.

To overcome these challenges, we propose a novel request-QS association (RQSA) swap matching algorithm, which is shown in Algorithm 1. The matching is initialized by a greedy strategy, i.e., all submitted requests are matched to QSs with the highest request utility (3) as long as constraints (6b) and (6c) are satisfied for each QS (lines 1 and 2). After initialization, the swap matching phase begins. The preference list of each request r is calculated and a more preferred QS q' than its currently matched QS q is identified (line 5). QS q' identifies a request r' from its associated requests that shares the same Tx or Rx node with another request r not associated to q' (line 7). Then swap matching $\eta_{r,r'}^{q}$ is considered, wherein r will be served by q' instead of $q \in \mathcal{Q}$ and r' will be served by $q \in \mathcal{Q}$ instead of q' (line 8). The swap is performed when the requests r and r' and QSs q and q' prefer the swap, with at least one participant strictly preferring the swap matching over its current matching (line 7-16). This procedure is repeated until no more swaps can be found in the network (line 17). In the last stage, the e2e EPR pair generation phase, the QSs solve optimization problem $\mathcal{P}1$, identify and perform the actions to serve their associated requests (lines 18 and 19). The stability of the resulting matching η follows from:

Lemma 1. *Upon convergence, RQSA reaches a swap stable matching according to Definition 2.*

Proof: To prove swap stability upon convergence, we have to show that no pair of submitted requests r and r' exists with their associated QSs q and q', such that a swap of r and r' is preferred by the submitted requests and QSs. RQSA checks for all combinations of r, r', q and q', whether a swap is preferred by all requests and QSs. If such a combination of r, r', q and q' is found, the swap is performed. This procedure is repeated until no more swaps are performed. Therefore, after the swap matching phase, the resulting matching η is swap stable as no more pairs of submitted requests remain that would prefer to be served by another QS.[3]

IV. Simulation Results and Analysis

For our simulations, we define the following *default* setup QCN parameters: 1) The number of Tx nodes is $K = 5$, the number of Rx nodes is $M = 5$, while the number of QSs is $Q = 3$; 2) Heralding stations perform $n = 10$ link-level EPR pair generation attempts, and the numbers of successfully-generated pairs in every time slot are binomial random variables $N_{k,q}^{\text{Tx}} \sim B(n = 10, p = p_{k,q})$ and $N_{q,m}^{\text{Rx}} \sim B(n = 10, p = p_{q,m})$ (see Sec. II). The probability of success is $p_{k,q} = e^{-d_{k,q}/L_0}$, for links between a QS $q \in \mathcal{Q}$ and a Tx node $k \in \mathcal{K}$, where $L_0 = 0.54\,\text{km}$ is the optical fiber's attenuation coefficient [17], and $d_{k,q}$ is the length of those links. Similarly, $p_{q,m} = e^{-d_{q,m}/L_0}$ for links between a QS q and an Rx node $m \in \mathcal{M}$. The lengths $d_{k,q}$ and $d_{q,m}$ are sampled from a uniform distribution between $100\,\text{m}$ and $1\,\text{km}$, $\mathcal{U}(0.1, 1)$; 3) Each request has a different minimum required fidelity $F_{k,m}^{\min}$ based on its intended quantum application.[4] Thus, we randomly sample such values from a uniform distribution $\mathcal{U}(0.5, 0.8)$; 4) Initial fidelities $F_{k,q}^{\text{Tx}}$ and $F_{q,m}^{\text{Rx}}$ of link-level EPR pairs depend on the hardware, so they are sampled from a uniform distribution $\mathcal{U}(0.83, 0.99)$ [17]; 5) The number of submitted requests lies in the range $R \in [0, 40]$. We perform 100 independent simulation runs wherein all aforementioned random variables are drawn from their respective distributions. Each run analyzes a single time slot where Tx nodes submit a set of requests, that we solve the request-QS association problem for. Unless stated otherwise, these default parameters are used in all simulation experiments.

We benchmark the proposed RQSA algorithm against the following baselines: 1) *Optimal*, which formulates the request-QS association problem as an integer optimization problem solved using an advanced solver [18]. This requires complete QCN information that is impractical due to classical communication delay, 2) *Greedy algorithm*, which selects the QS with the highest worst-case fidelity to serve a request, and when it lacks enough link-level EPR pairs, the next-best QS is chosen, and 3) *Random algorithm*, which randomly associates each request with a QS. We discuss the experimental results next.

[3]Regarding the proof of convergence, we refer the reader to a general proof for swap matching algorithms in [12] due to space limitations.

[4]Particularly, distillation protocols require a minimum fidelity of 0.5, while 0.8 is a typical value for quantum key distribution applications [9].

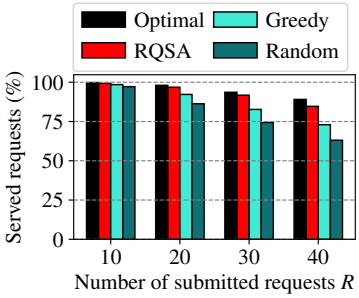

Fig. 2. Average percentage of served requests as a function of R.

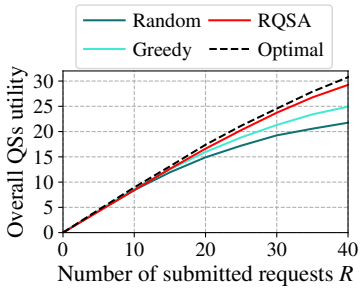

Fig. 3. Overall QSs' utility as a function of R.

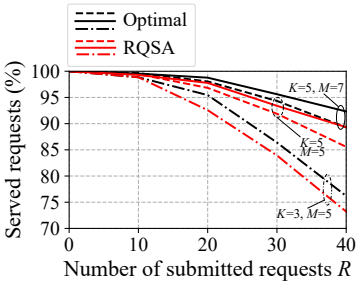

Fig. 4. Average percentage of served requests as a function of K,M and R.

1) Impact of Number of Requests on Served Requests: First, we analyze, the performance as the number R of submitted requests varies, and compare the percentage of served requests by the different algorithms. Corresponding results are shown in Fig. 2, where we observe that RQSA is within 5% of the optimal solution in terms of served requests. Moreover, RQSA achieves superior performance compared to both greedy and random algorithms, and the performance gap between the algorithms increases as the number of requests increases. For instance, when the number of submitted requests is 40, RQSA serves around 13% and 22% more requests than the greedy and random algorithms, respectively. From Fig. 2, we also observe that the optimal percentage of served requests decreases as the number of requests increases. This is due to limited QCN resources, namely link-level Tx/Rx EPR pairs, which leads to many submitted requests becoming infeasible as more requests are submitted.

2) Impact of Number of Requests on Overall QS Utility: Next, we show the effect of R on the overall achieved QSs' utility, i.e., *sum of served e2e fidelities*, in Fig. 3. We observe from Fig. 3 that RQSA achieves near-optimal performance, even for large R, e.g., $R = 40$. In such cases, RQSA achieves a performance within 5% of the optimal overall utility, unlike the greedy and random algorithms that start to diverge from the optimal solution as R becomes large. Note that, in contrast to the optimal solution algorithm, RQSA requires significantly smaller run time, and does not require full QCN information availability while being scalable.

3) Impact of QCN Size on Performance: Finally, in Fig. 4, we analyze the scalability of RQSA by showing the percentage of served requests as R varies while considering three different QCN sizes. In particular, we consider the cases in which $K > M$, $K = M$, and $K < M$ for a fixed number of QSs $Q = 3$. From Fig. 4, for small QCNs, e.g., $K = 3$, we observe that a small number of Tx nodes imposes a bottleneck on the maximum number of served requests, since the number of available link-level EPR pairs becomes insufficient to satisfy the increased number of requests. Additionally, we observe from Fig. 4 that RQSA is scalable across different (small and large) QCN sizes, and it achieves a near-optimal performance, that is within 4% of the optimal solution.

V. Conclusion

In this paper, we have studied the problem of requests-QSs association in QCNs with multiple QSs, which is crucial for QCN applications like quantum data centers. To develop a practical solution and overcome the challenges of partial information and the combinatorial complexity of the association problem, we have formulated the problem as a matching game. The proposed formulation takes into account practical QCN considerations such as limited memory capacity, heterogeneous fidelity requirements, and scheduling of QS operations. Moreover, we have developed a novel swap-matching based RQSA algorithm to solve the matching game while achieving stability. Simulation results show that the proposed approach is scalable and achieves a near-optimal performance.

References

[1] M. Chehimi and W. Saad, "Physics-informed quantum communication networks: A vision toward the quantum internet," *IEEE Network*, vol. 36, no. 5, pp. 32–38, Sep. 2022.

[2] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, "Quantum repeaters: the role of imperfect local operations in quantum communication," *Physical Review Letters*, vol. 81, no. 26, p. 5932, 1998.

[3] C. H. Bennett *et al.*, "Purification of noisy entanglement and faithful teleportation via noisy channels," *Physical review letters*, vol. 76, no. 5, p. 722, 1996.

[4] Y. Zhao and C. Qiao, "Redundant entanglement provisioning and selection for throughput maximization in quantum networks," in *IEEE Conf. on Computer Communications (INFOCOM)*, 2021, pp. 1–10.

[5] M. Pant, H. Krovi, D. Towsley, L. Tassiulas, L. Jiang, P. Basu, D. Englund, and S. Guha, "Routing entanglement in the quantum internet," *npj Quantum Information*, vol. 5, no. 1, p. 25, 2019.

[6] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the stochastic analysis of a quantum entanglement switch," *ACM SIGMETRICS Performance Evaluation Review*, vol. 47, no. 2, pp. 27–29, 2019.

[7] T. Vasantam and D. Towsley, "Stability analysis of a quantum network with max-weight scheduling," *arXiv preprint arXiv:2106.00831*, 2021.

[8] W. Dai, A. Rinaldi, and D. Towsley, "Entanglement swapping in quantum switches: Protocol design and stability analysis," *arXiv preprint arXiv:2110.04116*, 2021.

[9] N. K. Panigrahy, T. Vasantam, D. Towsley, and L. Tassiulas, "On the capacity region of a quantum switch with entanglement purification," *arXiv preprint arXiv:2212.01463*, 2022.

[10] P. Promponas, V. Valls, and L. Tassiulas, "Full exploitation of limited memory in quantum entanglement switching," *arXiv preprint arXiv:2304.10602*, 2023.

[11] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[12] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Algorithmic Game Theory*, G. Persiano, Ed. Springer Berlin Heidelberg, 2011, pp. 117–129.

[13] T. P. Cope, K. Goodenough, and S. Pirandola, "Converse bounds for quantum and private communication over holevo–werner channels," *Journal of Physics A: Mathematical and Theoretical*, vol. 51, no. 49, p. 494001, 2018.

[14] A. E. Roth and M. A. O. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, ser. Econometric Society Monographs. Cambridge University Press, 1990.

[15] D. Chen *et al.*, "Matching-theory-based low-latency scheme for multitask federated learning in MEC networks," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 415–11 426, 2021.

[16] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: fundamentals and applications," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 52–59, 2015.

[17] B. Hensen *et al.*, "Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres," *Nature*, vol. 526, no. 7575, pp. 682–686, 2015.

[18] N. V. Sahinidis, *BARON 2023.3.11: Global Optimization of Mixed-Integer Nonlinear Programs*, 2017.