Topology Optimization in All-Optical Switched Quantum Data Center Networks

Jiyao Liu*, Lei Fan†‡, Yuanxiong Guo§, Zhu Han‡, and Yu Wang*
*Department of Computer and Information Sciences, Temple University, Philadelphia, USA
†Dept. of Engineering Technology, ‡Dept. of Electrical and Computer Engineering, University of Houston, Houston, USA
§Department of Information Systems and Cybersecurity, University of Texas at San Antonio, San Antonio, USA
{jiyao.liu, wangyu}@temple.edu, {lfan8, zhan2}@uh.edu, yuanxiong.guo@utsa.edu.

***Abstract*—Quantum Data Centers (QDCs) have emerged as a promising architecture for scaling distributed quantum computing beyond the limitations of a single quantum processor. Although prior studies have demonstrated the potential of QDCs on fixed network topologies, it remains unclear how the topology itself should be designed to better support workload-dependent communication demands. In this paper, we study topology optimization in all-optical switched QDC networks and develop a unified framework that connects physical network design, probabilistic entanglement generation, and distributed circuit execution. We first model the QDC network as a directed multigraph with practical device and path constraints, and then analyze the stochastic latency of parallel entanglement generation under layered circuit execution. Based on this analysis, we formulate a latency-aware topology optimization problem and derive a tractable surrogate objective for efficient solving. Simulation results demonstrate the effectiveness of the proposed framework and provide new insights into the trade-offs among topology flexibility, resource utilization, and communication efficiency in QDC networks.**

***Index Terms*—Topology Design, Quantum Data Center, Quantum Network, Data Center Network, Quantum Computing**

I. Introduction

Large-scale quantum applications are now beyond the capability of a single monolithic quantum processor unit (QPU). As a result, *Quantum Data Centers* (QDCs) [1]–[3], which interconnect multiple quantum computers or QPUs through a high-speed optical quantum network, have become a practical direction for near-future scalable and fault-tolerant quantum computing. In this architecture, distributed circuit execution relies on timely and high-quality remote entanglement generation among connected QPUs. Therefore, the underlying quantum network is no longer a passive transport layer, and it directly determines application-level execution latency and reliability in such distributed quantum computing (DQC).

Most prior studies evaluate distributed quantum execution on fixed topologies [3]–[6] or with simplified communication assumptions [7]–[11]. This leaves an important system question underexplored: *how should we jointly design network topology and runtime scheduling to best support communication demands in quantum data centers?* In all-optical switched QDC networks, topology choices (e.g., switch connectivity, path multiplicity, and reconfiguration) strongly affect photon loss, successful entanglement rates, and ultimately circuit completion time. Ignoring the topology optimization can lead to significant performance loss even when other individual components (such as circuit partitioning, gate mapping or layered execution scheduling) are well optimized.

To overcome the above mentioned challenges, in this paper, we study topology optimization for all-optical switched QDC networks through a unified framework that links physical network design, probabilistic entanglement generation, and distributed circuit execution. Specifically, we model the QDC network as a directed multigraph with practical device, port, and path constraints, and characterize how topology decisions affect photon-path loss, parallel entanglement generation, and ultimately stage completion time under layered execution. Based on this model, we derive a probabilistic latency analysis and formulate a latency-aware topology optimization problem together with tractable surrogate objectives that can be solved efficiently. We then evaluate the resulting topologies under realistic workloads using Monte-Carlo simulation. Overall, the proposed framework provides a systematic way to understand how topology flexibility and resource allocation jointly influence communication efficiency in QDC deployments.

The main contributions of this work are threefold:

- We propose a unified framework for topology optimization in all-optical switched QDC networks that connects hardware constraints, directed photon-path construction, and layered distributed circuit execution.
- We develop a probabilistic analysis of parallel entanglement generation and derive latency-aware optimization objectives and tractable surrogates for topology design under practical device and path constraints.
- We conduct simulation-driven evaluation to quantify performance trade-offs and to compare the optimized topologies with representative baseline data center network architectures.

The remainder of this paper is organized as follows. Section II reviews related work on quantum data centers, quan-

The work is partially supported by the US National Science Foundation (Grant No. CNS-2106761, CNS-2318683, CNS-2318663, ECCS-2302469, ECCS-2541996, and OAC-2417716) and an internal grant provided by the Office of the Vice President for Research (OVPR) at Temple University. The authors used ChatGPT and Prism (OpenAI) to assist with language editing, and all content was reviewed and validated by the authors.

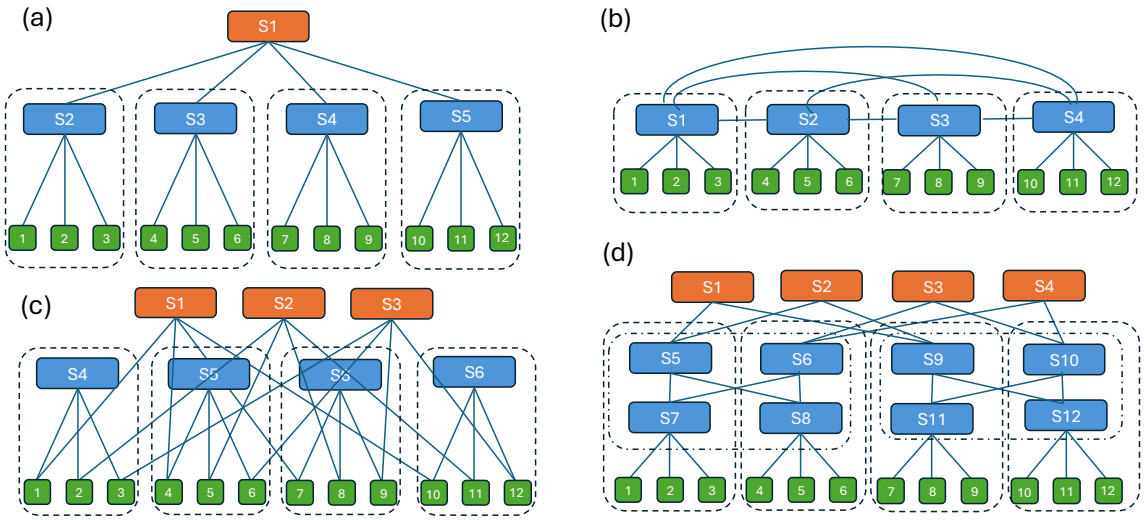

Fig. 1. Network topologies of current data center networks or quantum data center networks (QDCNs). (a) Classical Top-of-Rack (ToR), (b) QFly [12], (c) BCube [13], (d) Fat Tree [5], [6], [14]. Green squares are QPUs, while red or blue rectangles are quantum switches.

tum networks, and distributed circuit execution. Section III introduces our QDC network model and the overall framework for topology optimization and layered circuit execution. Section IV presents the latency-aware topology optimization problem and a probabilistic approximation of the objective to improve practical solvability. Section V reports the evaluation results. Finally, Section VI concludes the paper and discusses future research directions.

II. Background and Related Works

A. Quantum Data Centers

Distributed quantum computing (DQC) [16] is an emerging architecture for scaling beyond the qubit count, connectivity, and error-correction limits of a single quantum processor. Instead of relying on one large monolithic device, DQC partitions computation across multiple processors and coordinates them through entanglement-assisted communication. Building on this idea, Quantum Data Centers (QDCs) [1], [2] organize many quantum computers and networking devices in a shared infrastructure, analogous to classical data centers but with fundamentally different communication primitives.

Recent industrial and academic efforts have demonstrated rapid progress in this direction. Conceptional and prototype systems from major ecosystem players (such as Cisco [3], [4], [17] and IBM [18], [19]) indicate growing interest in modular quantum systems, photonic interconnects, and networked control planes. IBM has advanced a system-level vision through modular architectures such as IBM Quantum System Two, which integrates multiple quantum processors with shared cryogenic infrastructure, classical control, and cloud-based runtimes to enable scalable hybrid quantum–classical workloads [18], [19]. Cisco's Quantum Research [17] is developing a modular quantum data center network architecture [3], [4], scalable and independent from quantum computing platforms while facilitating a flexible interconnect setup. It envisions its ability to create multiple, co-existing, and independent logical interconnect topologies, so that each topology can be specialized for optimizing the performance of specific quantum algorithms. On the device side, IonQ [20] has shifted its strategy from laboratory-grade experimental systems toward data-center-ready quantum computers. Unlike traditional quantum computers that look like "chandeliers" in specialized labs, its Forte Enterprise is built on standard 19-inch racks for the typical, modern data center.

All of these efforts move towards a bright future of scalable QDCs. However, to enable the practical deployments resource constrained communication links are often the bottleneck. This motivates new design approaches that explicitly account for topology, switching behavior, and workload-dependent traffic demands in QDC environments.

B. Quantum Networks

Quantum networking [21]–[27] has been studied across multiple scales, including metropolitan and wide-area settings using fiber, free-space, and satellite links. These works emphasize long-distance entanglement distribution, repeater architectures, and protocol robustness under realistic channel loss and decoherence. While such research provides key theoretical and experimental foundations, its objectives and constraints differ from those of data-center scenarios.

QDC networks (QDCNs) are typically short-to-medium range, highly reconfigurable, and performance-driven. A network topology similar to classical data center network (such as top-of-rack, ToR) can be used for interconnection. For example, Cisco [3], [4] uses topologies of FatTree [14] and BCube [13] for their QDCN design and benchmarking, while [5] and [6] use FatTree as their QDCN architecture. QFly [12] adopts a flattened, low-diameter architecture, Dragonfly topology [28] for interconnection. See Figure 1 for illustration of some of these topologies. However, these topologies are

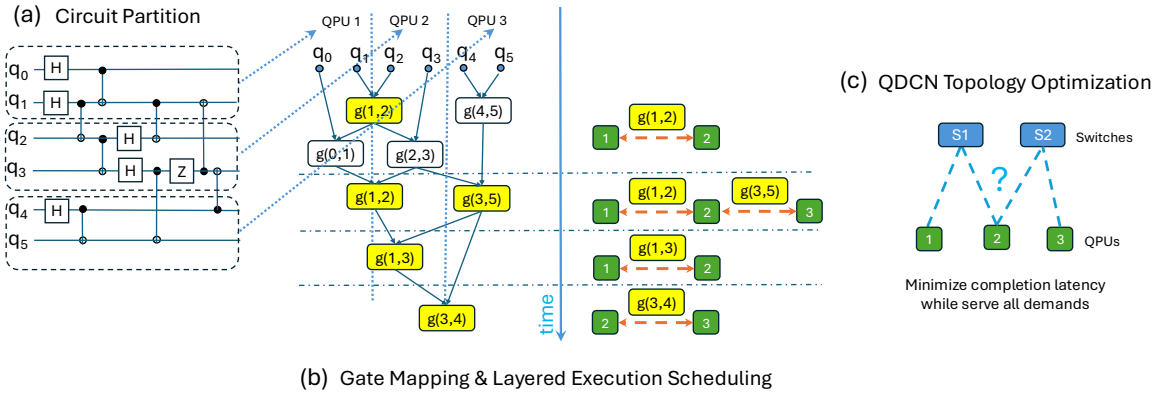

Fig. 2. Illustration of common workflow for distributed circuit execution: circuit partitioning, inter-processor gate mapping, and layered runtime scheduling. (a) A quantum circuit with 6 qubits [15] are partitioned into three parts. (b) Each of part is assigned to one of three QPUs, and yellow rectangles show the remote CNOT gates. Their dependency can be showed in a directed graph, which can be used for layered scheduling. Here single-qubit gates are ignored. Right part of this subfigure shows the entanglement demands from remote gates among QPUs over time. (c) Topology optimization of QDCN studied in this paper aims to find the optimal topology (using quantum switches to connect QPUs and allocating entranglement paths/channels for QPU pairs) which minimizes the completion latency of layered execution to serve all remote gate demands. It considers both physical constraints of devices and probabilistic entanglement generation.

static and designed for general traffic demands. In this paper, we consider the topology optimization via reconfigurable all-optical switched QDCN for specific quantum computing (circuits) demands.

In particular, all-optical switched fabrics are attractive in QDCNs because they can provide high path diversity and low control overhead for photonic communication. Yet this flexibility introduces a new optimization challenge: route and topology decisions directly determine end-to-end photon survival rates, entanglement generation throughput, and thus distributed quantum circuit execution latency. Therefore, QDCN design must tightly couple physical-layer loss modeling with system-level scheduling objectives, rather than treating communication as a fixed black box.

C. Distributed Circuit Execution

Current research in DQC (or distributed circuit execution) focuses on several key challenges such as optimizing qubit allocation [7], [10] or circuit distribution [8], [9], [11], [29]–[34] across nodes, designing efficient network topologies [35]–[37], optimizing quantum compiler framework [38]–[40], and overcoming the inherent noise and decoherence [41]–[43].

A common workflow for distributed circuit execution includes circuit partitioning, inter-processor gate mapping, and layered runtime scheduling, as shown in Figure 2. *Circuit partitioning* decomposes a monolithic quantum circuit into subcircuits assigned to different QPUs, typically by modeling qubit interactions as a graph and minimizing cross-partition edges to reduce costly inter-QPU operations. Following partitioning, *inter-processor gate mapping* determines how nonlocal gates (i.e., operations spanning multiple QPUs) are realized using quantum communication primitives such as entanglement-assisted gates or quantum teleportation, which require careful allocation of limited and error-prone entanglement resources. Finally, *layered runtime scheduling* orchestrates both local computation and communication by organizing operations into dependency-aware execution layers, enabling overlap between entanglement generation and gate execution while mitigating idle time and decoherence. These stages are inherently interdependent, and recent works emphasize their joint optimization to balance communication overhead, resource utilization, and overall execution fidelity in emerging distributed quantum systems. But many of them often assume a static communication substrate and therefore miss opportunities from adaptive topology reconfiguration.

Another limitation is that many evaluations rely on coarse communication abstractions, such as fixed per-link rates or deterministic delays, which do not capture the stochastic nature of parallel entanglement generation. For all-optical switched QDCNs, execution latency depends on both path-level success probabilities and contention across multiple communication pairs over time. This calls for an integrated framework that links probabilistic communication performance, topology construction, and scheduling decisions. In this work, we address this gap by combining a detailed hardware-aware model, a probabilistic latency analysis for parallel channels, and a unified optimization formulation for topology and scheduling to enable communication-efficient QDC.

III. System Model

A. Components of Quantum Data Center Network

We model an all-optical switched QDC network as a directed multigraph $G(V, E)$, where the node set $V = Q \cup S$ consists of quantum computers (QPUs) Q and optical switches S, and the edge set E represents candidate directed fiber links among them. Multiple parallel links between the same pair of devices are allowed, because practical QDC deployments may provision several fibers to increase path diversity and entanglement throughput. In our setting, the physical locations

of quantum computers and switches are assumed to be given, while the logical interconnection topology is determined by the selected links and their usage over time.

Quantum computers. Each quantum computer (QPU) is characterized by three main parameters: the number of network interfaces (ports), the number of communication buffer qubits, and the number of data qubits. The interfaces determine how many optical channels can be attached to the device, the communication qubits limit the number of concurrent remote-entanglement operations, and the data qubits support local circuit execution. For clarity, we assume homogeneous quantum computers in the same data center, although the framework can be extended to heterogeneous platforms.

Quantum links. A physical link is characterized by its length, attenuation, and propagation speed. These parameters determine two quantities that are most relevant to execution: communication latency and photon survival probability. We consider an emitter-scatter style entanglement generation protocol [3], which is compatible with photonic synchronization in short-range data-center environments. Following the standard loss model, if the total optical loss along a photon path is L dB, then the transmission efficiency is

$$\eta_t = 10^{-L/10}. \tag{1}$$

Combined with the detection efficiency η_d, the average entanglement generation rate of one channel is approximated by

$$R = \frac{\eta_t \eta_d}{2(\tau_0 + \tau_b)}, \tag{2}$$

where τ_0 is the round-trip propagation and control delay, and τ_b is the local buffering or reset time [3].

Optical switches and converters. Each optical switch is described by its number of ports, reconfiguration latency, and insertion loss. Switches enable the network to create different photon paths for different communication demands across time windows. We also allow path-dependent device loss introduced by components such as connectors, emitters, scatter modules, and, when needed, quantum frequency converters (QFCs). However, in the main QDC setting considered here, QFCs are not used as themselves introduce significant noise but give marginal benefit within the data center range.

Overall, a photon path between a pair of quantum computers may traverse multiple switches and links, and its survival rate is determined by the accumulated loss of all traversed components. This component-level abstraction enables us to connect physical topology choices directly to end-to-end entanglement performance.

B. Circuit Partition and Execution

Given a monolithic quantum circuit (could be a group of quantum circuits), we first partition it into subcircuits that can be executed on different quantum computers. The partition determines which qubits are assigned to which QPUs and identifies the two-qubit operations whose endpoints lie on different QPUs (such as those yellow gates in Figure 2(b)). These nonlocal operations must be supported through remote entanglement generation and quantum communication. After partitioning, the circuit is transformed into a layered execution schedule so that operations with no unresolved dependency can be grouped into the same execution stage.

We use global time slots to index these execution stages. At each global time slot t, the distributed circuit induces a communication demand matrix over QPU pairs. Specifically, $R_{i,j}^t$ denotes the number of entangled pairs required between quantum computers i and j in order to serve the nonlocal operations scheduled in stage t. Because several QPU pairs may request entanglement simultaneously, the network must decide how to allocate physical paths and parallel channels across these demands.

Within one global time slot, entanglement generation is repeated over smaller local time slots until the required number of successful entanglements has been accumulated for every active QPU pair. This separation between global scheduling intervals and local entanglement attempts is important: the switch configuration remains fixed during one global slot, while the stochastic entanglement process unfolds over many local trials. As a result, the completion time of a stage depends jointly on the requested entanglement volume, the number of parallel channels assigned to each QPU pair, and the success probability of each corresponding photon path.

The execution latency of the whole distributed circuit is therefore determined by the sum of stage-wise completion times. In addition to latency, the quality of served entanglement also matters because lower-fidelity links can degrade the reliability of remote gates. In this paper, we focus primarily on minimizing execution time while capturing the physical loss mechanisms that govern entanglement success rates.

C. Overall Framework

Our overall framework decouples topology design from circuit execution while preserving the interactions between them. The input consists of: 1) a set of quantum computers and candidate switch locations in the data center, 2) hardware parameters such as port counts, device loss, and link length, and 3) one or more distributed quantum workloads represented by their stage-wise entanglement demands as shown in Figure 2(b). Based on these inputs, the framework constructs a topology design problem that decides how many physical links to provision between devices and how these links should be used over time (with possible reconfiguration).

For a fixed topology, the execution layer provides the demand sequence $\{R_{i,j}^t\}$ across global time slots. This is usually the output from circuit partitioning, gate mapping and layered scheduling. They can be the output of existing schedulers, such as Cisco's [3]. In this paper, to enable topology optimization, we use a similar scheduler but only consider the quantum computer resource constraints. However, when optimizing, our solution still only use the available devices (switches) used in other topology. For each slot, the topology-control layer selects directed photon paths and parallel channels for active QPU pairs, which in turn determines their path loss $L_{i,j}^{t,k}$, success probability $p_{i,j}^{t,k}$, and expected entanglement generation rate.

These quantities are then fed into a probabilistic latency model to estimate how long the slot lasts before all demands in that stage are satisfied.

As shown in Figure 2(c), the optimization objective is to find a topology and corresponding time-varying path allocation that minimize end-to-end circuit completion time under practical device constraints, such as QPU interface limits, switch port capacities, and link availability. Because exact latency evaluation is expensive, we later derive a tractable approximation that retains the key dependence on path quality and parallelism. The final designs can then be compared using both analytical objectives and simulation-based evaluation.

Overall, this framework provides a clean interface among hardware modeling, distributed circuit execution, and topology optimization.

IV. Problem Formulation

This section formulates the latency-aware topology optimization problem for all-optical switched QDC networks. Our ultimate objective is to minimize the end-to-end execution latency of distributed quantum workloads. However, directly embedding the exact stochastic latency into a combinatorial optimization problem is difficult. We therefore proceed in two steps. First, we characterize the entanglement-generation latency induced by a given topology and path allocation. Second, we derive a tractable surrogate objective that preserves the dependence on path quality and parallelism, while remaining suitable for optimization. We also provide detailed constraints that connect topology design, path construction and channel usage.

A. Statistical Analysis - Entanglement Generation Latency and Circuit Completion Time

Consider $K_{i,j}^t$ parallel channels (i.e. quantum paths) between a QPU pair (i,j) during global time slot t, where channel k succeeds with probability $p_{i,j}^{t,k}$ in each local entanglement-generation attempt (depending on the path allocated in the topology). Let $S_{i,j}^{t,\tau}$ denote the number of successful entanglements generated for pair (i,j) in local time slot τ, and let $N_{i,j}^{t,\tau}$ denote the total number of successful entanglements accumulated after τ local slots.

Per-slot success distribution. In each local slot,

$$S_{i,j}^{t,\tau} = \sum_{k=1}^{K_{i,j}^t} S_{i,j}^{t,\tau,k}, \qquad S_{i,j}^{t,\tau,k} \sim \text{Bernoulli}(p_{i,j}^{t,\tau,k}).$$

Because the switch configuration is fixed within one global time slot, the success probability of each channel does not depend on the local index τ. Hence we write $p_{i,j}^{t,k} \triangleq p_{i,j}^{t,\tau,k}$ for any τ. Accordingly, $S_{i,j}^{t,\tau}$ is identically distributed across local slots and follows a Poisson–binomial distribution with probability generating function (PGF)

$$G_{S_{i,j}^t}(z) = \prod_{k=1}^{K_{i,j}^t} \left((1-p_{i,j}^{t,k}) + z p_{i,j}^{t,k}\right). \tag{3}$$

Accumulated successes. Since local slots are independent, the total number of successes after τ slots is

$$N_{i,j}^{t,\tau} = \sum_{\iota<\tau} S_{i,j}^{t,\iota}, \tag{4}$$

and its PGF is

$$G_{N_{i,j}^{t,\tau}}(z) = \left(G_{S_{i,j}^t}(z)\right)^{\tau}. \tag{5}$$

Therefore, the probability of obtaining exactly n successful entanglements after τ local slots is

$$\mathbb{P}(N_{i,j}^{t,\tau} = n) = [z^n]\left(G_{S_{i,j}^t}(z)\right)^{\tau}, \tag{6}$$

where $[z^n]f(z)$ denotes the coefficient of z^n in the expansion of $f(z)$.

Raising the PGF to the τ-th power corresponds to summing τ independent Poisson–binomial random variables. Thus, $N_{i,j}^{t,\tau}$ is itself a Poisson–binomial distribution over the multiset $\{p_1,\dots,p_K\}$, where each p_k is repeated t times.

Latency model. Let $R_{i,j}^t$ be the entanglement demand for pair (i,j) in global time slot t. The completion time for that pair is the first local slot at which the accumulated number of successes reaches the target demand:

$$D_{i,j}^t \triangleq \min\{\tau \mid N_{i,j}^{t,\tau} \geq R_{i,j}^t\}. \tag{7}$$

Since all communication demands in the same global slot must be satisfied before execution can proceed, the duration of slot t is determined by the slowest active pair,

$$D^t \triangleq \max_{(i,j)\in Q^{(2)}} D_{i,j}^t, \tag{8}$$

where Q is the set of quantum computers and $Q^{(2)}$ denotes all ordered QPU pairs with distinct endpoints. The total circuit completion time is then

$$D \triangleq \sum_t D^t = \sum_t \max_{(i,j)\in Q^{(2)}} D_{i,j}^t. \tag{9}$$

Note that in this work we focus on communication-induced latency and ignore local computational delay because i) local gate execution is typically much faster than remote entanglement generation; ii) local gates are already determined and does not rely on network topology optimization.

B. Optimization Objective - Latency Minimization and Approximation

Given the entanglement requests in each global interval, we need to suffice them with the resource networks. Among intervals, they are independent. Thus, to optimize the total execution time, we can optimize for each global interval, and the final execution time will be the summation of all intervals. The final topology is unchanged across intervals. To enable optimization, we need to first express the execution time in terms of the network topology to formulate the problem, which is non-trivial.

The exact quantity D depends on the stochastic evolution of all parallel entanglement-generation channels, making it difficult to optimize directly. To connect the physical topology

of QDCN with the execution latency, we first express the expected service rate for each QPU pair under a given path allocation.

Recall that we model the physical QDCN as a directed multigraph $G(V, E)$, where V contains all QPUs and switches, and E contains all candidate directed links, including parallel links. Suppose that $K_{i,j}^t$ channels are assigned to pair (i, j) in global time slot t. If channel k uses a photon path with total loss $L_{i,j}^{t,k}$, then its success probability is $p_{i,j}^{t,k} = 10^{-L_{i,j}^{t,k}/10}$. The expected number of successful entanglements produced in one local slot ($K_{i,j}^t$ channels attempt once together) is therefore

$$\mu_{i,j}^t = \sum_{k=1}^{K_{i,j}^t} p_{i,j}^{t,k} = \sum_{k=1}^{K_{i,j}^t} 10^{-L_{i,j}^{t,k}/10}. \tag{10}$$

Hence, topology optimization affects latency through two coupled mechanisms: the number of channels assigned to a QPU pair and the noise of the underlying photon paths.

Our original objective is to minimize the total completion time,

$$\min D = \min \sum_t \max_{(i,j)\in Q^{(2)}} D_{i,j}^t. \tag{11}$$

To make this objective tractable (solvable by a classic solver), we use the following bound on the expected completion time of each pair.

***Theorem** 1:* By overshoot arguments and Wald's identity, the expected completion time of pair (i, j) in slot t satisfies

$$\frac{R_{i,j}^t}{\mu_{i,j}^t} \le \mathbb{E}[D_{i,j}^t] \le \frac{R_{i,j}^t + K_{i,j}^t - 1}{\mu_{i,j}^t}. \tag{12}$$

The proof of this theorem is provided in the Appendix. This result shows that the expected delay is inversely proportional to the aggregate service rate $\mu_{i,j}^t$. Motivated by this observation, we introduce the following approximated latency (AL):

$$\hat{D} \triangleq \sum_t \hat{D}^t \tag{13}$$

$$\triangleq \sum_t \max_{(i,j)\in Q^{(2)}} \hat{D}_{i,j}^t \tag{14}$$

$$\triangleq \sum_t \max_{(i,j)\in Q^{(2)}} \frac{\sum_k (y_{i,j}^{t,k} + y_{j,i}^{t,k})}{\sum_k (p_{i,j}^{t,k} + p_{j,i}^{t,k})}, \tag{15}$$

where $y_{i,j}^{t,k}$ denotes the amount of demand assigned to channel k, and the two directions are both included because the physical network is modeled as a directed graph to accommodate the ports directions of switches.

Although (15) is much more informative than the exact stochastic objective, it is still difficult to optimize because the decision variables appear inside exponential and fractional terms. We therefore use a simpler weighted-noise objective (WN) as the optimization target,

$$\min \sum_t \sum_{(i,j)\in Q^{(2)}} \sum_k y_{i,j}^{t,k} L_{i,j}^{t,k}, \tag{16}$$

which favors topologies and channel assignments that allocate heavily used demands to low-loss photon paths. This surrogate is later evaluated against the more accurate latency metric in simulation. Note that (15) is an intuitive approximation to latency and does not yet have a clear relation with (16).

C. Constraints and Decision Variables

We next present the constraints that connect topology design, path construction, and channel usage.

Device port constraints. Let $w_{u,v}$ denote the number of physical fiber links provisioned from node u to node v. Then the available switch input ports, switch output ports, and QPU interfaces impose

$$\sum_{u\in V\setminus\{v\}} w_{u,v} \le P_v^{\text{in}}, \qquad \forall v \in S, \tag{17}$$

$$\sum_{v\in V\setminus\{u\}} w_{u,v} \le P_u^{\text{out}}, \qquad \forall u \in S, \tag{18}$$

$$\sum_{v\in V\setminus\{u\}} (w_{u,v} + w_{v,u}) \le P_u^{\text{io}}, \qquad \forall u \in Q. \tag{19}$$

These constraints ensure that the designed topology is physically realizable under device-level interface limits. Here, we assume that each switch has fixed in-port and out-port, but QPU's ports can serve as both in-port and out-port [17]

Link sharing across time slots. Let $e_{i,j,u,v}^{t,k}$ be a binary indicator that equals one if physical link (u, v) is used by channel/path k for pair (i, j) during global time slot t. Since all channels configured in the same slot must share the installed physical infrastructure, we require

$$\sum_{(i,j)\in R_{i,j}^t} \sum_{k<K_{i,j}} e_{i,j,u,v}^{t,k} \le w_{u,v}, \quad \forall (u, v) \in V^{(2)}, t < T. \tag{20}$$

This constraint guarantees that the number of simultaneously activated channels on any directed link does not exceed the number of installed fibers.

Directed photon path constraints. For each active pair (i, j) and channel k, the binary variable $x_{i,j}^{t,k}$ indicates whether that channel is established. The following constraints enforce a valid directed path from source QPU i to destination QPU j for $\forall (i, j) \in R_{i,j}^t, t < T, k < K_{i,j}$:

$$\sum_{v\in S(j)} e_{i,j,i,v}^{t,k} = x_{i,j}^{t,k}, \tag{21}$$

$$\sum_{v\in S(j)\setminus\{u\}} e_{i,j,u,v}^{t,k} = \sum_{v\in S(i)\setminus\{u\}} e_{i,j,v,u}^{t,k}, \forall u \in S, \tag{22}$$

$$\sum_{u\in S(i)} e_{i,j,u,j}^{t,k} = x_{i,j}^{t,k}, \tag{23}$$

$$\sum_{(u,v)\in S(i)\times S(j)} e_{i,j,u,v}^{t,k} \le x_{i,j}^{t,k}(|S| + 1). \tag{24}$$

The first and third constraints enforce source and destination incidence, the second imposes flow conservation at each intermediate switch, and the last prevents excessively long

paths. Here, $S(i)$ and $S(j)$ denote the switch sets excluding the corresponding endpoint definitions used by the model.

Path loss and success probability. For each feasible channel, the total path loss for $\forall(i,j) \in Q^{(2)}, t < T, k < K_{i,j}$ is defined by

$$L_{i,j}^{t,k} \triangleq (1 - x_{i,j}^{t,k})a_{\mathrm{m}} + \sum_{(u,v)\in S(i)\times S(j)} a_{u,v} e_{i,j,u,v}^{t,k}, \tag{25}$$

$$p_{i,j}^{t,k} \triangleq 10^{-L_{i,j}^{t,k}/10}. \tag{26}$$

Here, $a_{u,v}$ is the loss contributed by directed link (u,v) and a_{m} is a sufficiently large constant to prevent the usage of unestablished path. These expressions connect binary path decisions with the probabilistic entanglement model introduced earlier.

Channel usage variables. Let $y_{i,j}^{t,k}$ denote the amount of entanglement demand assigned to channel k for pair (i,j) in slot t. The following bounds couple usage and channel activation for $\forall(i,j) \in Q^{(2)}, t < T, k < K_{i,j}$:

$$x_{i,j}^{t,k} \le y_{i,j}^{t,k} \le x_{i,j}^{t,k} R_{\mathrm{m}}^{t}. \tag{27}$$

Therefore, an inactive channel cannot carry any demand, while an active channel can serve at most the maximum request R_{m}^{t} in one global slot.

Request fulfillment. Finally, the generated entanglements should be able to serve all requests

$$\sum_{k} (y_{i,j}^{t,k} + y_{j,i}^{t,k}) \ge R_{i,j}^{t}, \qquad \forall t, (i,j) \in Q^{(2)}. \tag{28}$$

Variable domains. The decision variables satisfy

$$w, y \in \mathbb{N}_0, \qquad x, e \in \{0,1\}.$$

In summary, our topology optimization aims to solve the topology decision and corresponding time-varying path allocation for objective (16) under the above constraints.

V. Evaluations

In this section, we evaluate the effectiveness of the proposed topology optimization framework for all-optical switched QDC networks. We aim to answer two key questions: (1) how well the proposed surrogate objective aligns with the simulated latency metric by the Monte-Carlo method (similar to Cisco's work [3]), and (2) how the optimized topology compares with existing data center network designs. We solve the formulated problem by Gurobi [44].

A. Experimental Setup

Workloads. We evaluate our framework using four representative distributed quantum circuits:

- **QFT**: Quantum Fourier Transform, which is an important operation for many algorithms, such as Shor's algorithm;
- **MCMT**: Multi-controlled multi-target gates, which is used for quantum compiler benchmarking to test their gate decomposition ability;
- **RCA**: Ripple-carry adder, which is an important component for many cryptographic algorithms;
- **Hybrid**: A mixed workload combining the above patterns.

All circuits are mapped to distributed settings and partitioned across multiple QPUs in the same way as Cisco's [3].

Baselines. We compare our optimized topology (denoted as **OPT**) with several representative data center network topologies:

- **OPT**: topology optimized using the proposed method;
- **ToR**: tree-based top-of-rack architecture;
- **QFly**: a flattened butterfly-style topology adapted for QDCNs;
- **Clos**: generalized Fat-Tree topology.

For fair comparison, these topologies are also used in Cisco [3], [4] and all baselines are configured under identical hardware constraints.

If not mentioned otherwise, each quantum computer has 4 communication and computing qubits, and 4 interface to emit and receive photons. The loss at interface is set to 5 dB. In-rack switches have 8 ports for both in and out directions, with 1 dB photon loss. Cross-rack switches have 16 ports for both directions with 2 dB noise. In-rack links and cross-rack links have 1 dB and 2 dB noise. These settings are different (typically smaller) than those in existing works because we only consider the quantum data center limited with a small range, e.g., within a building and no more than 1 km between any two components.

B. Impact of Objective Function

We first evaluate how the two proposed objectives compare in terms of solution quality and computational complexity. Figure 3 compares between the weighted-noise (WN) objective (i.e., objective (16)) and approximated latency (AL) objective (i.e., objective (15)) under different solver time limits. The x-axis shows the solver time limit (seconds), and the y-axis shows the resulting execution latency measured in average time slots (given by Monte-Carlo simulation as in Cisco [17]). Solid lines represent WN, while dashed lines represent AL. Different colors correspond to different problem sizes (e.g., 16, 24, and 32 qubits distributed to 2, 3, and 4 racks with 2 quantum computers in each rack).

Two key observations can be drawn from Figure 3. First, for small-scale problem instances, the weighted-loss objective (i.e., WN) can be solved to optimality within a short time. However, due to its approximate nature, the resulting solutions are suboptimal with respect to the true execution latency, and therefore do not achieve performance as good as the latency-aware objective (i.e., AL).

Second, although the latency-aware objective (i.e., AL) provides a more accurate representation of the execution delay, it is significantly more challenging to solve. The optimization converges much more slowly, and for larger problem instances, it may fail to find even a feasible solution within a reasonable time limit.

These results highlight a fundamental trade-off between accuracy and tractability. While WN is less accurate, it is computationally efficient and scalable, making it suitable for practical topology design. In contrast, AL offers higher fidelity but suffers from prohibitive computational overhead.

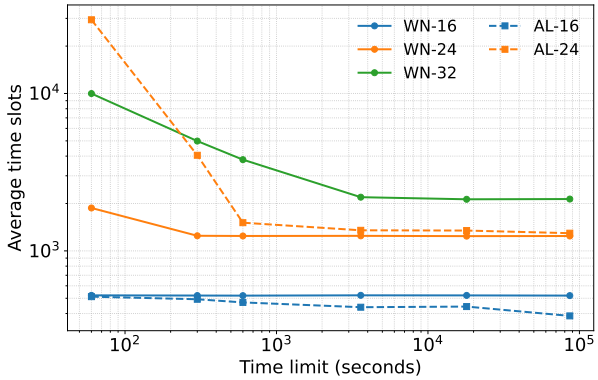

Fig. 3. Achieved average total latency (multiples of entanglement generation slot) against solver time limit by the two objectives with different sizes.

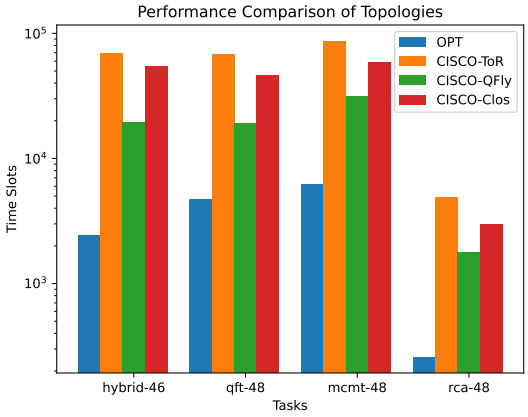

Fig. 4. Achieved total latency by different solutions for different quantum workload tasks on the selected topologies. For ToR, QFly, Clos, we use the scheduler provided in [3].

This observation motivates the need for improved surrogate objectives or solution methods that can better balance accuracy and efficiency.

C. Comparison with Baseline Topologies

Figure 4 shows performance comparison of different network topologies across representative quantum workloads. The x-axis shows the tested circuits (Hybrid, QFT, MCMT, and RCA), and the y-axis shows the execution latency in time slots. The legend includes the proposed optimized topology and scheduling (OPT) and Cisco's scheduler on baseline topologies (Cisco-ToR/QFly/Clos). From Figure 4, we make two key observations.

First, the optimized topology consistently achieves significantly lower execution latency across all tested circuits. This demonstrates that topology optimization is beneficial for both workload-specific (dedicated) scenarios and more general-purpose settings involving a mix of applications. In some cases, the improvement reaches up to an order of magnitude, which is particularly important in quantum systems where limited coherence time constrains the feasible execution window. Such a reduction can effectively determine whether a distributed quantum computation is practical or not on a given hardware platform.

Second, the results highlight the fundamental advantage of topology-aware design in quantum data centers. Conventional data center network topologies, which are designed for classical traffic patterns, do not adequately capture the unique characteristics of quantum communication, such as probabilistic entanglement generation and path-dependent loss. As a result, they are inherently suboptimal for QDC workloads. In contrast, the proposed approach adapts the topology to the underlying communication demands and physical constraints, leading to substantially improved performance.

D. Summary

Overall, the evaluation results demonstrate that the proposed topology optimization framework effectively improves distributed quantum circuit execution performance. The weighted-loss objective provides a practical and scalable approximation, while the optimized topology consistently outperforms conventional data center designs across diverse workloads. These findings highlight the importance of topology-aware design in QDCNs and confirm that adapting network structure to quantum communication characteristics can significantly reduce execution latency.

VI. Conclusion and Future Directions

In this paper, we studied topology optimization for all-optical switched QDCNs with the objective of minimizing distributed quantum circuit execution latency. We developed a unified framework that connects physical-layer characteristics, probabilistic entanglement generation, and system-level scheduling. By deriving a tractable latency surrogate and introducing a weighted-loss optimization objective, we enabled efficient topology design under realistic hardware constraints. Simulation results demonstrate that the proposed approach significantly outperforms conventional data center topologies by better exploiting path diversity and aligning network resources with workload demands.

There are several promising directions for future work. First, incorporating entanglement swapping into the model would enable more flexible and scalable communication patterns, especially in server-centric or multi-hop quantum network architectures. Second, improving the scalability of the optimization framework remains an important challenge, particularly for large-scale systems with many QPUs and deep circuit dependencies. Third, emerging entanglement generation technologies with heterogeneous characteristics could be integrated into the framework to further expand the design space. Finally, extending the model to explicitly account for fidelity and error accumulation, and jointly optimizing latency and reliability, would provide a more comprehensive foundation for practical QDC deployment.

REFERENCES

[1] J. Liu, C. T. Hann, and L. Jiang, "Data centers with quantum random access memory and quantum networks," *Physical Review A*, vol. 108, no. 3, p. 032610, 2023.

[2] J. Liu and L. Jiang, "Quantum data center: Perspectives," *IEEE Network*, vol. 38, no. 5, pp. 160–166, 2024.

[3] H. Shapourian, E. Kaur, T. Sewell, J. Zhao, M. Kilzer, R. Kompella, and R. Nejabati, "Quantum data center infrastructures: A scalable architectural design perspective," *arXiv preprint arXiv:2501.05598*, 2025. [Online]. Available: https://arxiv.org/abs/2501.05598

[4] S. Pouryousef, E. Kaur, H. Shapourian, D. Towsley, R. Kompella, and R. Nejabati, "Benchmarking quantum data center architectures: A performance and scalability perspective," *arXiv preprint arXiv:2601.01353*, 2026. [Online]. Available: https://arxiv.org/abs/2601.01353

[5] H. Zhang, Y. Xu, H. Hu, K. Yin, H. Shapourian, J. Zhao, R. R. Kompella, R. Nejabati, and Y. Ding, "SwitchQNet: Optimizing distributed quantum computing for quantum data centers with switch networks," in *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA '25)*, New York, NY, USA: Association for Computing Machinery, 2025, p. 1449–1463.

[6] Y. Xin and L. Zhang, "Towards a high-performance quantum data center network architecture," in *Proc. of IEEE International Conference on Communications (ICC 2025)*, 2025, pp. 2400–2405.

[7] Y. Mao, Y. Liu, and Y. Yang, "Probability-aware qubit-to-processor mapping in distributed quantum computing," in *Proceedings of the 1st Workshop on Quantum Networks and Distributed Quantum Computing*, 2023, p. 51–56.

[8] R. G Sundaram, H. Gupta, and C. Ramakrishnan, "Efficient distribution of quantum circuits," in *Proc. of 35th International Symposium on Distributed Computing (DISC 2021)*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[9] R. G. Sundaram and H. Gupta, "Distributing quantum circuits using teleportations," in *Proc. of 2023 IEEE International Conference on Quantum Software (QSW)*, 2023, pp. 186–192.

[10] Y. Mao, Y. Liu, and Y. Yang, "Qubit allocation for distributed quantum computing," in *Proc. of IEEE Conference on Computer Communications (INFOCOM 2023)*, 2023, pp. 1–10.

[11] R. G. Sundaram, H. Gupta, and C. Ramakrishnan, "Distribution of quantum circuits over general quantum networks," in *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*, 2022, pp. 415–425.

[12] D. Sakuma, A. Taherkhani, T. Tsuno, T. Sasaki, H. Shimizu, K. Teramoto, A. Todd, Y. Ueno, M. Hajdušek, R. Ikuta *et al.*, "Q-Fly: An optical interconnect for modular quantum computers," *arXiv preprint arXiv:2412.09299*, 2024. [Online]. Available: https://arxiv.org/abs/2412.09299

[13] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: a high performance, server-centric network architecture for modular data centers," in *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication*, 2009, pp. 63–74.

[14] C. E. Leiserson, "Fat-trees: Universal networks for hardware-efficient supercomputing," *IEEE transactions on Computers*, vol. 100, no. 10, pp. 892–901, 2012.

[15] L. Lao, H. van Someren, I. Ashraf, and C. G. Almudever, "Timing and resource-aware mapping of quantum circuits to superconducting processors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 2, p. 359–371, Feb. 2022.

[16] M. Caleffi, M. Amoretti, D. Ferrari, J. Illiano, A. Manzalini, and A. S. Cacciapuoti, "Distributed quantum computing: A survey," *Computer Networks*, vol. 254, p. 110672, 2024.

[17] Cisco Research, "Quantum data center," 2026. [Online]. Available: https://research.cisco.com/quantum/quantum-data-center

[18] IBM Quantum, "IBM quantum roadmap 2033," 2023. [Online]. Available: https://www.ibm.com/roadmaps/quantum/

[19] R. Mandelbaum, J. Gambetta, J. Chow, T. Mittal, T. J. Yoder, A. Cross, and M. Steffen, "How IBM will build the world's first large-scale, fault-tolerant quantum computer," *IBM Research Blog*, 2025. [Online]. Available: https://www.ibm.com/quantum/blog/large-scale-ftqc

[20] IonQ, "IonQ Forte Enterprise," 2026. [Online]. Available: https://www.ionq.com/quantum-systems/forte-enterprise

[21] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, p. eaam9288, 2018.

[22] A. S. Cacciapuoti, M. Caleffi, F. Tafuri, F. S. Cataliotti, S. Gherardini, and G. Bianchi, "Quantum internet: Networking challenges in distributed quantum computing," *IEEE Network*, vol. 34, no. 1, pp. 137–143, 2019.

[23] X. Fan, Y. Yang, H. Gupta, and C. R. Ramakrishnan, "Distribution and purification of entanglement states in quantum networks," in *Proc. of 2025 International Conference on Quantum Communications, Networking, and Computing (QCNC)*, 2025, pp. 74–82.

[24] J. Liu and Y. Wang, "Statistical modeling and latency optimization for entanglement routing in quantum networks," in *Proceedings of IEEE International Conference on Quantum Computing and Engineering (QCE 2025)*, 2025.

[25] J. Liu, X. Zhang, X. Wei, X. Liu, Y. Chen, H. Gao, and Y. Wang, "Swapping and purification scheme optimization for entanglement distribution in quantum networks," *IEEE Transactions on Networking*, vol. 34, pp. 3378 – 3392, 2026.

[26] J. Liu, L. Fan, Y. Guo, Z. Han, and Y. Wang, "Minimum cost swapping and purification scheduling over quantum repeater chains," in *Proceedings of IEEE International Conference on Quantum Communications, Networking, and Computing (QCNC 2026)*, 2026.

[27] X. Fan, C. Zhan, H. Gupta, and C. R. Ramakrishnan, "Optimized distribution of entanglement graph states in quantum networks," *IEEE Transactions on Quantum Engineering*, vol. 6, pp. 1–17, 2025.

[28] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," *ACM SIGARCH Computer Architecture News*, vol. 36, no. 3, pp. 77–88, 2008.

[29] D. Dadkhah, M. Zomorodi, S. E. Hosseini, P. Plawiak, and X. Zhou, "Reordering and partitioning of distributed quantum circuits," *IEEE Access*, vol. 10, pp. 70 329–70 341, 2022.

[30] J. M. Baker, C. Duckering, A. Hoover, and F. T. Chong, "Time-sliced quantum circuit partitioning for modular architectures," in *Proceedings of the 17th ACM International Conference on Computing Frontiers*, 2020, pp. 98–107.

[31] O. Daei, K. Navi, and M. Zomorodi-Moghadam, "Optimized quantum circuit partitioning," *International Journal of Theoretical Physics*, vol. 59, no. 12, pp. 3804–3820, 2020.

[32] Z. Davarzani, M. Zomorodi-Moghadam, M. Houshmand, and M. Nouri-Baygi, "A dynamic programming approach for distributing quantum circuits by bipartite graphs," *Quantum Information Processing*, vol. 19, pp. 1–18, 2020.

[33] D. Dadkhah, M. Zomorodi, and S. E. Hosseini, "A new approach for optimization of distributed quantum circuits," *International Journal of Theoretical Physics*, vol. 60, pp. 3271–3285, 2021.

[34] P. Andres-Martinez and C. Heunen, "Automated distribution of quantum circuits via hypergraph partitioning," *Physical Review A*, vol. 100, no. 3, p. 032308, 2019.

[35] J. Liu, X. Liu, X. Wei, and Y. Wang, "Topology design with resource allocation and entanglement distribution for quantum networks," in *Proc. of 21st Annual IEEE International Conference on Sensing, Communication, and Networking (SECON 2024)*, 2024.

[36] Y. Mao, Y. Liu, X. Shang, and Y. Yang, "Network topology design for distributed quantum computing," in *Proc. of 2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, 2024, pp. 1213–1223.

[37] J. Liu, L. Fan, Y. Guo, Z. Han, and Y. Wang, "Co-design of network topology and qubit allocation for distributed quantum computing," in *Proceedings of IEEE International Conference on Quantum Communications, Networking, and Computing (QCNC 2025)*, 2025.

[38] D. Ferrari, A. S. Cacciapuoti, M. Amoretti, and M. Caleffi, "Compiler design for distributed quantum computing," *IEEE Transactions on Quantum Engineering*, vol. 2, pp. 1–20, 2021.

[39] D. Ferrari, S. Carretta, and M. Amoretti, "A modular quantum compilation framework for distributed quantum computing," *IEEE Transactions on Quantum Engineering*, vol. 4, pp. 1–13, 2023.

[40] D. Cuomo, M. Caleffi, K. Krsulich, F. Tramonto, G. Agliardi, E. Prati, and A. S. Cacciapuoti, "Optimized compiler for distributed quantum computing," *ACM Transactions on Quantum Computing*, vol. 4, no. 2, pp. 1–29, 2023.

[41] J. Avron, O. Casper, and I. Rozen, "Quantum advantage and noise reduction in distributed quantum computing," *Physical Review A*, vol. 104, no. 5, p. 052404, 2021.

[42] A. M. Gomez, T. L. Patti, A. Anandkumar, and S. F. Yelin, "Near-term distributed quantum computation using mean-field corrections and auxiliary qubits," *Quantum Science and Technology*, vol. 9, no. 3, p. 035022, 2024.

[43] M. Prest and K.-C. Chen, "Quantum-error-mitigated detectable Byzantine agreement with dynamical decoupling for distributed quantum computing," *arXiv preprint arXiv:2311.03097*, 2023. [Online]. Available: https://arxiv.org/abs/2311.03097
[44] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: https://www.gurobi.com

APPENDIX

We prove the following two inequalities in Theorem 1 separately.

$$\frac{R_{i,j}^t}{\mu_{i,j}^t} \le \mathbb{E}[D_{i,j}^t] \le \frac{R_{i,j}^t + K_{i,j}^t - 1}{\mu_{i,j}^t},$$

where $\mu_{i,j}^t = \sum_{k=1}^{K_{i,j}^t} p_{i,j}^{t,k}$.

Recall that the per-slot number of successful entanglements is

$$S_{i,j}^{t,\tau} = \sum_{k=1}^{K_{i,j}^t} S_{i,j}^{t,\tau,k},$$

and obviously

$$\mathbb{E}[S_{i,j}^{t,\tau}] = \mu_{i,j}^t.$$

By Wald's identity,

$$\mathbb{E}[N_{i,j}^{t,D_{i,j}^t}] = \mathbb{E}\left[\sum_{\iota < D_{i,j}^t} S_{i,j}^{t,\iota}\right] = \mathbb{E}[S_{i,j}^{t,\iota}]\mathbb{E}[D_{i,j}^t] = \mu_{i,j}^t \mathbb{E}[D_{i,j}^t].$$

By definition of the stopping time,

$$N_{i,j}^{t,D_{i,j}^t} \ge R_{i,j}^t,$$

thus

$$\mu_{i,j}^t\, \mathbb{E}[D_{i,j}^t] = \mathbb{E}[N_{i,j}^{t,D_{i,j}^t}] \ge R_{i,j}^t,$$
$$\mathbb{E}[D_{i,j}^t] \ge \frac{R_{i,j}^t}{\mu_{i,j}^t}.$$

For the upper bound, note that just before stopping,

$$N_{i,j}^{t,D_{i,j}^t - 1} \le R_{i,j}^t - 1.$$

Moreover, in any local slot, at most all $K_{i,j}^t$ parallel channels can succeed, so

$$S_{i,j}^{t,D_{i,j}^t} \le K_{i,j}^t.$$

Hence

$$N_{i,j}^{t,D_{i,j}^t} = N_{i,j}^{t,D_{i,j}^t - 1} + S_{i,j}^{t,D_{i,j}^t} \le (R_{i,j}^t - 1) + K_{i,j}^t = R_{i,j}^t + K_{i,j}^t - 1.$$

Taking expectations and applying Wald's identity again gives

$$\mu_{i,j}^t\, \mathbb{E}[D_{i,j}^t] = \mathbb{E}[N_{i,j}^{t,D_{i,j}^t}] \le R_{i,j}^t + K_{i,j}^t - 1,$$

and therefore

$$\mathbb{E}[D_{i,j}^t] \le \frac{R_{i,j}^t + K_{i,j}^t - 1}{\mu_{i,j}^t}.$$

This completes the proof.