

Title	Trade-off between Fidelity and Latency for Entanglement Routing Design in Quantum Networks
Author(s)	NGUYEN, Thu Trang
Citation	
Issue Date	2025-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/20031
Rights	
Description	Supervisor: リム 勇仁, 先端科学技術研究科, 修士 (情報科学)

Master's Thesis

Trade-off between Fidelity and Latency for Entanglement Routing Design in Quantum Networks

NGUYEN Thu Trang

Supervisor: Professor Yuto LIM

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2025

TRADE-OFF BETWEEN FIDELITY AND LATENCY FOR
ENTANGLEMENT ROUTING DESIGN IN QUANTUM NETWORKS

By NGUYEN Thu Trang (2310436)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Science

Supervisor : Professor LIM, Yuto
Main Examiner : Professor LIM, Yuto
Examiners : Professor TAN, Yasuo
Associate Professor UDA, Satoshi
Associate Professor BEURAN, Razvan Florin

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2025

Abstract

Quantum networks intersect a paradigmatic shift in communication technologies, offering unprecedented capabilities for secure communication and distributed computation through the exploitation of quantum mechanical phenomena. These networks facilitate revolutionary and unlimited of possibility applications such as quantum key distribution (QKD), distributed quantum computing (DQC), quantum teleportation (QT), quantum clock synchronization (QCS), etc. —all fundamentally predicated on quantum entanglement as their operational cornerstone. Moreover, on the contrary with classical computer communication protocol, quantum networks have the fundamental no-cloning theorem in the quantum mechanic prohibits conventional signal amplification techniques, necessitating the development of quantum-native routing strategies that operate within these unique constraints.

However, the practical realization of quantum networks until now still faces a crucial challenge: The reliability and efficient distribution of high-fidelity entangled quantum states across long distances. Due to this distribution, the process faces intrinsic limitations arising from quantum decoherence, photonic loss in transmission channels, and the inherently probabilistic nature of entanglement generation protocols. In quantum communication, fidelity serves as a fundamental metric, representing the degree of similarity or correlation between the actual distributed entangled state and an ideal maximally entangled Bell pair. Fidelity is a crucial measure of the quality of quantum states in a network since it directly affects the reliability and effectiveness of quantum communication protocols. High fidelity ensures that the transmitted entangled states maintain their quantum properties, closely resembling the ideal Bell pair, and are thus suitable for performing secure and efficient quantum operations.

Latency, on the other hand, refers to the total time delay involved in the entire process of entanglement establishment, from the initial creation of the entangled pairs to the final verification and confirmation of the entanglement quality. This includes several distinct components: the physical propagation delay, which accounts for the time it takes for quantum information to travel through the communication channel (such as optical fibers); the entanglement generation time, which involves the probabilistic creation of entangled states between quantum nodes; and the quantum memory storage time, which reflects the duration for which quantum states

are held in memory while awaiting further operations such as entanglement swapping or purification. Although significant research has yielded solutions aimed at achieving high-fidelity entanglement despite these obstacles, the equally critical factor of time delay, or latency, in quantum routing links remains largely unaddressed in current investigations. This oversight is notable, given that latency is a crucial performance bottleneck for real-world quantum applications. Together, fidelity and latency define the efficiency and effectiveness of quantum networks. Achieving a balance between these two factors is a fundamental challenge in quantum routing, wherein high fidelity typically necessitates increased complexity of the routing processes, and conversely, high latency often compromises the quality of the quantum state, so it needs many purification steps.

This thesis addresses this critical challenge by introducing Q_{FiLa} (Quantum Fidelity-Latency), a novel quantum link metric specifically engineered to optimize the fundamental trade-off between quantum state fidelity and communication latency in entanglement routing protocols. Q_{FiLa} tackle this dichotomy by providing a unified optimization framework that enables adaptive, application-aware routing decisions essential for next-generation quantum communication infrastructures. Our research systematically integrates the Q_{FiLa} metric within the established Q-LEAP routing framework, enabling dynamic path selection through parameterized weighted combinations of fidelity and latency objectives. This integration facilitates adaptive routing strategies that can be tailored to diverse network conditions and heterogeneous application requirements. Specifically, Q_{FiLa} empowers routing protocols to intelligently prioritize either minimal latency (critical for time-sensitive applications such as real-time QKD and quantum sensing) or maximal fidelity (essential for high-precision applications including quantum teleportation and fault-tolerant distributed quantum computing), thereby aligning network resource allocation with application-specific requirements.

Before proceeding with the analysis of the Q_{FiLa} metric, we first examine the latency within the Nested Purification Protocol (NPP). This investigation allows us to understand the overall structure of latency in a specific purification protocol. Subsequently, we will extend our analysis to explore latency in the context of individual link routing, providing a detailed examination of latency at a more granular level within the quantum network architecture. To evaluate Q_{FiLa} 's performance characteristics, we conduct comprehensive simulation studies in both realistic and synthetic quantum network topologies. The Japan Photonic Network Model (JPNM) serves as our primary real-world benchmark, accurately reflecting the complexities and constraints of contemporary telecommunications infrastructure. Additionally, we add the simulation to another level by employing systematically

generated random topologies with varying scales and connection densities to assess the metric’s generalizability across diverse network configurations. Our experimental evaluation demonstrates that Q_{FiLa} -enhanced routing algorithms achieve substantial performance improvements over baseline Q-LEAP implementations, exhibiting significant reductions in end-to-end latency and computational overhead while maintaining equivalent throughput characteristics. These efficiency gains are particularly pronounced in high-fidelity operational regimes, attributable to Q_{FiLa} ’s capacity to prioritize near-optimal communication links and effectively prune the routing search space through intelligent weighting mechanisms—an advantage that becomes increasingly critical in large-scale networks where routing computational overhead represents a primary performance bottleneck.

Furthermore, Q_{FiLa} demonstrates remarkable adaptability across diverse network topologies and scales. Comprehensive sensitivity analyses conducted on random network topologies with varying scale factors consistently reveal robust performance characteristics, even in sparse and topologically irregular configurations. This versatility underscores Q_{FiLa} ’s practical viability for deployment across both established telecommunications-inspired quantum networks and emerging decentralized quantum internet architectures.

A principal contribution of this work lies in demonstrating that joint optimization of fidelity and latency parameters yields substantial improvements in overall quantum network performance metrics. Q_{FiLa} provides a computationally efficient, algorithmically flexible, and architecturally scalable solution for addressing this fundamental optimization challenge, thereby establishing a foundation for advanced quality-aware quantum routing protocols. Beyond these immediate contributions, this thesis identifies promising research directions for incorporating additional performance metrics—including hop count, entanglement generation success probability, quantum memory coherence time, and purification protocol overhead—into future routing optimization frameworks, advancing toward more comprehensive and resource-efficient quantum network protocols.

Despite these significant advances, several limitations remain that warrant future investigation. The current study operates under assumptions of static network topologies and homogeneous quantum link characteristics, without explicit modeling of dynamic network conditions, hardware heterogeneity, or entanglement purification failure mechanisms. Future research should incorporate adaptive link characterization based on real-time network feedback, explore multi-flow concurrent routing scenarios, and investigate hybrid classical-quantum routing architectures. Additionally, the development of machine learning-based adaptive parameter tuning for Q_{FiLa} weight optimization represents a particularly promising avenue for further

performance enhancement.

In conclusion, this thesis introduces Q_{FiLa} as a novel and demonstrably effective quantum link metric that substantially advances the state-of-the-art in quantum network routing through joint optimization of fidelity and latency parameters. The proposed approach provides a robust theoretical and practical foundation for scalable, quality-aware quantum communication systems, contributing significantly to the realization of practical quantum internet infrastructure.

Acknowledgment

This thesis is more than just a collection of chapters and data, but a testament to a journey marked by growth, discovery, and the profound generosity of many. As I reflect on this past year, my heart swells with gratitude for the incredible individuals and institutions whose belief, guidance, and unwavering support transformed a challenging academic pursuit into a deeply rewarding experience.

At the foundation of this achievement stands my esteemed supervisor, Professor Yuto Lim. From the flicker of an initial idea to the comprehensive analysis presented within these pages, Professor Lim has been far more than an advisor; he has been my academic compass, charting the course and illuminating every crucial step of my research. His profound insights, tireless dedication, and a unique ability to foster independent thought have been the defining influences on my work. The vibrant intellectual ecosystem within the Lim Lab, nurtured by his vision, has not only honed my scientific skills but also instilled a deeper appreciation for the collaborative spirit of research.

My sincere appreciation extends to other key mentors who broadened my horizons. I am deeply thankful for the insightful guidance of Professor Ikeda Kokolo, my minor research project supervisor, whose expertise provided crucial perspectives that enriched the early stages of my research.

The journey would have been far less enriching without the constant support and camaraderie of my labmates at Lim Lab. Their spirited discussions, shared challenges, and genuine encouragement created an invaluable sense of community. We navigated the complexities of research together, and for their friendship and intellectual companionship, I am truly grateful.

The comprehensive support from JAIST was instrumental in allowing me to fully immerse myself in my studies. I am deeply indebted for the access to cutting-edge facilities, essential resources, and vital funding opportunities. Beyond the tangible, JAIST's efficient administrative systems, from seamless documentation assistance to expert consultancy services, provided a crucial backbone, ensuring that my focus remained firmly on my academic pursuits.

On a deeply personal note, this entire endeavor was buoyed by the unwavering love and belief of my family. To my incredible mother, my supportive little brother, and my steadfast father – your unconditional love, your patience during demanding periods, and your immense trust from afar were the quiet strength that propelled me forward. Knowing you were there, holding me in your hearts, was my constant motivation.

The vibrant tapestry of friendships woven during my time here has also been a profound source of strength. My heartfelt thanks go to the Vietnamese community and the broader international community at JAIST, especially my cherished group of Vietnamese friends. You provided a welcoming haven, a home away from home, filled with understanding, laughter, and shared experiences that made my time in Japan truly special.

Finally, a heartfelt thank you goes to my friend - Nguyen Thanh Canh. Even though he wasn't in the same lab as me, his generosity extended far beyond just lending his computer during crucial months, his willingness to freely offer his time in helping me and expertise was an invaluable practical help that smoothed many technical hurdles and kept my work flowing.

To every individual who has touched this journey, whether through profound guidance or simple acts of kindness, thank you. This thesis is a testament not only to my efforts but to the collective support that made it possible.

This thesis was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Posts and Telecommunications Institute of Technology.

List of Abbreviations and Symbols

List of Symbols

α, β	Complex probability amplitudes
χ	Purity of a quantum state
$\hat{\rho}_{XY}(t)$	Estimated number of maximally entangled states
$\mathbb{C}$	Set of complex numbers
μ	Average fidelity decay baseline
ρ	Density matrix
τ'_{BSM}	Time taken to perform BSM at quantum repeater
τ_c	Memory coherence time in NPP protocol
τ_{BSM}	Time taken to perform BSM at elementary link center
τ_{e2e}	End-to-end latency in NPP protocol
τ_{elem}	Time for an elementary link operation
τ_{tot}	Total latency in NPP protocol
dist_{ij}	Euclidean distance between nodes i and j
c	Speed of light in fiber
d	Distance between two nodes
D_0	Number of unused memory links to be entangled
d_{ij}	Scaled physical distance between nodes i and j (km)
$E_C(\cdot)$	Entanglement cost
F	Fidelity of a quantum state

f	Fidelity value of an entangled link
F'	Normalized fidelity to the maximum of an entangle link
$f_{\max}$	Maximum observed fidelity
k	Order of the NPP protocol
L	End-to-end connection length (km)
l	End-to-end latency for a path
L'	Normalized latency to the minimum of an entangle link
L_0	Length of an elementary link (km)
$l_{\min}$	Minimum latency observed across all available links
L_{att}	Photon attenuation length
L_{dec}	Decoherence length
M	Total number of cycles
m	Index for cycles
n	Nesting level for NPP protocol
$p(t)$	Conditional fidelity in NPP protocol
P_M	Maximum probability of conclusive success for a given BSM
P_S	Success probability
Q_n	Steady-state rate of entanglement generation per ideal memory in n th-order NPP
Q_{FiLa}	Composite link metric for fidelity and latency
R_n	Achievable rate (of entanglement generation) in NPP protocol
s	Linear scale factor for distance conversion
$t(d)$	Transmission time
$t(E)$	Entanglement generation time

$t(N)$	Quantum memory storage time
T_{ED}	Fundamental time period for entanglement distribution attempts
T_{NPP}	Total time taken in NPP protocol
w_f	Application-defined weight for fidelity
w_l	Application-defined weight for latency
x_i, y_i	2D coordinates of node i
x_j, y_j	2D coordinates of node j

List of Abbreviations

2EDP	Two-way Entanglement Distillation Protocol
BSM	Bell-State Measurements
DQC	Distributed Quantum Computing
HEG	Heralded Entanglement Generation
HEGP	Entanglement Generation Protocol
HEP	Heralded Entanglement Purification
JPNM	Japan Photonic Network Model
NPP	Nested Purification Protocol
Q-LEAP	Quantum Low-complexity routing Algorithm
QCS	Quantum Clock Synchronization
QEC	Quantum Error Correction
QKD	Quantum Key Distribution
QT	Quantum Teleportation

List of Figures

1.1	Comparison of classical and quantum routing logic. Classical routing selects the shortest or lowest-latency path, while quantum routing must consider fidelity, purification, and resource availability.	2
2.1	Illustration of a basic quantum key distribution (QKD) setup.	11
2.2	Visualization of the BB84 protocol.	12
2.3	Comparison between classical bits and quantum bits (qubits).	15
2.4	Bloch sphere illustrating the computational z -basis ($ 0\rangle, 1\rangle$), and the orthogonal x - and y -basis states. Any pure qubit state corresponds to a point on the surface of the sphere, parameterized by angles θ and ϕ	16
2.5	Illustration of entangled photon pairs.	17
2.6	Steps in entanglement swapping using three stations. The intermediate station (Station 1) performs a Bell State Measurement to extend entanglement between Station 0 and Station 2.	19
2.7	Two-round entanglement purification protocol.	20
2.8	Illustration of the fidelity–latency tradeoff.	23
2.9	Workflow of the Q-LEAP routing algorithm.	29
3.1	Illustration of the fidelity–latency tradeoff.	32
3.2	Trade-off between fidelity and latency in quantum entanglement routing. Longer paths with purification improve fidelity but increase latency, while shorter paths reduce latency but may fall below the fidelity threshold.	33
3.3	Illustration of the nested purification protocol (NPP) over the chain of quantum repeater in quantum networks	34
3.4	Performance of achievable rate versus distance with (a) Different maximum success probability; and (b) Different memory coherence time	39
3.5	Performance of entanglement generation rate versus (a) Distance and (b) Nesting level with different maximum success probability	40
3.6	Performance of success probability versus distance with different nesting levels	41

3.7	Performance of latency versus distance with different maximum success probability	41
3.8	Performance of latency versus nested level with different maximum success probability	42
3.9	Performance of conditional fidelity versus the time period with different memory coherence time	43
3.10	Performance of latency versus distance with different entanglement success probability (p_s)	46
3.11	Q_{FiLa} Choice in Quantum Network Topology	50
4.1	Average link distance vs. scale factor in a 50-node k -nearest neighbor topology with $k = 3$ under Q-LEAP with Q_{FiLa} ($w_f = 0.5, w_l = 0.5$).	56
4.2	Average latency vs. fidelity threshold in random topologies (25–100 nodes): (a) Original Q-LEAP; (b) Q-LEAP with $Q_{FiLa}, w_f = 0.5, w_l = 0.5$	58
4.3	Path computation time across fidelity thresholds in random topologies of 25–100 nodes: (a) Original Q-LEAP; (b) Q-LEAP with Q_{FiLa} ($w_f = 0.5, w_l = 0.5$).	60
4.4	Total throughput vs. fidelity threshold in random topologies (25–100 nodes): (a) Original Q-LEAP; (b) Q-LEAP with $Q_{FiLa}, w_f = 0.5, w_l = 0.5$	61
4.5	Visualization based on JPNM: (a) Fidelity vs. distance and (b) Fidelity vs. latency.	63
4.6	Illustration of the Japan Photonic Network Model (JPNM) [1].	64
4.7	Distribution of the 82 quantum links in JPNM: (a) Link distances and (b) Fidelity.	66
4.8	Latency Distribution of the 82 quantum links in JPNM.	67
4.9	Performance comparison on JPNM under different fidelity thresholds: (a) Total throughput and (b) Average latency. . .	68
4.10	Path computation time performance comparison on JPNM under different fidelity thresholds.	69

List of Tables

1.1	Application Viewpoints and Trade-offs	4
3.1	Simulation Parameters and Settings for NPP Protocols	38
4.1	Distance Statistics Across Scale Factors	55
4.2	Simulation Parameters for Random Topologies	57
4.3	Simulation Parameters for JPNM Topology	65

Contents

Abstract	I
Acknowledgment	V
Chapter 1 Introduction	1
1.1 Overview	1
1.2 Motivation And Objectives	4
1.3 Thesis Approach	6
1.4 Thesis Contributions	8
1.5 Thesis Structure	9
Chapter 2 Background and Related Work	11
2.1 Background	11
2.1.1 Quantum Communication and Network Foundations . .	11
2.1.2 Qubit Representation and Basis States	14
2.1.3 Quantum Entanglement and Applications	16
2.1.4 Entanglement Distribution Techniques	18
2.1.5 Fidelity and Latency in Quantum Networks	23
2.2 Related Work	26
2.2.1 Toward Joint Optimization of Fidelity and Latency in Quantum Routing	26
2.2.2 Fidelity-Guaranteed Routing: Q-LEAP	28
2.3 Summary	30
Chapter 3 Novel Quantum Link Metric Design	31
3.1 Problem Statement	31
3.2 Latency Analysis of Entanglement Distribution Protocol . . .	33
3.2.1 Latency in Nested Purification Protocol	33
3.2.2 Latency in Entanglement-guaranteed Distribution Pro- tocol	43
3.3 Proposed Quantum Link Metric: Q_{FiLa}	46
3.3.1 Fidelity Model	47
3.3.2 Q-LEAP Overview	47
3.4 Summary	50

Chapter 4	Evaluation Studies and Discussions	52
4.1	Evaluation Overview	52
4.1.1	Simulation Setup and Parameters	52
4.1.2	Experiment Scenarios	53
4.2	Experiment Results	55
4.2.1	Scenario 1: Link Density Sensitivity Analysis	55
4.2.2	Scenario 2: Performance of Q-LEAP with and without Q_{FiLa} in Random Topologies	57
4.2.3	Scenario 3: Fidelity-Latency Trade-off Visualization	62
4.2.4	Scenario 4: Performance of Q-LEAP with and without Q_{FiLa} in JPNM	64
4.3	Summary	70
Chapter 5	Conclusions and Outlook	71
5.1	Summary of Findings	71
5.2	Study Limitations	72
5.3	Directions for Future Work	73
	List of Publications	75
	References	76

Chapter 1

Introduction

1.1 Overview

Quantum networks hold the potential to revolutionize ultra-secure communication by harnessing the fundamental principles of quantum mechanics. These networks are poised to enable a diverse array of security-critical applications, including quantum key distribution (QKD) [2], secure authentication [3], quantum clock synchronization (QCS) [4], etc. and the distribution of entangled states for advanced cryptographic tasks.

However, a significant challenge in quantum networking, which has garnered considerable research attention, is the reliable distribution of entanglement over long distances [5]. This endeavor is inherently constrained by factors such as quantum decoherence, optical channel loss, and the probabilistic nature of entanglement generation. Unlike their classical routing that is illustrated in Figure 1.1, quantum systems cannot employ conventional signal amplification or replication techniques due to the no-cloning theorem [6]. This fundamental limitation necessitates the development of entirely new strategies for routing and resource management within quantum network architectures.

To overcome these limitations, quantum repeaters are considered as a fundamental solution to long-distance quantum communication [7]. They elevate long-distance entanglement by performing entanglement swapping and purification operations across intermediate nodes. In such a repeater chain, adjacent nodes first generate Bell pairs, which are then connected via entanglement swapping to form a longer-distance entangled state. However, each operation introduces potential degradation of fidelity, which quantifies how closely the shared entangled state resembles an ideal Bell state [8].

Traditional entanglement routing approaches often focus on optimizing a single metric—typically either throughput fidelity. For instance, the Q-LEAP algorithm [9] ensures fidelity guarantees by adaptively applying purification based on a predefined fidelity threshold. However, this often leads to longer path lengths or more purification steps, thereby increasing latency in the

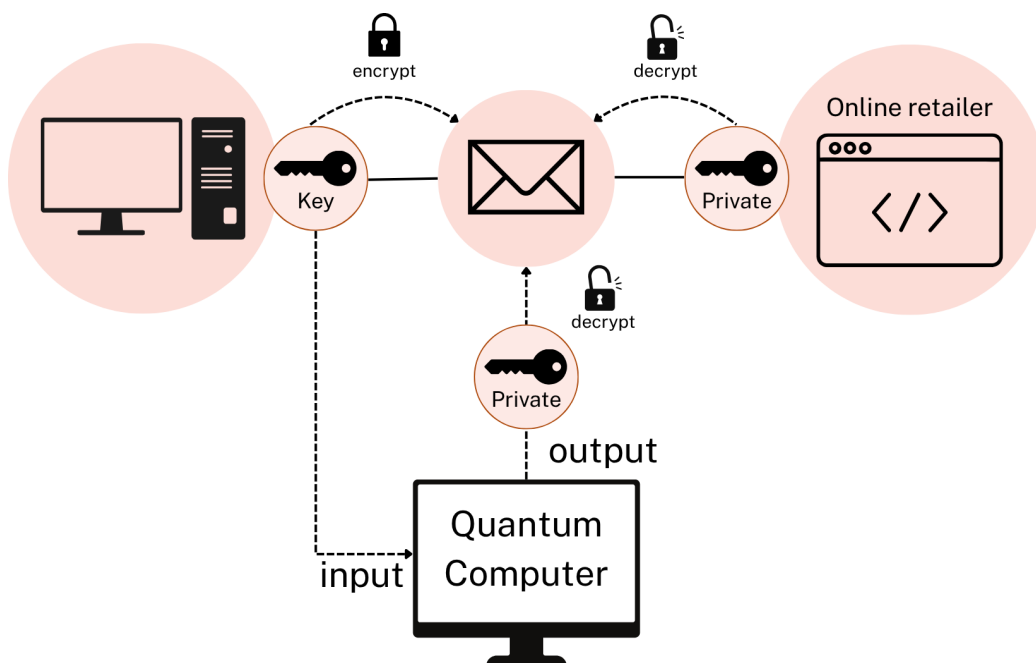


Figure 1.1: Comparison of classical and quantum routing logic. Classical routing selects the shortest or lowest-latency path, while quantum routing must consider fidelity, purification, and resource availability.

entire routing. In contrast, latency-sensitive applications such as quantum key distribution (QKD) or distributed quantum computing (DQC) require timely delivery of entangled states, making routing decisions based solely on fidelity suboptimal in practice.

In the need for a more adaptable approach to routing in long-distance quantum networks. This thesis will introduce a novel quantum link metric— Q_{FiLa} (Quantum Fidelity-Latency)—which will integrate both fidelity and latency into a unified routing framework. The Q_{FiLa} metric enables dynamic path selection based on weighted combinations of normalized to the maximum fidelity and minimum latency values, thereby explicitly capturing the trade-off between these two competing metrics. This tradeoff is central to quantum networking, where higher fidelity often requires greater latency due to longer paths or additional purification steps, while low-latency paths may lack sufficient fidelity for reliable quantum communication.

Q_{FiLa} extends the Q-LEAP framework by incorporating a fidelity–latency trade-off into the routing decision process. By dynamically adjusting the balance between fidelity and latency, Q_{FiLa} enables the selection of entanglement paths that better align with application-specific performance requirements. Through simulations on realistic quantum network topologies, such as the Japan Photonic Network Model (JPNM), our evaluations show that the Q_{FiLa} metric provides a stable, adaptive, and resource-efficient approach to entanglement routing, particularly under diverse application constraints and varying network conditions.

Table 1.1 summarizes key quantum applications and highlights their differing sensitivity to latency and fidelity, thereby motivating the need for an adaptive routing metric such as Q_{FiLa} .

This thesis focuses on the joint optimization of fidelity and latency in quantum entanglement routing, and introduces a scalable, simulation-validated link metrics designed to efficiently select routing paths under varying fidelity thresholds. The proposed methodology not only addresses the trade-off between fidelity and latency, but also lays the foundation for incorporating additional performance metrics in future research. Beyond fidelity and latency, other relevant factors such as the number of hops, entanglement generation success probability, purification overhead, throughput, and memory storage time can also influence routing decisions and network performance. By exploring these dimensions, this work contributes to the broader goal of advancing practical and efficient quantum internet infrastructure.

Table 1.1: Application Viewpoints and Trade-offs

Quantum Application	Latency Priority	Fidelity Priority	Trade-off Insight
Quantum Key Distribution (QKD)	Medium–High	High	Too low fidelity leads to insecure keys; too much latency reduces throughput
Quantum Teleportation (real-time)	High	High	Must coordinate classical and quantum layers; fidelity loss ruins transferred state
Entanglement Swapping in Repeaters	Medium	High	Swapping too early causes fidelity loss; waiting too long may exceed memory lifetime
Quantum Sensing (real-time MRI/MEG)	High	Medium	Requires tight timing; fidelity can sometimes be compensated by classical processing
Quantum Machine Learning for Trading	High (if real-time)	High	Low latency models must still learn with high-quality qubits/circuits
Quantum Secure Messaging (QKD + encryption)	High	High	Critical to balance low key delivery delay with strong error correction

1.2 Motivation And Objectives

Routing in quantum networks involves selecting routing paths that satisfy application-level requirements for fidelity, timing, and resource usage. The core purpose of this task lies the challenge of distributing high-quality entangled states across a network where quantum operations are inherently noisy,

and physical constraints like channel loss, decoherence, and probabilistic entanglement generation significantly impact performance. In practice, the quality of entanglement (fidelity) degrades over long distances or after many quantum operations, while purification to improve fidelity introduces latency overhead and consumes additional network resources.

This fidelity-latency tradeoff is not widely recognized in recent literature yet. For example, the Q-LEAP algorithm [9] provides fidelity-guaranteed routing by dynamically triggering purification, yet it often results in high latency due to path selection that prioritizes fidelity alone. Conversely, low-latency routing paths may bypass purification, leading to entangled states that fall below fidelity thresholds. Hence, a core problem in quantum routing lies in optimizing fidelity and latency jointly, rather than independently.

This thesis is motivated by the limitations of single-metric quantum routing strategies and aims to design a fidelity-latency tradeoff aware framework that balances quantum state quality and delivery efficiency. Our goal is to support flexible routing decisions based on the application’s needs, which may prioritize latency (e.g., real-time QKD) or fidelity (e.g., teleportation for quantum computing), or both in different application needs and network conditions.

Due to the existing routing approaches lack of adaptability, also from classical routing typically focuses on shortest path or least-latency [10–12]. While this is straightforward in classical networks, it is inadequate for quantum systems that require high-quality entangled states for correctness. The no-cloning theorem and fragile nature of quantum information imply that fidelity cannot be ignored [13, 14]. Moreover, dynamic network conditions such as link failures, channel fidelity fluctuations, or memory decoherence demand routing strategies that are context-aware and application-driven.

To address the challenges of optimizing both fidelity and latency in quantum networks, this thesis introduces the Q_{FiLa} (Quantum Fidelity-Latency) link metric. This composite routing metric enables tunable routing behavior by adjusting fidelity and latency weights, supporting multi-objective optimization across diverse network conditions. The main objectives of this research are as follows:

The aim of this thesis is to propose an innovative approach for quantum entanglement routing in quantum networks by addressing the tradeoffs between fidelity and latency. The objectives of this research are as follows:

- **To develop a quantum link metric:** This thesis introduces the Q_{FiLa} metric, which integrates both fidelity and latency into a unified framework. This metric is designed to support dynamic routing decisions that can balance quality (fidelity) and delay time (latency)

based on the specific requirements of different quantum applications. The goal is to enable the routing algorithm to flexibly manage these competing objectives, allowing for optimization across diverse network conditions.

- **To support dynamic tradeoffs based on application needs:** One of the key objectives is to create a routing protocol that supports dynamic tradeoffs, allowing quantum networks to adjust based on the specific demands of an application. For example, while banking and security applications may prioritize fidelity, time-sensitive applications like real-time quantum computing may prioritize latency. The proposed solution will allow the network to shift between these priorities based on real-time conditions, optimizing the performance of the quantum network to meet varying requirements.
- **To extend quantum network design with additional resource parameters:** This thesis will also explore additional parameters, such as coherence time and the number of hops, which will allow for more nuanced and efficient routing decisions. The aim is to extend beyond traditional routing protocols, which often focus solely on fidelity, and instead consider multiple factors for a more robust, resource-efficient, and scalable quantum network.

1.3 Thesis Approach

To achieve these objectives, this thesis aims to propose a research approach to address the trade-off between fidelity and latency in entanglement routing for long-distance quantum networks. Our methodology centers on extending Q-LEAP, an existing fidelity-guaranteed routing framework by introducing a novel multiobjective link metric, and evaluating its performance through comprehensive simulation studies.

Our research builds directly by improving the principles of the Q-LEAP algorithm [9]. Although the original routing algorithm effectively minimizes resource consumption and ensures a predefined fidelity threshold through adaptive purification and an extended Dijkstra algorithm [9], it does not integrate other crucial factors like the number of hops, entanglement generation success probability, latency,...and so on. For Q-LEAP, they have not considered the latency, which makes the routing path high in fidelity but not so good in latency due to the complex routing processes. Our approach aims to extend this foundational framework by introducing latency as a critical, explicit optimization metric, thereby enabling a more comprehensive and practical balance between these two competing objectives.

A central component of our thesis is the design and implementation of a state-of-the-art quantum fidelity-latency Q_{FiLa} link metric. This approach focuses on capturing the overall quality of a quantum link by concurrently considering both its fidelity and latency characteristics. Therefore, its design integrates fidelity modeling, where we model the fidelity (f) of an entangled link as an exponential decay function of physical distance (d), accounting for photon loss and decoherence effects. We also include latency modeling, defining link latency (l) to encompass entanglement generation time ($t(E)$), physical propagation delay ($t(d)$), and quantum storage time ($t(N)$), ensuring a realistic representation of end-to-end entanglement establishment. Furthermore, we apply a normalization process to both fidelity and latency values, scaling them into comparable ranges, which ensures the routing algorithm can make balanced decisions, irrespective of their differing physical units and scales. Finally, we combine the normalized to the maximum fidelity (F') and normalized to the minimum latency (L') using a weighted sum, enabling flexible optimization strategies tailored to specific application requirements.

In addition, we integrate the proposed Q_{FiLa} link metric into Q-LEAP, treat it as an enhanced routing algorithm, developed as a modified version of the original algorithm. This modified algorithm utilizes the Q_{FiLa} metric during its path-search phase, allowing the identification of paths that offer an optimal trade-off between fidelity and latency. The algorithm maintains Q-LEAP's computational efficiency while enabling multi-objective optimization. Furthermore, the purification decision process is adapted to align with these dual objectives, influencing the number of purification rounds based on both desired fidelity and acceptable latency.

Subsequently, to evaluate scalability and adaptability, we begin by applying our method to randomly generated network topologies ranging from 25 to 100 nodes. These synthetic topologies allow us to examine routing behavior across diverse network densities and link configurations. Following this large-scale evaluation, we transition to a more realistic, smaller-scale topology based on the Japan Photonic Network Model (JPNM), which includes 48 nodes and 82 links with physical distances varying from 10 km to 673 km. This real-world topology enables simulation of both metropolitan and long-distance quantum communication scenarios. The routing performance will be access under a range of fidelity thresholds (e.g., 0.55 to 0.95) to observe how the algorithm responds to different quality-of-service requirements. Key performance metrics include total throughput, average latency, and path computation time, offering a comprehensive view of the algorithm's ability to balance fidelity and latency.

Finally, we benchmark the proposed Q_{FiLa} -based approach against the original Q-LEAP algorithm and other routing strategies that optimize single

objectives. This comparative analysis highlights the effectiveness of joint fidelity-latency optimization and provides insights into the tradeoffs involved. Overall, our simulation methodology ensures a systematic investigation into the complex interplay between fidelity and latency in quantum routing, contributing to the development of practical and efficient solutions for future quantum internet infrastructure.

1.4 Thesis Contributions

This thesis contributes to the development of a novel composite routing metric, Q_{FiLa} , that addresses the critical trade-off between entanglement fidelity and latency in quantum networks. Through the integration of both fidelity and latency considerations into the existing Q-LEAP framework, this work optimizes the quality and efficiency of entanglement distribution in quantum communication systems. The goal is to provide a flexible and scalable routing solution, enabling the deployment of large-scale quantum networks while maintaining both high fidelity and low latency.

The key contributions of this thesis are outlined as follows:

- **Proposing a Novel Quantum Fidelity-Latency Link Metric:** We introduce the Q_{FiLa} link metric, a comprehensive solution that balances the competing objectives of maximizing fidelity and minimizing latency in quantum routing. By incorporating realistic models of fidelity degradation (due to distance, photon loss, and decoherence) and latency accumulation (due to entanglement generation, propagation, and storage), the metric is designed to accommodate varying application requirements through robust normalization and a tunable weighted combination approach.
- **Finding the Balance for Application-Specific Needs:** A crucial contribution of this work is the ability to dynamically adjust the balance between fidelity and latency based on the needs of specific quantum applications. For instance, applications such as QKD require low latency to maintain real-time security, which calls for higher weight on latency (w_l) and lower weight on fidelity (w_f). In contrast, distributed quantum computing may prioritize high fidelity to ensure the accuracy of quantum states over the need for minimal latency. This research shows how the Q_{FiLa} metric's tunable weights can be tailored to meet these different needs, enabling a more adaptable and application-aware quantum routing protocol. By incorporating additional network parameters like memory coherence time, number of hops, and entanglement success probability, this approach goes beyond

traditional single-objective routing, offering a multi-dimensional trade-off that supports the diverse and evolving requirements of future quantum networks.

1.5 Thesis Structure

This thesis is organized into five chapters, each focusing on a critical aspect of quantum routing in quantum networks, with particular emphasis on the proposed Q_{FiLa} link metric. Below is an overview of each chapter:

Chapter 1: Introduction: This chapter provides an introduction to the research, outlining the key challenges in quantum entanglement distribution within quantum networks. It discusses the motivation behind this work, followed by the objectives and approach taken to address limitations in existing quantum routing strategies. The chapter highlights the novel contributions of the thesis, including the development of the Q_{FiLa} metric, and concludes with an overview of the structure of the thesis.

Chapter 2: Background and Related Work: This chapter offers a comprehensive background on quantum communication, including the principles of quantum networks, entanglement, and the critical role of quantum repeaters. It delves into quantum network models, the representation of qubits, and quantum entanglement applications such as quantum key distribution (QKD) and distributed quantum computing. The related work section reviews existing research in quantum routing algorithms, with a particular focus on Q-LEAP and other strategies, discussing their strengths and limitations in optimizing fidelity and minimizing latency.

Chapter 3: Novel Quantum Link Metric Design: Chapter 3 introduces the primary focus of the thesis—the Q_{FiLa} metric. It provides the problem statement and outlines the methodology used to develop a novel composite link metric that integrates both fidelity and latency for quantum routing. The chapter details the formulation of the Q_{FiLa} metric, the modeling of fidelity and latency in quantum links, and the network topologies used for evaluation, including the Japan Photonic Network Model (JPNM) and synthetic random topologies. It further discusses the application of the metric in adaptive routing algorithms.

Chapter 4: Evaluation and Discussion: This chapter presents an in-depth evaluation of the Q_{FiLa} metric, comparing its performance using the Q_{FiLa} to the Q-LEAP algorithm. It includes simulations on both real-world and synthetic quantum network topologies, covering various scenarios such as the fidelity-latency trade-off distributions, Q_{FiLa} in general cases like the random topology to a real topology like the Japan Photonic Network Model

(JPNM) analysis. The chapter also addresses link density sensitivity and explores the impact of network scalability on routing performance.

Chapter 5: Conclusions and Outlook: The final chapter summarizes the key findings and contributions of the research. It discusses the limitations of the study, including assumptions made and constraints faced, and suggests directions for future work. The chapter highlights the importance of integrating advanced quantum functionalities, such as quantum error correction and entanglement purification, into future routing strategies. It also emphasizes the need for further studies on large-scale network testing and dynamic routing algorithms.

Chapter 2

Background and Related Work

2.1 Background

2.1.1 Quantum Communication and Network Foundations

The advent of quantum information technology marks a major shift in the landscape of communication systems, leading to the emergence of the quantum Internet. Unlike classical networks, which transmit classical bits through physical copying and amplification, quantum networks exploit unique quantum mechanical principles—such as superposition, entanglement, and quantum teleportation—to exchange quantum states known as *qubits* [6]. These qubits, unlike classical bits, can exist in a superposition of multiple states [15], enabling fundamentally new paradigms for secure communication and distributed computation.

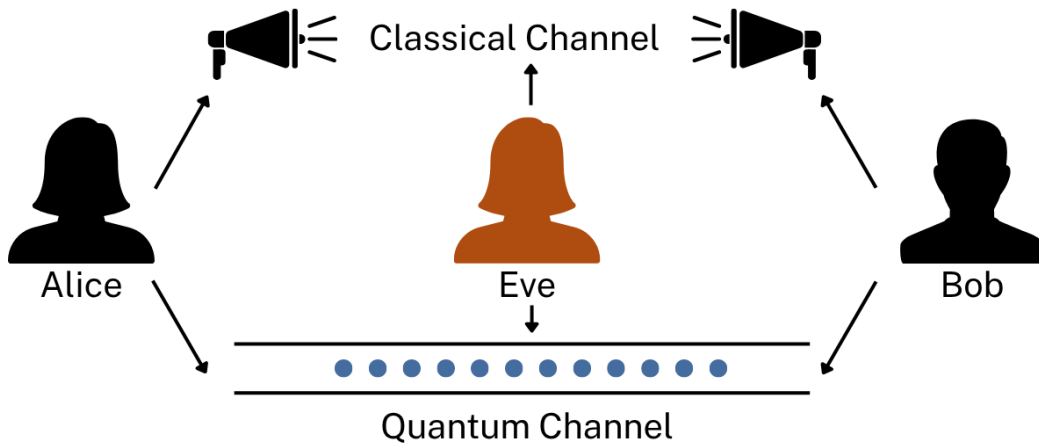


Figure 2.1: Illustration of a basic quantum key distribution (QKD) setup.

A wide range of applications in quantum networks rely on entanglement as a fundamental resource [16–18]. Among these, one of the most well-

established and practically pursued is *quantum key distribution* (QKD), which allows two parties to generate a shared secret key with security rooted in the laws of quantum mechanics [19, 20]. In QKD protocols, any attempt by an eavesdropper to intercept transmitted quantum states unavoidably disturbs them due to the no-cloning theorem [13], making intrusion detectable to the legitimate parties. Figure 2.1 illustrates two legitimate parties, Alice and Bob, communicate over a classical public channel and a quantum channel. An eavesdropper, Eve, is assumed to have full access to both channels. Any attempt by Eve to intercept or measure the quantum states unavoidably introduces disturbances, which Alice and Bob can detect by comparing subsets of their measurement outcomes.

The BB84 protocol, proposed by Bennett and Brassard in 1984 [19], is the earliest and most widely adopted QKD scheme. In this protocol showed in Figure 2.2, Alice generates a random bit sequence and independently selects a basis—either rectilinear (+) or diagonal ($\times$)—to encode each bit into the polarization of a photon. Bob, upon receiving each photon, also randomly selects a basis to measure it. If Bob’s measurement basis matches Alice’s, he obtains the correct bit value; otherwise, the outcome is random.

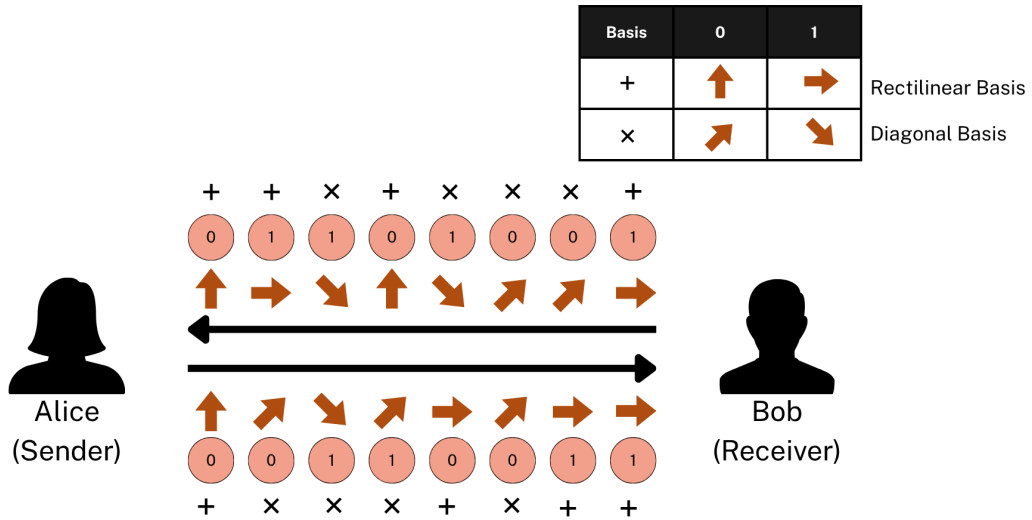


Figure 2.2: Visualization of the BB84 protocol.

After the quantum transmission phase, Alice and Bob publicly announce their basis choices over the classical channel and discard all bits where the bases differ. The remaining bits, corresponding to matching bases, form the *sifted key*. To verify security, a subset of the sifted key is compared publicly. If the error rate exceeds a threshold, the presence of an eavesdropper

is inferred. Due to the no-cloning theorem and Heisenberg’s uncertainty principle, Eve cannot measure or replicate the qubits without introducing detectable disturbances. This inherent sensitivity to measurement gives BB84 its unconditional security guarantees.

Quantum networks enable additional applications such as quantum teleportation, distributed quantum computing, and clock synchronization. Quantum teleportation allows for the transfer of an unknown quantum state between distant parties using entanglement and classical communication [21]. Distributed quantum computing relies on shared entanglement among separate quantum processors to jointly execute computational tasks beyond the capacity of a single node [22]. Similarly, entangled particles can be used to synchronize clocks across large distances with quantum-level precision, offering advantages for navigation, metrology, and scientific experimentation [23].

The core function of a quantum network is to establish remote entanglement between arbitrary nodes. This is often achieved via multi-hop connections involving intermediate nodes known as quantum repeaters, which help mitigate the exponential decay of entanglement fidelity over long distances [22]. Unlike classical repeaters, which amplify signals, quantum repeaters employ techniques such as entanglement swapping and purification. Entanglement swapping enables two end nodes to become entangled without direct interaction by connecting adjacent entangled pairs through intermediate nodes. Purification, on the other hand, improves fidelity by combining several low-quality entangled pairs to produce a higher-fidelity one [9].

Nonetheless, constructing large-scale quantum networks remains a challenge. Quantum states are inherently fragile and prone to errors due to photon loss, decoherence, and imperfect operations [22]. Additionally, entanglement generation is a probabilistic process, and the fidelity of entangled pairs degrades significantly with distance [6]. Unlike classical networks, where signals can be freely amplified or copied, quantum information cannot be cloned due to the no-cloning theorem [13]. Thus, quantum routing must be designed to work within these physical limitations while managing resource consumption efficiently.

In summary, the foundational properties of quantum communication systems—entanglement, no-cloning, and probabilistic operations—necessitate new paradigms in network design, error handling, and routing. The goal of quantum networking is not simply to transfer data, but to deliver high-fidelity entangled states between distributed users on demand. Addressing these challenges is essential to unlock the full potential of the quantum Internet.

2.1.2 Qubit Representation and Basis States

A qubit (quantum bit) is the fundamental unit of quantum information and serves as the quantum analogue of a classical bit. While a classical bit can exist in only one of two deterministic states, 0 or 1, a qubit can exist in a linear superposition of both basis states simultaneously [6]. This unique property enables richer computational possibilities in quantum systems.

Mathematically, a general qubit state is described using Dirac (bra-ket) notation [24] as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad (2.1)$$

where $\alpha, \beta \in \mathbb{C}$ are complex probability amplitudes satisfying the normalization condition:

$$|\alpha|^2 + |\beta|^2 = 1. \quad (2.2)$$

Upon measurement, the qubits collapses to either $|0\rangle$ or $|1\rangle$ with probabilities $|\alpha|^2$ and $|\beta|^2$, respectively. This probabilistic nature, governed by the Born rule, distinguishes quantum systems from classical deterministic ones [6].

As illustrated in Figure 2.3, a classical bit can be interpreted as a binary switch, whereas a qubit can encode both 0 and 1 simultaneously in a complex weighted manner. The figure also depicts intuitive metaphors (such as colored beads and polarizations) to visualize superposition.

Dirac notation, also known as bra-ket notation, provides a compact and expressive mathematical framework for describing quantum states. A *ket* $|\psi\rangle$ denotes a column vector in a complex Hilbert space, while its conjugate transpose—called a *bra*—is denoted $\langle\psi|$. Together, these notations allow elegant formulation of inner products, projections, and operator actions.

The computational basis of a qubit consists of two orthonormal states:

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \quad (2.3)$$

In this basis, the qubits $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ can be represented as a 2D column vector:

$$|\psi\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \quad (2.4)$$

The corresponding bra vector is defined as:

$$\langle\psi| = [\alpha^* \quad \beta^*], \quad (2.5)$$

where α^* and β^* are the complex conjugates of the amplitudes.

The inner product $\langle\psi|\psi\rangle$ of a qubit state with itself yields its norm squared:

$$\langle\psi|\psi\rangle = |\alpha|^2 + |\beta|^2 = 1. \quad (2.6)$$

This is known as the normalization condition and guarantees that the total probability of all possible outcomes sums to unity. It also ensures the qubits resides on the surface of the Bloch sphere, which represents all pure qubits states geometrically.

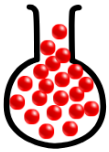
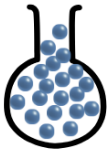
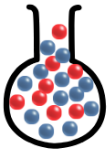
Classical bits		Quantum bits (Qubits)
Bit 1  Empty = “0”	Bit 2  Filled = “1”	Qubit 1  1/3 of “0” and 2/3 of “1”
 20 red beads = “0”	 20 blue beads = “1”	 8/20 of “0” and 12/20 of “1”
 Head = “0”	 Tail = “0”	 50% of chance of landing on “0” 50% of chance of landing on “1”

Figure 2.3: Comparison between classical bits and quantum bits (qubits).

This geometric picture (Fig. 2.4) provides intuitive insight into quantum operations and state evolution. For instance, rotations around the axes correspond to quantum gates, and antipodal points represent orthogonal states.

These foundations—Dirac notation, matrix representation, and inner products—are essential for the later treatment of entanglement, teleportation, and routing in quantum networks.

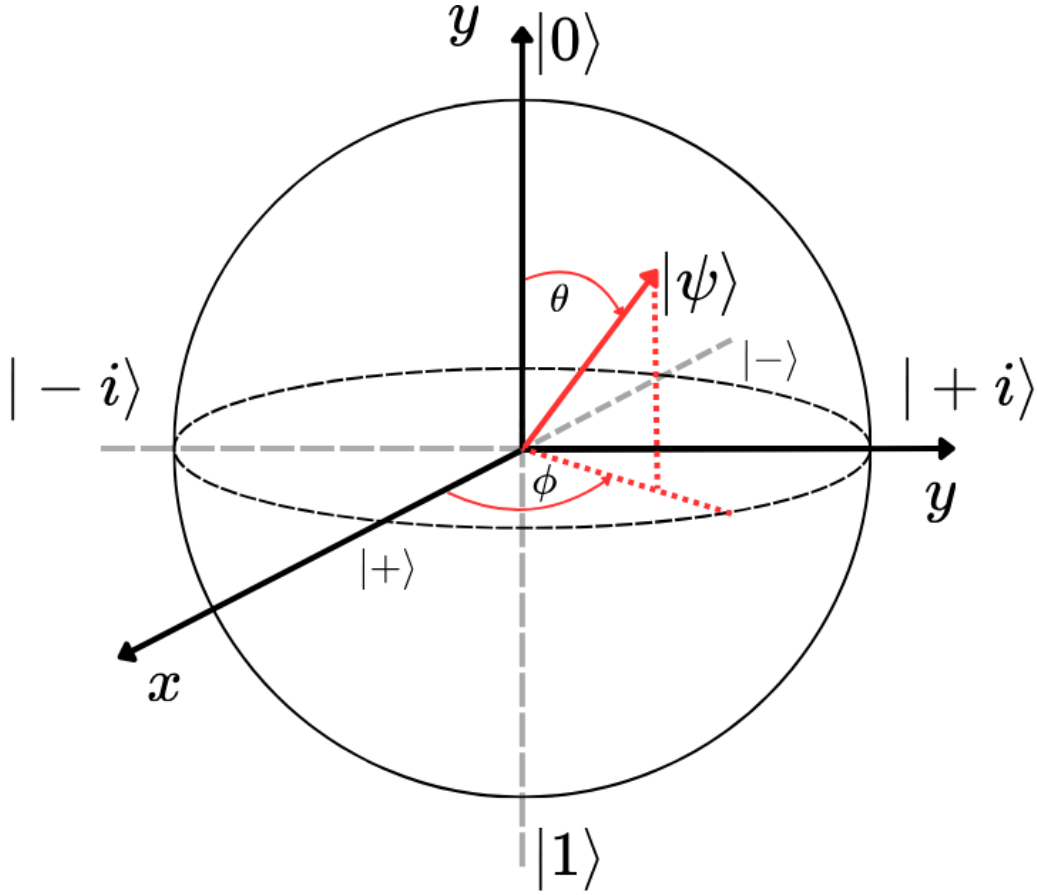


Figure 2.4: Bloch sphere illustrating the computational z -basis ($|0\rangle, |1\rangle$), and the orthogonal x - and y -basis states. Any pure qubit state corresponds to a point on the surface of the sphere, parameterized by angles θ and ϕ .

2.1.3 Quantum Entanglement and Applications

Quantum entanglement is a uniquely quantum phenomenon in which the states of two or more particles become so strongly correlated that the state of each particle cannot be described independently, regardless of the spatial separation between them. This nonlocal correlation enables powerful protocols in quantum communication and forms the basis of distributed quantum networking [6].

Entanglement also plays a key role in ensuring the security of the quantum key distribution (QKD). As described in Section 2.2, protocols such as BB84 [19] exploit the quantum properties of measurement and basis selection to detect eavesdropping. Although BB84 does not require explicit entanglement, its extensions—such as the E91 protocol [20]—leverage

entangled pairs to improve security guarantees.

To intuitively illustrate entanglement, consider Figure 2.5, which shows a pair of entangled photons shared between two parties. Although each photon's color (representing a quantum property such as polarization) appears uncertain prior to measurement, observing one photon immediately determines the state of the other, even if they are separated by vast distances. If one is observed as "red", the other must be "blue", and vice versa. This perfect correlation, independent of measurement order or location, cannot be explained by classical physics and is a hallmark of quantum entanglement.

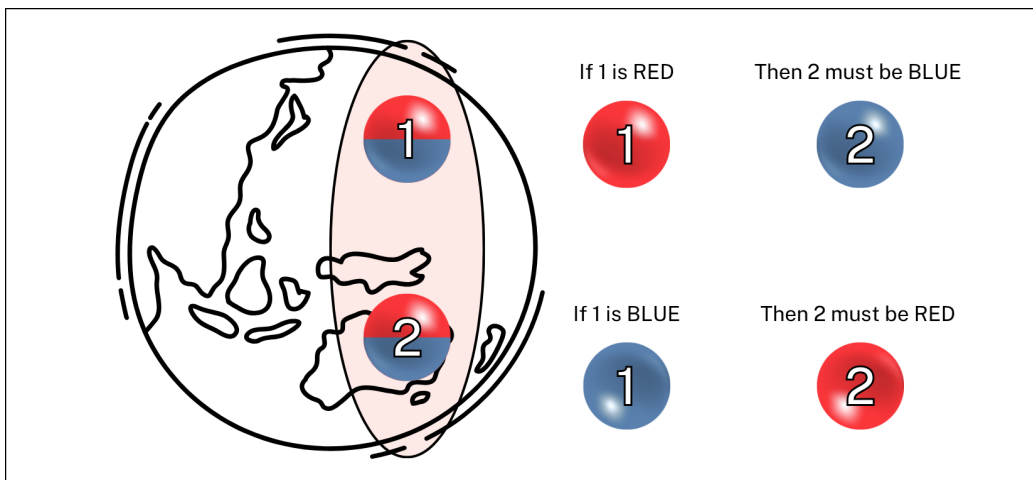


Figure 2.5: Illustration of entangled photon pairs.

Beyond QKD, entanglement enables several distributed quantum applications:

- **Quantum teleportation** [21]: Allows transmission of a quantum state between two nodes by consuming an entangled pair and sending classical information.
- **Clock synchronization** [23]: Utilizes shared entangled photons to synchronize distant atomic clocks with quantum-level accuracy.
- **Distributed quantum computing** [6]: Connects multiple quantum processors through entanglement to enable cooperative computation beyond individual node capacity.

These entanglement-enabled capabilities demonstrate the transformative potential of quantum networks. However, the fragility of entanglement over distance and the need for high-fidelity correlations present significant technical challenges—necessitating advanced techniques like purification and entanglement routing, which are the focus of subsequent chapters. of routing

design.

2.1.4 Entanglement Distribution Techniques

Establishing entanglement between distant nodes is essential for enabling large-scale quantum networks. However, entanglement generation over long distances is severely limited by channel losses and decoherence, which degrade the fidelity of entangled states. To overcome this challenge, quantum networks employ two fundamental techniques: *entanglement swapping* and *entanglement purification*. These operations are the building blocks of quantum repeaters, which enable a scalable distribution of entanglement across the network [6].

2.1.4.1 Entanglement Swapping

Quantum repeaters enable scalable entanglement distribution by dividing a long-distance link into shorter segments and then “stitching” them together using a process known as *entanglement swapping*. This process is central to extending entanglement over long distances in quantum networks, where direct transmission suffers from exponential photon loss [25].

Entanglement swapping enables two remote nodes, initially unentangled, to become entangled through the mediation of an intermediate station. This process is illustrated in Figure 2.6, where we consider three stations labeled Station 0, Station 1, and Station 2. Station 0 and Station 2 each hold one qubit, while Station 1 possesses two qubits, each entangled with the corresponding qubit at Station 0 and Station 2, respectively. If Station 0 is entangled with one of Station 1’s qubits, and Station 2 is entangled with the other, then Station 1 can perform a *Bell State Measurement* (BSM) on its two local qubits. This operation consumes the two initial entangled pairs and projects the outer qubits—held by Station 0 and Station 2—into an entangled state, despite the absence of any direct interaction between them.

The full sequence proceeds as follows: in Step 1, entangled pairs are established between Station 0 and Station 1, and separately between Station 1 and Station 2. In Step 2, Station 1 performs a BSM on its two local qubits. This measurement collapses the intermediate entanglement links, as shown in Step 3. Finally, Step 4 reveals the successful establishment of a new entangled pair between Station 0 and Station 2. By iterating this process across multiple repeater stations, entanglement can be reliably extended over long distances, forming the foundation for scalable quantum communication [6].

By enabling such long-range connectivity, entanglement swapping is essential for building large-scale quantum networks and is particularly useful in routing decisions where fidelity must be preserved while minimizing latency.

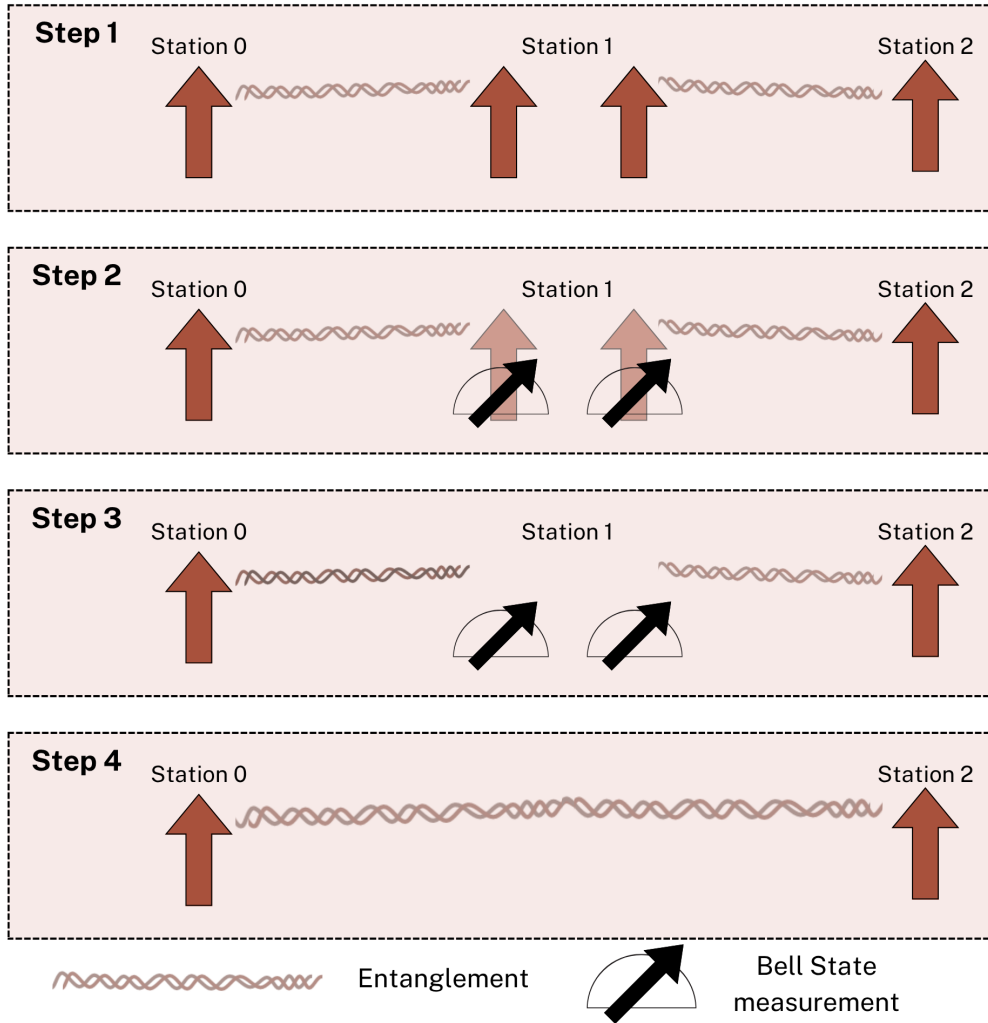


Figure 2.6: Steps in entanglement swapping using three stations. The intermediate station (Station 1) performs a Bell State Measurement to extend entanglement between Station 0 and Station 2.

2.1.4.2 Entanglement Purification

To counteract the degradation of entanglement quality due to noise and decoherence, *entanglement purification* is employed as a key technique to improve fidelity. This process consumes multiple low-fidelity entangled pairs

and, through a series of quantum operations and classical communication, probabilistically distills a smaller number of higher-fidelity pairs [26].

The purification protocol typically operates in multiple rounds. In each round, two or more identically prepared noisy entangled pairs are locally processed using entangling gates (e.g., CNOT) and measurement, followed by classical post-selection. Pairs yielding successful measurement outcomes are retained, while others are discarded. This iterative procedure continues until the remaining pair(s) reach the target fidelity threshold.

As illustrated in Figure 2.7, the first purification round begins with entangled pairs of fidelity $F = 0.8$. Two such pairs are combined, resulting in one purified pair with fidelity approximately $F = 0.8941$. This higher-quality pair is then used in a second purification round with another $F = 0.8$ pair, further boosting fidelity to $F = 0.9846$. This cascading approach highlights the probabilistic nature and resource demands of purification.

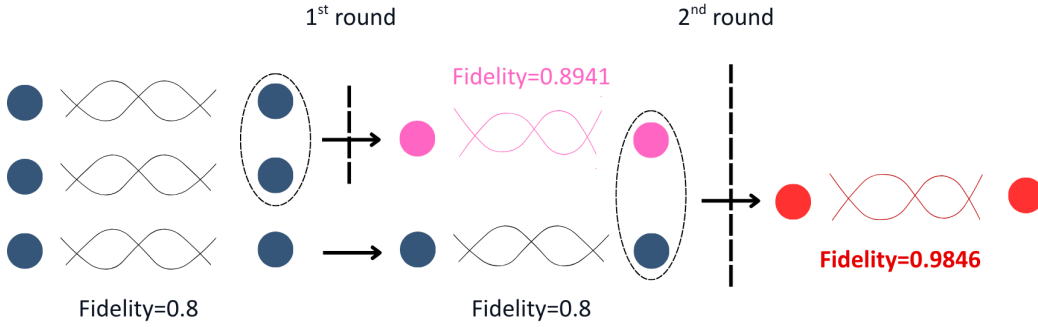


Figure 2.7: Two-round entanglement purification protocol.

Entanglement purification is critical for fidelity-sensitive quantum networking tasks such as quantum teleportation, clock synchronization, and especially quantum key distribution (QKD). However, it introduces nontrivial latency and resource overhead, which must be carefully considered in fidelity-aware routing strategies [6, 9].

2.1.4.3 Quantum Repeater: Definition and Overview

Quantum repeaters are essential components for enabling long-distance quantum communication [27]. They play a crucial role in overcoming the challenges posed by photon loss and noise in quantum channels, such as optical fibers. Quantum communication relies on the transmission of entangled quantum states, which, over long distances, suffer from degradation due to photon loss. Quantum repeaters address this issue by utilizing intermediate nodes to create, store, and swap entanglement between distant quantum

systems, thus enabling the establishment of entanglement over large-scale quantum networks [6, 28].

The operation of a quantum repeater typically involves generating entanglement between quantum systems over short distances, storing the entanglement in memory, and then swapping this entanglement between intermediate nodes to extend the range of quantum communication. This method is vital for mitigating the loss of entanglement that occurs over long distances and overcoming the limitations imposed by photon decay. Quantum repeaters also employ techniques such as quantum error correction (QEC) and entanglement distillation to purify the entangled states and preserve their quality across the network. By ensuring the fidelity and reliability of the entanglement, these repeaters enable the construction of scalable quantum networks.

2.1.4.4 Generations of Quantum Repeaters

Quantum repeaters have evolved through several generations, each improving on the previous by introducing more sophisticated techniques for error suppression and entanglement preservation.

The first generation of quantum repeaters (1st Gen) primarily relied on probabilistic error suppression methods to address photon loss and operational errors. The key protocol used in this generation was the Heralded Entanglement Generation Protocol (HEGP) [28–30], which generates entangled photon pairs in a probabilistic manner. While effective, this approach suffers from a low success rate, necessitating multiple attempts to generate successful entanglement. Additionally, entanglement swapping was employed to extend entanglement over multiple intermediate stations, allowing the entanglement to be distributed over longer distances. Despite these advances, first-generation quantum repeaters are limited by the probabilistic nature of the methods, leading to significant overhead and inefficiency.

The second generation of quantum repeaters (2nd Gen) marked a major improvement by incorporating deterministic error suppression techniques alongside the probabilistic methods used in the first generation. In this generation, quantum error correction (QEC) was applied to detect and correct operational errors, such as those arising from imperfect gates and measurements. This advancement significantly improved the fidelity of the entangled states, ensuring that the generated entanglement remained high quality even under error-prone conditions. Furthermore, second-generation repeaters made use of the Two-way Entanglement Distillation Protocol (2EDP), which was designed to purify the entangled states, enhancing their reliability. Although second-generation repeaters represented a significant

advancement, they still relied on probabilistic methods for photon loss suppression, limiting their overall performance and efficiency.

The third generation of quantum repeaters (3rd Gen) represents the most advanced iteration of quantum repeater technology. These repeaters utilize deterministic error suppression techniques for both photon loss and operational errors. By incorporating quantum error correction (QEC) for both types of errors, third-generation repeaters ensure high-fidelity entanglement across the entire quantum network. This development significantly improves the reliability of quantum communication systems, allowing for faster and more efficient communication over long distances. Third-generation repeaters also operate with higher coupling efficiencies and faster local operations, which enhance the overall performance of the system. However, these repeaters require sophisticated quantum hardware and advanced error-correcting codes, making them more technologically demanding. Despite these challenges, they offer the highest performance for long-distance quantum communication.

The image in Figure 2.8 provides a concise overview of the evolution of quantum repeaters across their three generations, detailing the error types they address, the approaches used, and illustrative schematics. Quantum repeaters have undergone significant advancements from their first-generation counterparts to the current third-generation systems. The figure categorizes error suppression strategies based on whether they combat "Loss Error" or "Operation Error". For each error type, it outlines "Approaches" such as *Heralded Entanglement Generation (HEG)*, *Quantum Error Correction (QEC)*, and *Heralded Entanglement Purification (HEP)*, along with their conceptual "Examples" and "Schematics".

The columns labeled "1G", "2G", and "3G" indicate which generation of quantum repeater utilizes a particular error suppression approach. The first generation (1G) is shown to rely on probabilistic methods, such as *Heralded Entanglement Generation (HEG)* for Loss Error and *Heralded Entanglement Purification (HEP)* for Operation Error. These methods, while useful, have limited efficiency. The second generation (2G) integrates *Quantum Error Correction (QEC)* for Operation Error, representing a significant improvement in reliability, but still relies on HEG for Loss Error. This reliance on probabilistic methods for photon loss suppression limited its overall performance. The third generation (3G) represents the most advanced iteration, indicated by green checkmarks for *Quantum Error Correction (QEC)* in both the Loss Error and Operation Error categories, signifying a fully deterministic error suppression. This provides the highest performance in terms of fidelity and communication rates. These advancements continue to drive the development of scalable quantum networks and represent a critical

step towards realizing a global quantum communication infrastructure.

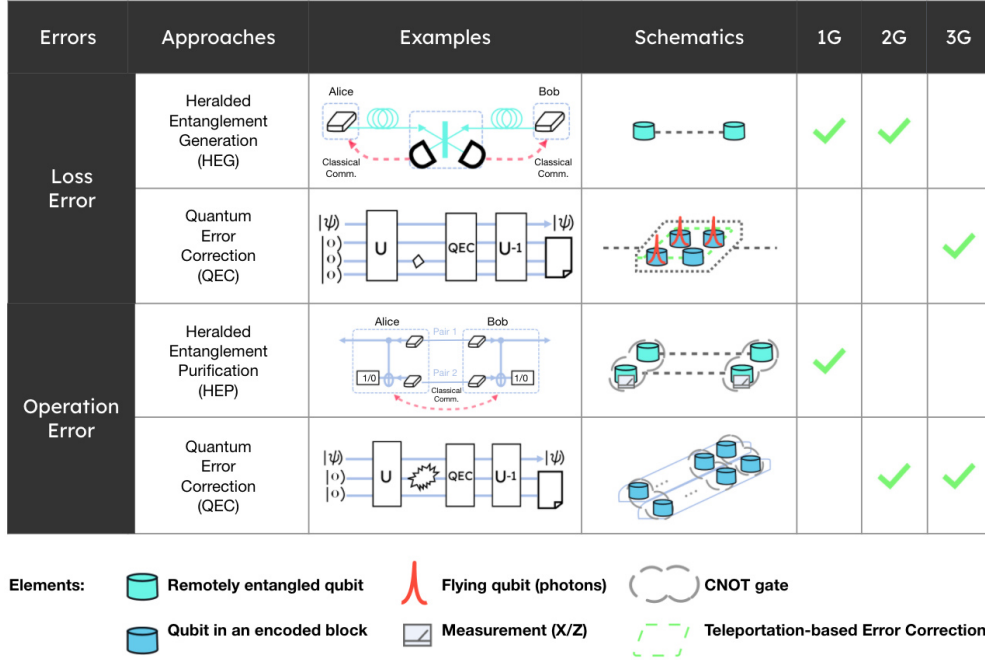


Figure 2.8: Illustration of the fidelity–latency tradeoff.

Quantum repeaters have undergone significant advancements from their first-generation counterparts to the current third-generation systems. The first generation relied on probabilistic methods, which, while useful, had limited efficiency. The second generation integrated quantum error correction to improve the reliability of entanglement generation, but still faced limitations due to its reliance on probabilistic methods for photon loss suppression. The third generation introduced fully deterministic error suppression, providing the highest performance in terms of fidelity and communication rates. These advancements continue to drive the development of scalable quantum networks and represent a critical step towards realizing a global quantum communication infrastructure.

2.1.5 Fidelity and Latency in Quantum Networks

In quantum communication systems, two essential performance metrics are fidelity and latency. These metrics capture a fundamental trade-off that impacts the design of entanglement distribution protocols and routing strategies.

2.1.5.1 Fidelity

Fidelity is a measure of the quality of an entangled quantum state. It quantifies how closely a prepared or distributed quantum state resembles the ideal maximally entangled state [31]. Given an intended target Bell state $|\Phi^+\rangle$, the fidelity of a state ρ is defined as:

$$F(\rho, |\Phi^+\rangle) = \langle \Phi^+ | \rho | \Phi^+ \rangle \quad (2.7)$$

A fidelity value of $F = 1$ indicates a perfect entangled state, while lower values reflect noise, decoherence, or imperfect operations during transmission or entanglement swapping [32].

For example, let's assume we aim to achieve the Bell state $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, but due to imperfections in the quantum system, each qubit has an independent 1% chance of being mis-set. The resulting density matrix ρ might look like this:

$$\rho = \begin{pmatrix} 0.9801 & 0 & 0 & 0 \\ 0 & 0.0099 & 0 & 0 \\ 0 & 0 & 0.0099 & 0 \\ 0 & 0 & 0 & 0.0001 \end{pmatrix}$$

This is a 4x4 matrix, representing the probabilities of the system being in the states $|00\rangle$, $|01\rangle$, $|10\rangle$, and $|11\rangle$, as indicated by the diagonal elements.

We now compute the fidelity $F(\rho, |\Phi^+\rangle)$, using the target state $|\Phi^+\rangle$. The fidelity is calculated as follows:

$$F = \langle \Phi^+ | \rho | \Phi^+ \rangle$$

We substitute the target state $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ and its conjugate transpose:

$$F = \frac{1}{2} (\langle 00 | \rho | 00 \rangle + \langle 00 | \rho | 11 \rangle + \langle 11 | \rho | 00 \rangle + \langle 11 | \rho | 11 \rangle)$$

We can now evaluate each term:

- $\langle 00 | \rho | 00 \rangle = 0.9801$ - $\langle 00 | \rho | 11 \rangle = 0$ (since ρ does not have off-diagonal elements between $|00\rangle$ and $|11\rangle$) - $\langle 11 | \rho | 00 \rangle = 0$ (same reason as above) - $\langle 11 | \rho | 11 \rangle = 0.0001$

Thus, the fidelity becomes:

$$F = \frac{1}{2} (0.9801 + 0 + 0 + 0.0001)$$

$$F = \frac{1}{2} \times 0.9802 = 0.9801$$

This shows that the fidelity is approximately $F = 0.9801$, indicating that the prepared state is very close to the ideal Bell state, with a small error due to imperfections in the quantum system.

A fidelity value close to 1 indicates that the quantum states are highly accurate and resemble the ideal maximally entangled state, while lower values reflect the presence of noise, decoherence, or imperfections in the system. The fidelity calculation provides valuable insight into the reliability and effectiveness of entanglement-based protocols. Understanding and optimizing fidelity is essential for improving the performance of quantum communication networks, as high fidelity ensures the success of entanglement distribution, key exchange, and other quantum operations. As quantum technologies advance, it is crucial to minimize errors and imperfections to achieve high-fidelity entanglement, ultimately enabling the realization of large-scale, efficient quantum networks.

2.1.5.2 Latency

Refers to the total time delay involved in establishing entanglement between two distant quantum nodes [33]. It is an essential factor to consider in quantum communication networks, as high latency can lead to slower communication speeds and diminished performance of entanglement-based protocols. Latency in quantum networks accounts for several contributing factors, including photon transmission delay, classical communication requirements, and the time overhead of routing through intermediate nodes. Each of these factors adds complexity and delays to the process of entanglement distribution, making it crucial to minimize them for efficient quantum communication.

One of the primary contributors to latency is the photon transmission delay. This delay arises due to the time it takes for photons, which carry quantum information, to travel between quantum nodes. The transmission delay is directly related to the physical distance between the nodes and the speed of light in the medium, typically optical fibers or free space. As the distance increases, the photon transmission time increases proportionally, resulting in higher latency. In real-world systems, this is one of the most significant bottlenecks, especially for long-distance quantum networks. For instance, in fiber optic networks, the speed of light in glass fibers is approximately 2×10^8 m/s, which can lead to significant delays over large distances.

Another important contributor to latency is the classical communication required in entanglement generation and purification steps. While quantum entanglement is typically established through quantum operations, classical communication plays a vital role in verifying and synchronizing quantum

operations across distant nodes. This process is known as entanglement swapping, where quantum information is shared between nodes via classical channels. The classical communication channel, which operates at the speed of light, adds further delays, especially when multiple steps of verification or classical communication are required between quantum nodes.

Additionally, quantum networks often require intermediate nodes to facilitate the distribution of entanglement over long distances [34]. These intermediate nodes may need to route quantum information through various pathways or perform quantum memory operations, all of which can introduce additional overhead. For example, in a quantum repeater network, entanglement must be swapped multiple times at intermediate nodes before it can be established between the two end nodes. Each additional intermediate step introduces a delay, and as the number of intermediate nodes increases, so does the overall latency.

Finally, the processing time at each node can also contribute to latency. For instance, when quantum measurements, entanglement generation, or error-correction protocols are implemented, these processes require computational resources and time, which can add to the overall delay. The efficiency of these processes is critical in reducing latency, particularly in real-time quantum communication systems.

2.2 Related Work

2.2.1 Toward Joint Optimization of Fidelity and Latency in Quantum Routing

The design of routing algorithms for quantum networks has become a critical area of research due to the intrinsic limitations of quantum systems, such as decoherence, photon loss, and the fidelity degradation of entangled states over long distances. Many significant prior work has investigated entanglement distribution and routing optimization, primarily focusing on enhancing either the throughput or the fidelity of quantum links, often neglecting latency as a first-class performance metric.

Early efforts in quantum network routing concentrated on improving end-to-end entanglement throughput. For instance, the authors in [35] proposed a throughput-optimized routing framework that aimed to maximize the entanglement distribution rate by jointly considering entanglement generation and classical control signaling. However, this work does not consider the fidelity degradation that occurs during transmission or purification processes. While effective in high-rate settings, such models are insufficient for latency-

sensitive applications, especially when fidelity must be maintained above a threshold to support reliable quantum operations.

Additionally, in an effort to address reliability in lossy network environments, the work in [36] introduced the concept of redundant entanglement provisioning. Their method proactively generates additional entangled pairs to improve success rates and enhance throughput under probabilistic loss. Although this technique improves robustness, it still focuses exclusively on entanglement generation rates, without integrating fidelity or latency guarantees into the routing process.

More recently, there has been a growing recognition of the need to incorporate fidelity-aware strategies into routing protocols. The study that R. Zhou et al. [37] proposes Q2R, a QoS-aware routing framework designed to meet heterogeneous QoS requirements in quantum networks, focusing on maximizing "goodput" instead of just throughput. It formulates the QoS-aware entanglement routing problem considering latency, delivered entanglements, fidelity, and purification. While comprehensive in its framework and scheduling, a limitation of this work is that it does not explicitly define or model latency as a direct link metric for routing decisions at the granular level proposed. Instead, it includes latency as a general request constraint only,

These limitations underscore a key gap in the existing literature: most routing algorithms are designed to optimize fidelity or latency, but not both simultaneously. This dichotomy is increasingly inadequate for emerging quantum network applications that demand low-latency and high-fidelity communication, such as distributed quantum sensing and interactive quantum cloud services. Addressing this challenge, Q-LEAP [9] represents a notable advance by introducing purification-aware routing decisions that maintain fidelity above a guaranteed threshold. However, Q-LEAP still does not incorporate latency-awareness into its metric, leaving room for further refinement in environments where responsiveness is essential.

In contrast to these prior works, our thesis proposes a novel link metric that jointly evaluates fidelity, latency, and resource availability. By incorporating both fidelity guarantees and latency constraints into the routing process, our approach enables more flexible and application-specific decision-making in quantum networks. This hybrid optimization strategy is especially relevant for dynamic environments with varying noise levels, decoherence times, and application demands. In doing so, our work aims to bridge the gap between high-fidelity and low-latency routing strategies, moving toward a more holistic and practical design for quantum internetworking.

2.2.2 Fidelity-Guaranteed Routing: Q-LEAP

Establishing high-fidelity entangled connections over long distances is a fundamental challenge in building scalable quantum networks. As quantum states traverse multiple hops, fidelity degrades due to photon loss, decoherence, and imperfect operations, often rendering the end-to-end entangled state unusable for applications such as quantum teleportation or secure key distribution. Addressing this challenge, Li et al. [9] proposed Q-LEAP (Quantum Low-complExity routing Algorithm) is designed from the perspective of a "multiPlicative" routing metric, is a fidelity-guaranteed entanglement routing algorithm that explicitly incorporates purification decisions into routing to ensure the fidelity of the final entangled state exceeds a given threshold.

Unlike early entanglement routing protocols which focused primarily on throughput or robustness while assuming fidelity would be managed externally, Q-LEAP integrates fidelity into the routing decision process itself. This shift recognizes that routing decisions directly influence fidelity degradation, particularly in multi-hop topologies. Therefore, Q-LEAP was developed to minimize the resource consumption required to meet fidelity guarantees while maintaining low computational complexity, making it suitable for large-scale quantum networks.

At the core of Q-LEAP is a routing strategy based on the multiplicative nature of fidelity degradation. Given that fidelity across a path decays as the product of link fidelities, Q-LEAP employs a modified version of Dijkstra's algorithm to identify paths that maximize end-to-end fidelity. The algorithm then evaluates whether the selected path can satisfy the desired fidelity threshold through limited purification. To avoid the computational overhead of exhaustively exploring all purification combinations, Q-LEAP estimates a target average fidelity for each hop based on the overall fidelity requirement and path length. Links falling below this average are assigned purification levels according to a precomputed purification cost table, ensuring a lightweight evaluation.

The complete Q-LEAP workflow consists of four primary phases:

1. **Initialization:** Gather fidelity information of available links and construct a purification cost table mapping fidelity levels to resource consumption.
2. **Path Search:** Use a fidelity-aware Dijkstra variant to find a candidate path with the highest multiplicative fidelity.
3. **Purification Decision:** Evaluate whether the candidate path meets the fidelity requirement using the average fidelity rule. Assign purification levels only where necessary.

2.3 Summary

This chapter introduced the fundamentals of quantum communication, focusing on the key principles such as superposition, entanglement, and quantum teleportation. The chapter also covered the representation of qubits, their probabilistic nature, and the mathematical framework for quantum states.

Additionally, the chapter explored the challenges in quantum networks, particularly the loss of entanglement over long distances due to photon loss and decoherence. To address this, quantum repeaters were introduced, utilizing techniques like entanglement swapping and purification. The evolution of quantum repeaters was also discussed, from the first generation relying on probabilistic methods to the third generation incorporating deterministic error correction for improved performance.

Finally, the chapter highlighted the trade-off between fidelity and latency in quantum networks, with a focus on optimizing both metrics for efficient communication. The Q-LEAP algorithm, which guarantees fidelity while maintaining low computational complexity, was presented as an example of a fidelity-aware routing approach. However, the chapter concludes by pointing out the need for further research to incorporate latency into routing decisions for practical quantum networking applications.

Chapter 3

Novel Quantum Link Metric Design

3.1 Problem Statement

Quantum networks face inherent trade-offs between fidelity and latency that significantly impact the performance of entanglement distribution and quantum communication. Fidelity, which measures how closely an entangled quantum state approximates an ideal Bell pair, is a fundamental metric for ensuring the reliability of quantum applications such as quantum teleportation, secure quantum key distribution (QKD), and distributed quantum computing. However, fidelity deteriorates exponentially with distance due to photon loss, decoherence, and imperfect operations [6].

To compensate for fidelity degradation over long distances, protocols often employ entanglement swapping and purification, which introduce operational overhead and increase the time required to establish end-to-end entanglement. As a result, latency becomes a critical performance bottleneck. Latency encompasses not only the physical propagation delay, but also the probabilistic nature of entanglement generation, queueing delays, and the cost of entanglement-enhancing operations. These delays are particularly problematic in latency sensitive scenarios, such as distributed quantum algorithms and real-time QKD [38, 39].

Crucially, fidelity and latency are not independent. Enhancing fidelity through additional purification rounds or longer multihop routes tends to increase latency [38]. Conversely, minimizing latency by preferring shorter paths may yield insufficient fidelity for successful quantum operations, especially in low-quality or long-range links.

This fidelity-latency interaction introduces a challenging trade-off in quantum network routing decisions, as graphically illustrated in Figure 3.1. The figure depicts a generalized network topology with a designated source and destination, where each node functions as a quantum repeater. Thin lines within the network represent shorter path segments, while bold lines

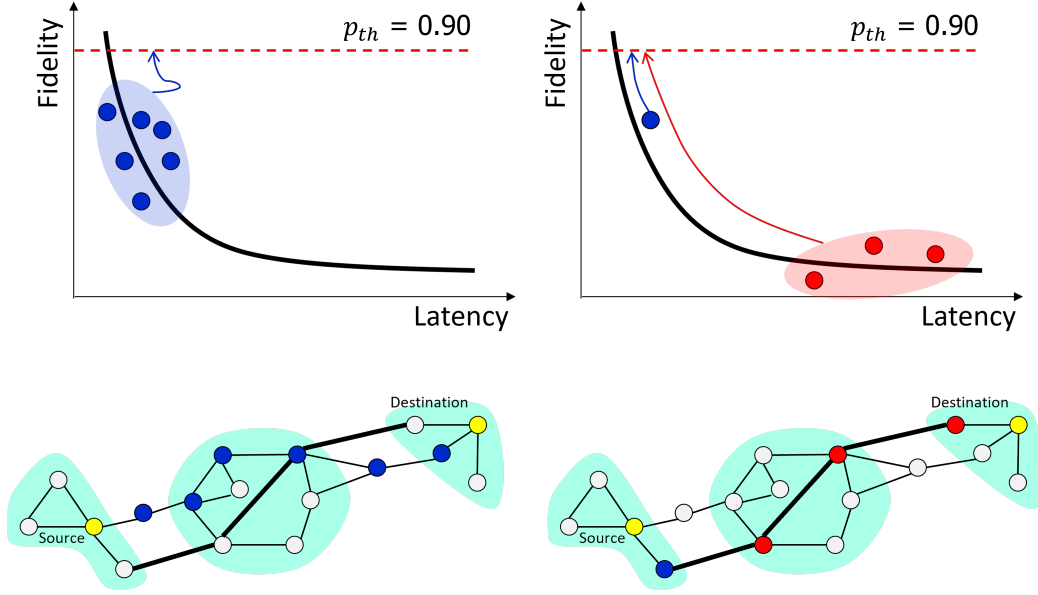


Figure 3.1: Illustration of the fidelity-latency tradeoff.

denote longer path segments.

The left scenario in Figure 3.1 exemplifies the challenge of prioritizing fidelity. It highlights potential routing paths (blue nodes and highlighted sub-networks) that, while achieving a high quantum fidelity (approaching or exceeding a defined threshold, $p_{th} = 0.90$, as p_{th} refers to the minimum acceptable level of fidelity for a quantum entangled state in a quantum network, below which the entanglement is considered unreliable for further use in communication or computation), inherently incur substantial latency. This increased latency is typically a consequence of either extensive entanglement purification operations required to boost quality, or a greater number of hops/repeaters along the path, each contributing to propagation and operational delays.

Conversely, the right scenario in Figure 3.1 showcases routing paths (red nodes and highlighted sub-networks) designed for low-latency transmission. While these paths offer quicker delivery of quantum states, they demonstrably fail to meet the required fidelity threshold. This often occurs because minimizing latency might involve sacrificing purification steps or traversing inherently noisier links, resulting in a degraded quantum state quality that falls below acceptable application requirements.

Existing routing algorithms tend to optimize for a single objective, either maximizing fidelity or minimizing latency, but rarely both simultaneously. Fidelity-centric approaches, such as Q-LEAP [9], ensure quality by prioritiz-

ing fidelity, but risk high end-to-end latency. This trade-off of latency and fidelity is specify in Figure 3.2.

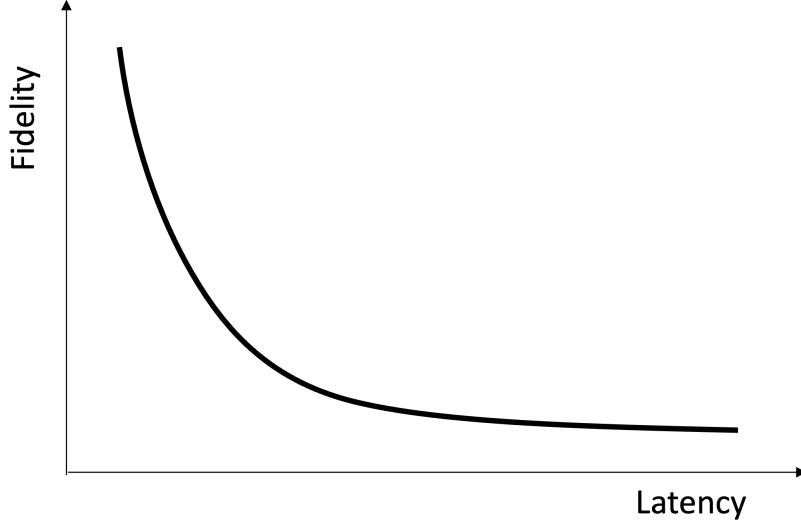


Figure 3.2: Trade-off between fidelity and latency in quantum entanglement routing. Longer paths with purification improve fidelity but increase latency, while shorter paths reduce latency but may fall below the fidelity threshold.

However, latency-oriented strategies favor speed and responsiveness, but can compromise quantum communication reliability. This gap highlights the inadequacy of unidimensional routing metrics to balance the nuanced requirements of real-world quantum network applications [17].

Thus, there is a critical need for a routing strategy that jointly considers fidelity and latency, enabling more adaptable and efficient path selection under varying network conditions. The development of such a composite metric is the core focus of this work.

3.2 Latency Analysis of Entanglement Distribution Protocol

3.2.1 Latency in Nested Purification Protocol

3.2.1.1 System Description

The illustration of quantum networks with the chain of quantum repeaters is shown in Figure 3.3, The total distance from Alice to Bob is denoted

as the end-to-end connection of length (L) km, which is divided into 2^n elementary links, called L_0 km. In order to keep the entanglement on L_0 elementary link, the nested purification protocol (NPP), which merges the methods of entanglement swapping and purification into a unified protocol, can be used to perform BSM at each quantum repeater, to extend the existing entanglement on $S_{2^{n-1}}$ links to S_{2^n} links. We must perform BSMs at quantum repeaters S_i , for $i = 0, 1, 2, \dots, n$, where n indicates the nesting level at which n th times of Bell-state measurement (BSM) are performed in the NPP protocol.

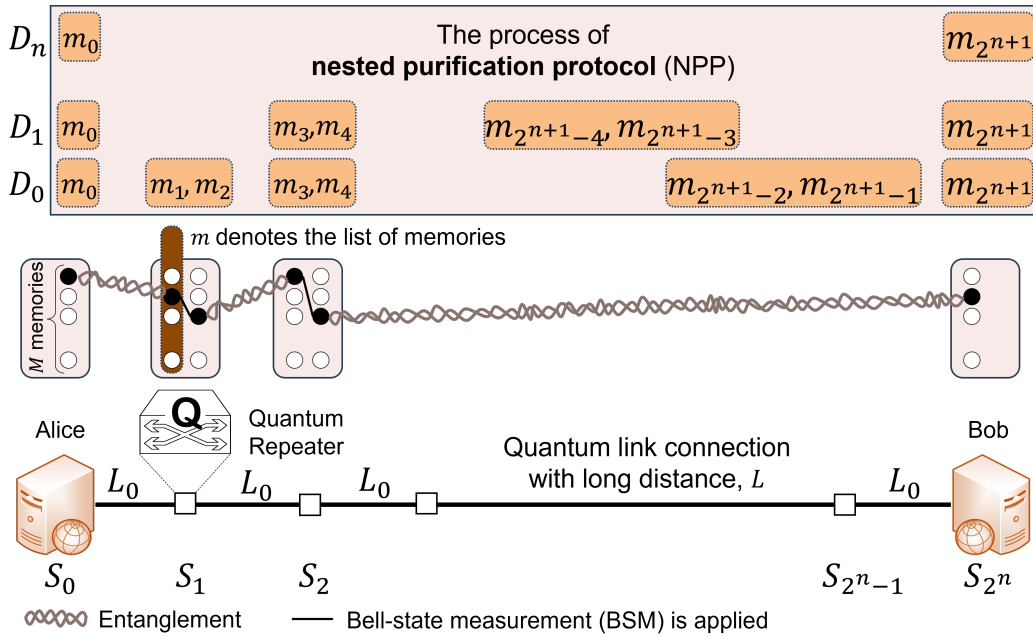


Figure 3.3: Illustration of the nested purification protocol (NPP) over the chain of quantum repeater in quantum networks

The NPP protocol combines the methods of entanglement swapping and purification into single protocol that needs the number of memory copies (M) to perform its purification mechanism in parallel fashion. Thus, the NPP with the list of memories $m = 1, 2, \dots, M$ operates in cycles with a period of T_{ED} , which denotes the fundamental time period for making entanglement distribution attempts. The NPP protocol involves multiple levels of nesting, where each level corresponds to a different stage of entanglement swapping and purification. The goal is to extend entanglement over long distances by dividing the total distance into smaller segments and progressively connecting these segments through higher levels of nesting.

Based on the research in [40], our study will inherit and focus on the concept of multiple quantum repeater links with 2^n elementary links, which are arranged equally at each distance L_0 with the total length L of the chain and n nesting level ($n > 0$) of NPP protocol. As shown in Figure 3.3, each BSMs is applied at each quantum repeater S_{2^n} with the list of memories, in which the entanglement swapping for each quantum repeater has the success probability (P_S) and can achieve the maximum probability of conclusive success for a given BSM as $P_M < 1$. As noted in [40], P_M cannot achieve one because an inefficiency of photodetectors may result in inconclusive outcomes when the BSM is performed. The assumption in our study that is our entanglement distribution scheme is heralding with success probability (P_S), shows the confirmation of our entanglement process has succeeded or not.

Figure 3.3 also shows the NPP protocol, stated that once entanglement is established across all D_0 links, it becomes possible to perform all necessary BSMs simultaneously at each quantum repeater, but there is no assurance that a BSM at one nesting level will yield the entangled states needed for the next level. With the k th-order NPP protocol, for $k = 1, 2, \dots, n$, called a cyclic protocol at each period of time, will try to entangle unused memory of D_0 links, the result will show the success or failure of each attempt in the next cycle. In our study, we assume that $k = n$, with the entanglement is established over D_0 link and BSMs are performed at all quantum repeaters without knowing the success or failure of previous entangling attempts.

In our study, the architecture of quantum repeater in the quantum networks has two connected elementary links. The quantum repeater has a single-photon detector that can spectrally resolve in sequence on each of two time bins of the qubit in single frequency. Upon successful projection by the BSM on one of the Bell states in at least one of the multiple frequencies. Upon receiving a pair of which-frequency information from the adjoining elementary links, two memories translate their qubits to one predetermined common frequency. A BSM at a single frequency is then performed on this pair. P_S is the success probability for a single frequency. We assume that a universal synchronized clock is available with the clock rate of the system, i.e., T_{ED} is limited by the time it takes to perform the BSM. We denote that τ_{BSM} is the time taken to perform the BSM at the elementary link center, whereas τ'_{BSM} is the time taken to perform the BSM at the quantum repeater and the time for loading (readout) of the qubits to (from) the memories. Therefore, both τ_{BSM} and τ'_{BSM} values between Alice and Bob in long-distance connection are contributing to the total latency of making entanglement distribution attempts.

3.2.1.2 Latency-Distance Analysis

In the entanglement distribution, Alice and Bob obtain mismatched bits, which is indicated as error probability. This error probability are functions of loss-noise parameters of photodetectors in the elementary links and memory quantum repeaters. In our study, we consider the source of error is memory dephasing, whose effect on the rate in conjunction with the purification technique. With the ideal quantum memories in a quantum repeater [40], i.e., there is no source of error and no purification is needed to be performed in the system, the steady-state rate of entanglement generation per each of m ideal memories used in the n th-order NPP protocol is given by

$$Q_n = \frac{P_S P_M^n}{2L/c} \quad (3.1)$$

where c is the speed of light in the optical fiber.

Suppose we consider the case of imperfect memories, the memory degradation by a dephasing process can be modeled as the qubit state of memory after decaying for a time period. Therefore, we need to perform BSM at the quantum repeater to ensure two memories of qubit state are maximally entangled. This process is also called entanglement swapping, which requires some entanglement cost ($E_C(\cdot)$) to do the distillable entanglement. With the aid of purification in the NPP protocol, we can purify memories and obtain the achievable rate according to the multiplication of $Q_n E_C(\hat{\rho}_{XY}(t))$. $\hat{\rho}_{XY}(t)$ is the estimate number of maximally entangled states that is obtained out of the m copies of $\hat{\rho}_{XY}$ in a period of time t . For sake of simplification, the achievable rate is given by

$$R_n = Q_n E_C(\hat{\rho}_{XY}(t)) \quad (3.2)$$

Suppose entanglement distribution scheme on a single photon along the quantum channel is optimally accomplished by the NPP protocol. We consider a long-distance transmission, then the success probability degrades exponentially with L . In particular, the achievable rate can be expressed as

$$R_n = e^{\left(-2\sqrt{\frac{L \ln(1/P_M)}{c\tau_c}}\right)} \quad (3.3)$$

where τ_c is the memory coherence time.

It was stated in [41] that latency plays a significant role in determining the overall key generation rate. In principal, the latency affects the conditional fidelity of photons, which in turn influences the success probabilities of measurements and entanglement generation. This chain of dependencies ultimately determines the key generation rate. Therefore, reducing latency

is crucial to improve the overall performance and efficiency of the quantum repeater through the purification protocol. Hereby, the total latency is given by

$$\tau_{tot} = T_{(*)} + \tau_{e2e} \quad (3.4)$$

where

$$\tau_{e2e} = \frac{L}{c} \quad (3.5)$$

and $(*)$ is the purification protocol for performing the entanglement distribution and entanglement swapping in between Alice and Bob connection. For example, T_{NPP} is referred to the NPP protocol [8]. Therefore, the total time taken in NPP protocol can be expressed as

$$T_{(NPP)} = n(\tau_{elem} + \tau'_{BSM}) \quad (3.6)$$

where

$$\tau_{elem} = \frac{L_0}{c} + \tau_{BSM} \quad (3.7)$$

where τ_{BSM} is BSM operation at the elementary link, usually 150 ns [42]. Since the BSM operation is inversely proportional to the achievable rate, therefore, we can express the BSM time taken at the quantum repeater is given by

$$\tau'_{BSM} = \frac{2^n}{R_n} \quad (3.8)$$

In order to formulate the conditional fidelity of a quantum repeater with imperfect memory, we can re-write the generation rate in the form of entanglement cost as follow:

$$R_n = Q_n E_C(\hat{\rho}_{XY}(T_n)) \quad (3.9)$$

where $T_n = \max(T_{ED}, T_{n-1})$ and $T_{ED} = \frac{L_0}{c}$.

Suppose we consider $p(t)$ represents the conditional fidelity of the measured quantum state in the memory, in which it is a measure of how close a quantum state is to a desired pure state, showing how well the memory can preserve the entangled state over time despite memory dephasing [40]. In this context, it measures how well the quantum memory preserves the entangled state over time t . As time t increases, the exponential term $e^{(-t/\tau_c)}$ decreases, indicating that the memory's coherence is decaying. The conditional fidelity $p(t)$ approaches 0.5 as t becomes large, indicating a 50% chance of the memory retaining its quantum state. Hence, the conditional fidelity decreases over time as the coherence of the memory is lost. Thus, the conditional fidelity is given by

$$p(t) = \frac{1 + e^{(-t/\tau_c)}}{2} \quad (3.10)$$

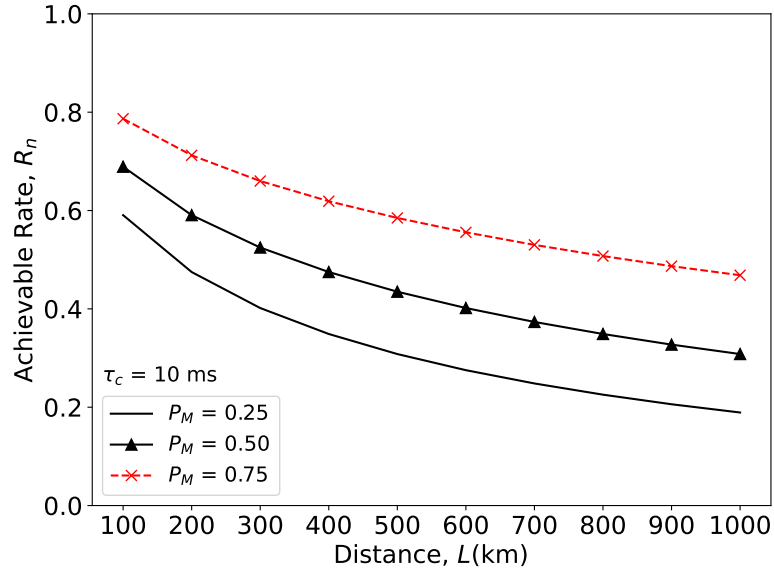
3.2.1.3 Numerical Studies

Table 3.1: Simulation Parameters and Settings for NPP Protocols

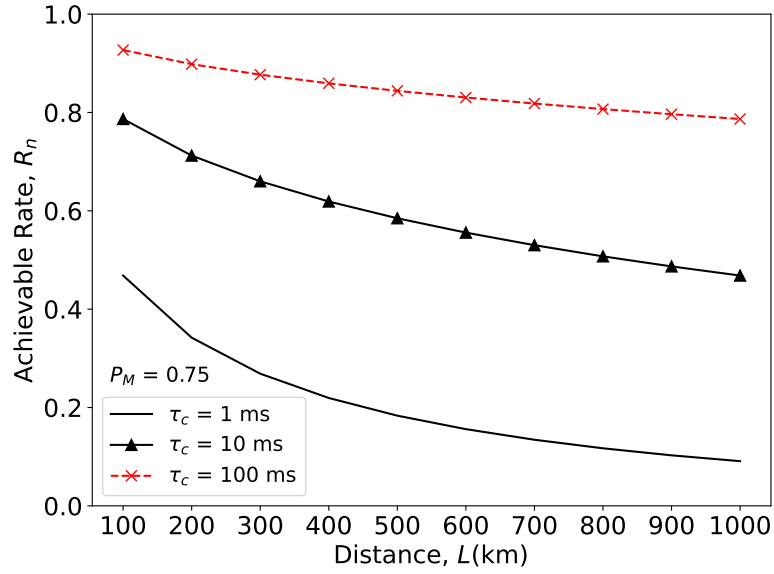
Parameter	Symbol	Value
Speed of light in a fiber	c	2×10^8 m/s
Memory coherence time	τ_c	10 ms
BSM time taken at link	τ_{BSM}	150 ns
Maximum Success Probability	P_M	0.25, 0.50, 0.75

We quantitatively study the influence of latency and conditional fidelity on the long-distance quantum networks based on the derivation of achievable rate. Beside with the setting parameters in Table 3.1, we assume that $P_S = 0.2 \times 10^{-0.01L_0}$, which is the best choice for the NPP protocol in [40]. The optimum values for NPP is set to n to maximize the achievable rate for the required memory coherence time is 10 ms. Figure 3.4 demonstrates to us the two subfigures illustrating the achievable rate R_n as a function of distance L . Figure 3.4 (a) analyzes the impact of different values of the maximum success probability, P_M . The three curves represent $P_M = 0.25$ (black, no markers), $P_M = 0.50$ (black, triangle markers), and $P_M = 0.75$ (red, cross markers), while keeping the memory coherence time fixed at $\tau_c = 10$ ms. The result of the achievable rate decreases with increasing distance but improves with higher P_M .

While in Figure 3.4 (b) investigates the impact of different τ_c on the achievable rate while keeping $P_M = 0.75$ is fixed. The three curves correspond to $\tau_c = 1$ ms (black, no markers), $\tau_c = 10$ ms (black, triangle markers), and $\tau_c = 100$ ms (red, cross markers). The graph shows that the longer memory coherence time maintains a higher achievable rate over larger distances. Similar to that is in Figure 3.5, presents the entanglement generation rate Q_n in 2 cases, with nesting level is fixed at $n = 2$. Fig. 3.5 (a) shows the entanglement generation rate Q_n in different distance L with various values of P_M . The graph result showing a similar trend to R_n , which is with higher P_M leading to better performance. For Figure 3.5 (b) analyzes the impact of the nesting level n at a fixed distance of $L = 1000$ km. As n increases, the entanglement generation rate improves, particularly for higher values of P_M . These results highlight the dependence of entanglement generation on both P_M and n , with larger values leading to enhanced performance.

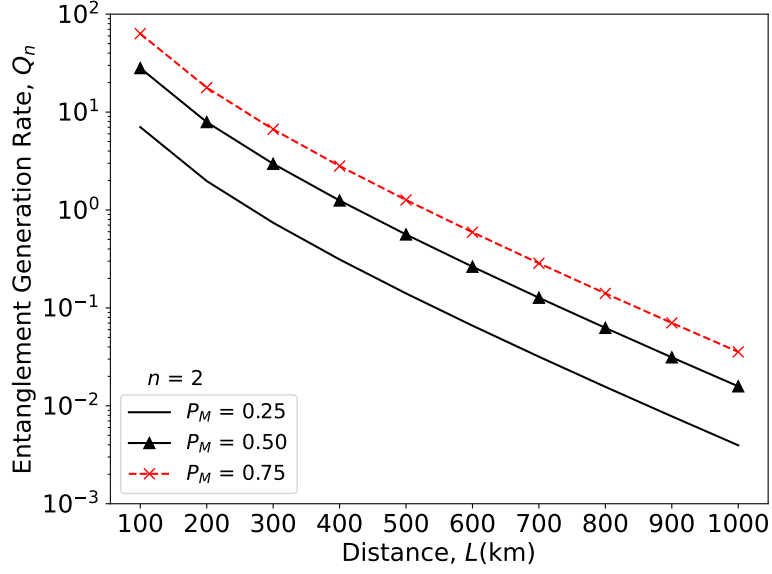


(a)

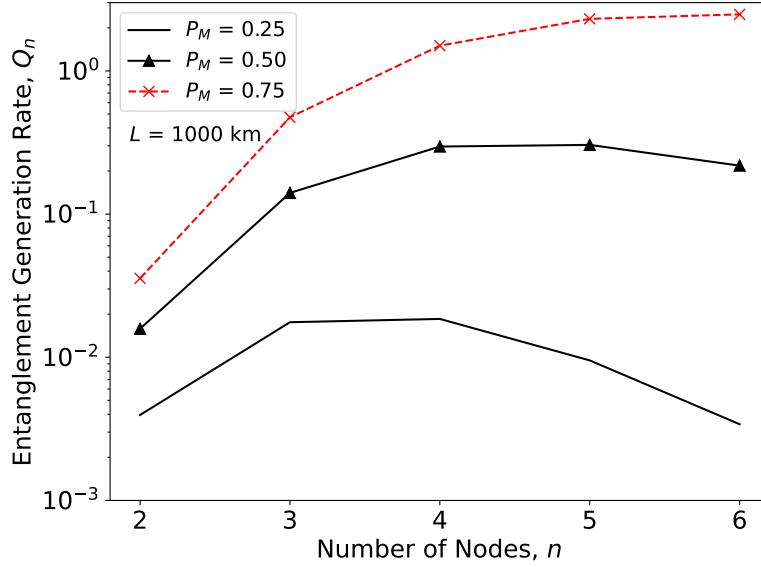


(b)

Figure 3.4: Performance of achievable rate versus distance with (a) Different maximum success probability; and (b) Different memory coherence time



(a)



(b)

Figure 3.5: Performance of entanglement generation rate versus (a) Distance and (b) Nesting level with different maximum success probability

Meanwhile, Figure 3.7 illustrates the success probability P_S as in different distance L and nesting levels n . Our results show that the P_S decreases with increasing distance across all nesting levels. The curves demonstrate the higher the nesting levels, the higher the success probability over long

distances.

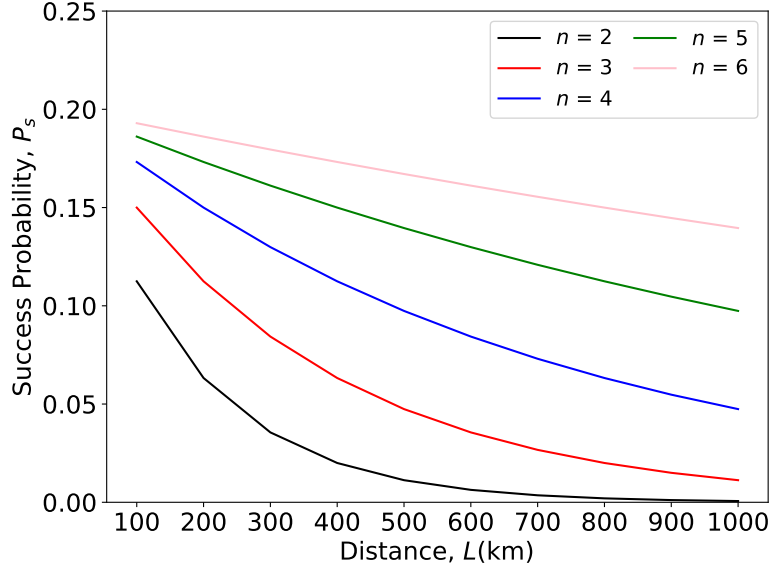


Figure 3.6: Performance of success probability versus distance with different nesting levels

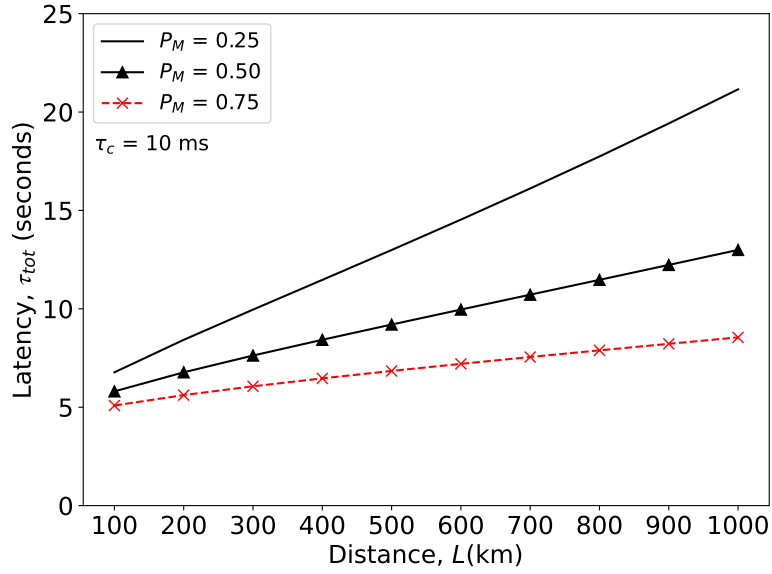


Figure 3.7: Performance of latency versus distance with different maximum success probability

Figure 3.7 shows the latency is plotted against the distance that includes three different curves, corresponding to P_M values of 0.25, 0.5, and 0.75, with

a fixed $\tau_c = 10$ ms. The latency values increase with distance, and the curves show that higher P_M values result in lower latency across the entire distance range.

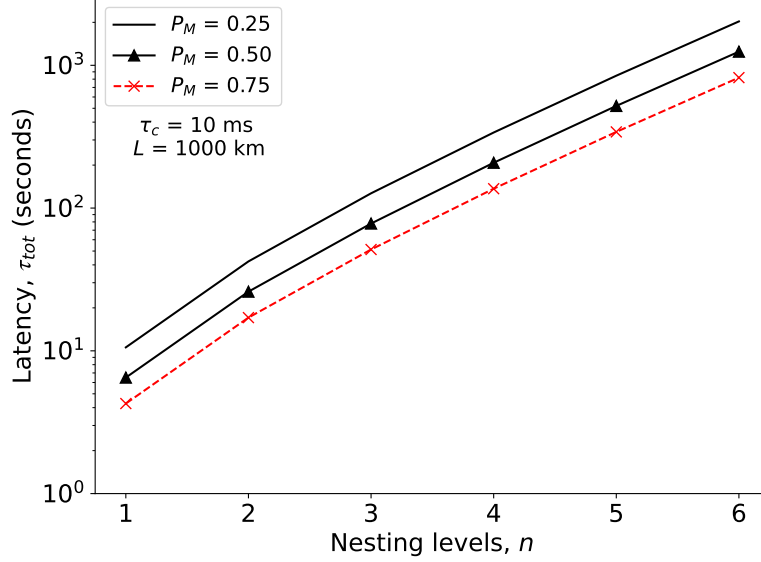


Figure 3.8: Performance of latency versus nested level with different maximum success probability

Figure 3.8 shows the performance of latency versus nested level with different P_M with a fixed τ_c and L . The figure includes three different curves, corresponding to P_M values of 0.25, 0.5, and 0.75. The curves demonstrate that as the distance increases, the latency exponentially increases, and the lower the P_M value, the higher the latency for a given distance.

The given line graph in Figure 3.9 illustrates the relationship between the conditional fidelity $p(t)$ and the time period t for three different values of memory coherence time τ_c (i.e., 1 ms, 10 ms, and 100 ms). Overall, the conditional fidelity decreases as the time period increases for all three cases. However, the rate of decline varies depending on τ_c . The highest fidelity is observed when $\tau_c = 100$ ms, while the lowest is seen for $\tau_c = 1$ ms. Initially, at $t = 0.0001$ seconds, all three curves start close to a fidelity value of 1.0. As time progresses, the fidelity for $\tau_c = 1$ ms (black, no markers) declines rapidly, dropping below 0.7 by $t = 0.01$ seconds. In contrast, the fidelity for $\tau_c = 10$ ms (black, triangle markers) also decreases but at a slower pace, maintaining a value above 0.8 throughout the range. Meanwhile, the fidelity for $\tau_c = 100$ ms (red, cross markers) remains close to 1.0, showing only a slight decline over time. In summary, a longer memory coherence time results in

better retention of conditional fidelity over time, whereas a shorter memory coherence time leads to a faster deterioration in performance, leading to short distance.

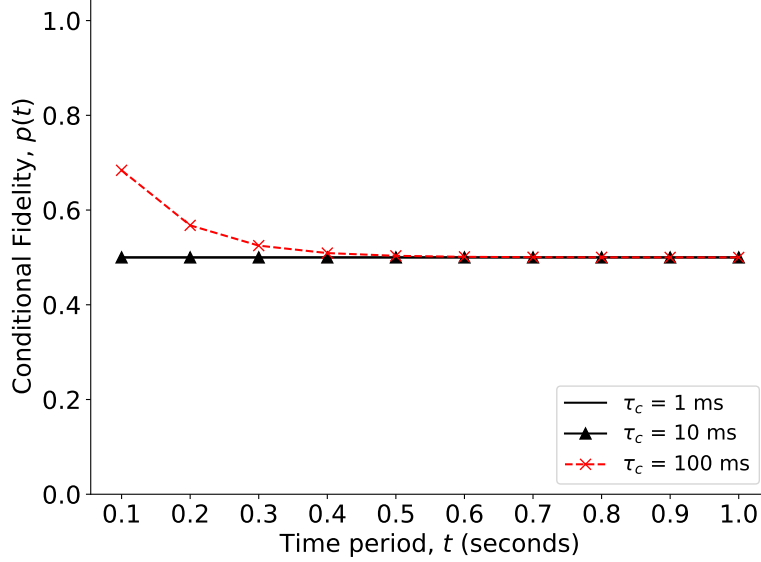


Figure 3.9: Performance of conditional fidelity versus the time period with different memory coherence time

3.2.2 Latency in Entanglement-guaranteed Distribution Protocol

3.2.2.1 System Description

The overall system architecture is a quantum network modeled as a graph $G = (V, E)$, where nodes $v \in V$ are quantum repeaters and edges $e \in E$ are quantum communication links. This framework, which is the basis for the Q-LEAP algorithm, operates on a time-slotted model where entangled pairs are generated on each link at the start of a time slot. Unlike the multi-level, nested approach of NPP, the Q-LEAP-based system focuses on finding a single, end-to-end path that satisfies specific fidelity threshold requirements for a given source-destination pair. The primary entanglement operations employed are entanglement generation, entanglement purification, and entanglement swapping.

Each quantum repeater in the network is assumed to have the necessary functionalities of a first-generation repeater, including the ability to generate entangled pairs, perform Bell-state measurements (BSMs) for swapping,

and execute purification operations. The Q-LEAP algorithm, as a routing strategy, makes intelligent decisions on which links to use and which links require purification to satisfy a specified fidelity threshold. It selects a single path, and purification decisions are made on a link-by-link basis along that path to ensure the final end-to-end fidelity is met. A key distinction is that, unlike the hierarchical connection-purification cycles of NPP, the Q-LEAP model's latency is a summation of delays across a series of individually-purified links, as entanglement swapping is performed sequentially at each repeater node along the chosen path. This approach provides a more direct and computationally efficient method for routing in dynamic network environments.

3.2.2.2 Latency-Distance Analysis

In the context of Q_{FiLa} , and as utilized by the Q-LEAP algorithm, latency reflects the time required to generate and deliver entangled states. This end-to-end latency (l) for a path is modeled as the sum of three primary components: entanglement generation time ($t(E)$), propagation delay ($t(d)$), and quantum memory storage time ($t(N)$):

$$l = t(E) + t(d) + t(N) \quad (3.11)$$

Each component contributes to the overall communication delay, reflecting the inherent complexities of quantum networks operating with first-generation repeater functionalities.

$t(d)$: Transmission Delay Time This component represents the physical time light takes to traverse the quantum channel. It is directly proportional to the total distance (L) of the end-to-end connection and inversely proportional to the speed of light (c) in the optical fiber:

$$t(d) = \frac{L}{c} \quad (3.12)$$

The speed of light in fiber is typically 2×10^8 m/s. This is a fundamental physical constant that applies universally to any quantum communication scheme.

$t(N)$: Quantum Memory Delay or Storage Time This term accounts for the time qubits must be stored in quantum memories while awaiting synchronization or the successful completion of other probabilistic operations in the network. In this study, $t(N)$ is treated as a constant overhead, set to 500 ms based on the assumed synchronization timestep. However, the efficiency of this storage time is fundamentally linked to the memory coherence time (τ_c) that we have discussed.

$t(E)$: Entanglement Generation Time Representing the time required for entanglement generation, our model captures the cumulative operational delays across the chosen quantum path. This time is fundamentally influenced by the probabilistic nature of entanglement operations within first-generation quantum repeaters, meaning multiple attempts may be necessary on any given link before a successful entangled pair is established. Specifically, for each segment along the path, $t(E)$ encompasses the initial preparation time for qubits, the crucial round-trip communication delay for classical control signals between adjacent nodes (which scales with physical distance), and any necessary processing or reset time at the repeater nodes. The overall contribution of a link to $t(E)$ is then determined by the average number of attempts required, inversely proportional to the success probability per entanglement attempt (P_s). By summing these average times over all links in the path, the model aggregates the total operational overhead inherent in preparing the quantum resource for an end-to-end connection, as utilized by algorithms like Q-LEAP.

It should be noted that $t(E)$, $t(d)$, and $t(N)$ remain relatively stable across different network scenarios and thus represent the baseline link latencies inherent to the quantum network infrastructure. In contrast, the path computation time (i.e., the time taken by the Q-LEAP algorithm itself to find a path) varies depending on the complexity of the routing algorithm and the size of the search space. This algorithmic computation time is a critical performance metric when evaluating different routing strategies, but it is distinct from the quantum communication latency (l) modeled here.

3.2.2.3 Numerical Studies

The given line graph in Figure 3.10 illustrates the relationship between the latency of an entangled link between two nodes (l) and distance (L) for a connection established using Q_{FiLa} across three different values of entanglement success probability (p_s). Overall, latency increases as distance increases for all three cases. However, the rate of increase varies depending on p_s . The lowest latency is observed when $p_s = 0.75$, while the highest is seen for $p_s = 0.25$.

Initially, at $L = 100$ km, all three curves start with relatively similar latency values. As distance progresses, the latency for $p_s = 0.25$ (black line, no markers) increases rapidly, reaching approximately 0.7 seconds by $L = 1000$ km. In contrast, the latency for $p_s = 0.50$ (black line, triangle markers) also increases but at a slower pace, staying below 0.6 seconds throughout the range. Meanwhile, the latency for $p_s = 0.75$ (red line, cross markers) remains significantly lower than the two remaining curves across the entire distance

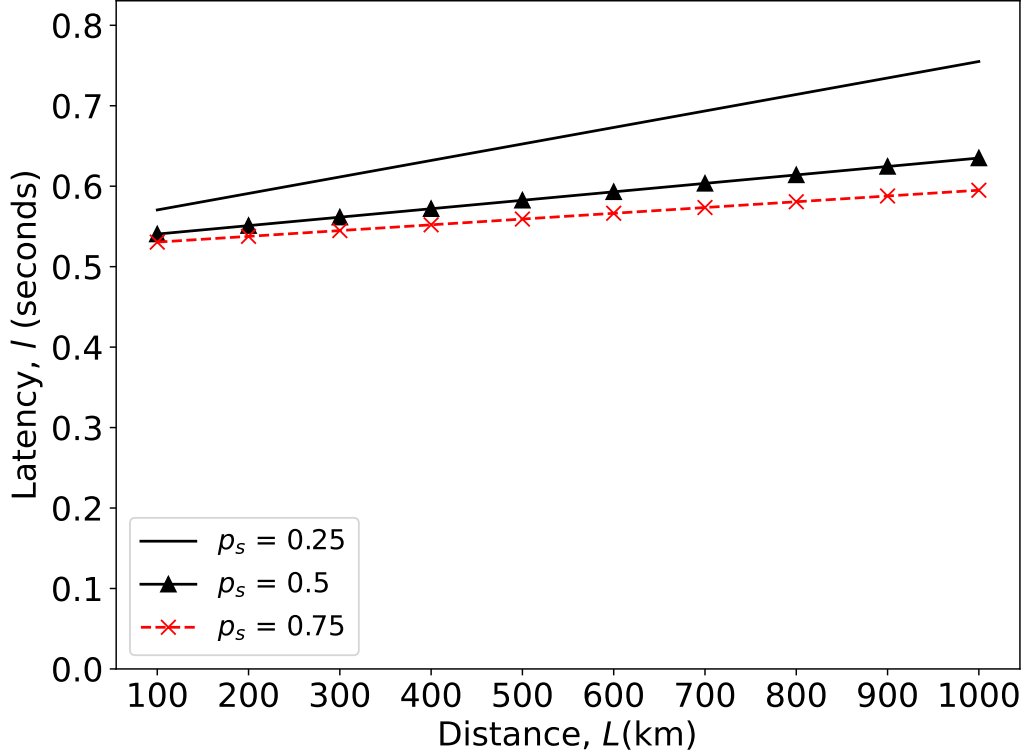


Figure 3.10: Performance of latency versus distance with different entanglement success probability (p_s)

range.

In summary, a higher success probability (p_s) results in lower latency across varying distances. This is because higher p_s values directly translate to a more efficient execution of probabilistic quantum operations (such as entanglement generation, purification, and swapping) within first-generation quantum repeaters. A higher success probability means fewer retries are needed on average for each operation, which significantly reduces the waiting time incurred by failed attempts and the associated two-way classical communication for confirmation. Conversely, a lower success probability leads to a faster increment in latency over long distances due to the increased number of required attempts and resultant delays.

3.3 Proposed Quantum Link Metric: Q_{FiLa}

To address the challenge of jointly optimizing fidelity and latency in quantum routing, we introduce a novel link metric named Q_{FiLa} (Quantum Fidelity-

Latency). This composite metric is specifically designed to guide routing decisions by quantifying the trade-off between fidelity degradation and communication delay over quantum links. By incorporating both fidelity and latency in a normalized scalar function, Q_{FiLa} enables balanced and flexible path selection under heterogeneous quantum network conditions and application demands.

3.3.1 Fidelity Model

Fidelity in quantum networks measures how well an entangled state preserves its intended quantum properties after transmission. It is a key performance metric in applications such as quantum teleportation, QKD, and distributed quantum computing. Due to the cumulative effects of photon attenuation and decoherence, fidelity degrades with distance. This behavior can be captured by an exponential decay model [6]:

$$f = \mu \left(1 + e^{-d \left(\frac{L_{att} + L_{dec}}{L_{att} L_{dec}} \right)} \right) \quad (3.13)$$

where:

- d : Is the distance between two nodes
- $\mu \in [0.5, 1]$: Represents the average fidelity decay baseline (typically 0.5),
- L_{att} : Is the photon attenuation length,
- L_{dec} : Is the decoherence length.

Both L_{att} and L_{dec} are treated as fixed constants, with values of 100 km and 500 km, respectively. As distance increases, the exponential decay dominates, causing fidelity to approach the lower bound. This model accurately reflects the deterioration observed in practical quantum links.

3.3.2 Q-LEAP Overview

Building upon the Q-LEAP routing algorithm introduced by Li et al. [9], we extend the algorithm to incorporate the Q_{FiLa} link metric, which balances both fidelity and latency in routing decisions. Q-LEAP was originally designed to guarantee fidelity over multi-hop quantum communication paths by incorporating purification decisions directly into the routing process. The key steps of the Q-LEAP algorithm remain the same, but we adapt it to integrate the composite metric for enhanced performance in real-world quantum networks.

In this work, we extend Q-LEAP by incorporating the Q_{FiLa} metric, which includes both fidelity and latency as factors in path selection. The modified Q-LEAP algorithm now not only searches for high-fidelity paths but also accounts for the latency along each path. This modification allows for adaptive and flexible routing decisions that balance the two key performance metrics, making Q-LEAP suitable for dynamic, large-scale quantum networks.

3.3.2.1 Q_{FiLa} Formulation for Q-LEAP

In quantum communication networks, it is essential to evaluate the performance of links that balance multiple factors, such as the quality of entanglement (fidelity) and the time delay (latency). To achieve this, we introduce a unified routing metric that combines both fidelity and latency into a single value, allowing for a more comprehensive assessment of the link's overall performance. This is done by first normalizing both the fidelity and latency values to the same range, specifically $[0, 1]$, ensuring that they are comparable and can be combined effectively in subsequent calculations. The normalization of these values ensures consistency across different scales and units of measurement.

The fidelity normalization process is described by the following equation:

$$F' = \frac{f}{f_{\max}} \quad (3.14)$$

Here, the fidelity value f of the entangled link between two quantum nodes is normalized by dividing it by $f_{\max}$, which represents the maximum observed fidelity among all available links. This ensures that the normalized fidelity, F' , falls within the range $[0, 1]$, where a value of $F' = 1$ indicates the highest possible fidelity.

Similarly, latency is normalized using the following equation:

$$L' = \frac{l}{l_{\min}} \quad (3.15)$$

In this case, the latency value l of the entangled link between two nodes is normalized by dividing it by $l_{\min}$, the minimum latency observed across all available links. By normalizing in this way, the resulting L' value also falls within the range $[0, 1]$, with a value of $L' = 1$ corresponding to the lowest latency.

To summarize the variables used in these equations:

- f and l : These represent the fidelity and latency values, respectively, for a particular entangled link between two quantum nodes in the network.

- $f_{\max}$ and $l_{\min}$: These denote the maximum observed fidelity and the minimum observed latency across all available entangled links in the network.

The normalization process ensures that both fidelity and latency are on comparable scales, making it possible to combine them into a single composite metric. By normalizing the values, we eliminate the issue of differing units and magnitudes between these two important performance indicators.

After normalizing both fidelity and latency, we combine them into a single composite metric, Q_{FiLa} , that balances the trade-off between these two parameters. The final formula for the Q_{FiLa} metric is given by:

$$Q_{FiLa} = w_f F' + w_l L' \quad (3.16)$$

In this formula, w_f and w_l are application-defined weights that control the relative importance of fidelity and latency in the overall metric. These weights must satisfy the condition $w_f + w_l = 1$, ensuring that the combined metric reflects a normalized weighting of both factors. The weight w_f prioritizes fidelity, while w_l prioritizes low latency. Depending on the specific requirements of the application, these weights can be adjusted to favor either high-quality entanglement or low transmission delay, or a balance of both.

Thus, the Q_{FiLa} metric allows us to rank quantum communication links based on their overall performance, considering both the fidelity of the entanglement and the latency of the connection. The overall metric score increases when both fidelity and latency improve, making this composite metric highly useful for decision-making in routing algorithms, especially in scenarios that require a balance between high-quality entanglement and low-latency transmission.

Figure 3.11 illustrates a random network topology where each node represents a quantum repeater, and the connections between nodes represent quantum communication links. The image highlights three distinct cases of routing decisions based on the Q_{FiLa} metric, which combines both fidelity and latency considerations for link selection.

1. **Left Case - Best Fidelity:** In this scenario, the routing algorithm prioritizes selecting links that offer the highest fidelity. These links are typically shorter, as indicated by the light lines in the image. The objective is to maximize the quality of entanglement between quantum nodes, ensuring that the entangled states remain as close to ideal as possible.
2. **Right Case - Best Latency:** Here, the routing algorithm focuses on minimizing the time delay in communication by selecting links with

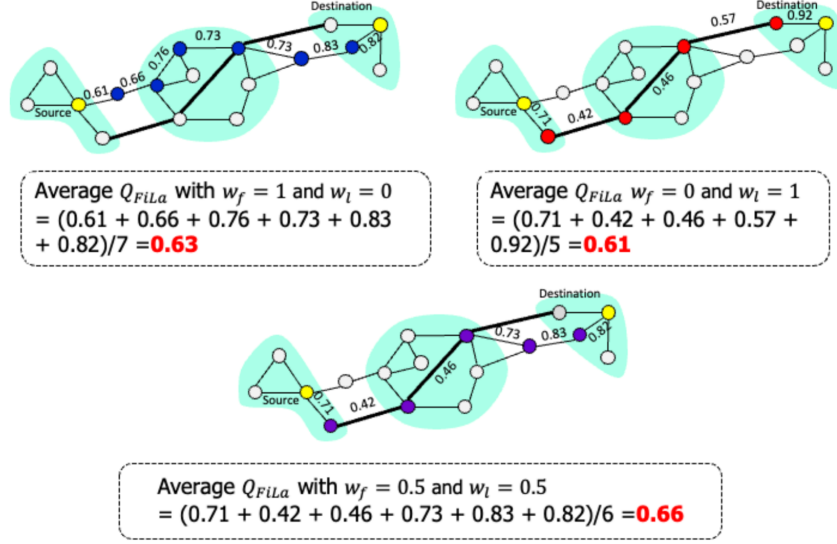


Figure 3.11: Q_{FiLa} Choice in Quantum Network Topology

the lowest latency, which are depicted by the bold lines. The aim is to optimize the speed of entanglement distribution, ensuring quick communication between quantum nodes.

3. **Bottom Case - Best Fidelity Combined with Latency:** The third case demonstrates the combined approach, where the Q_{FiLa} metric incorporates both fidelity and latency into a single routing decision. The algorithm balances the trade-off between achieving high fidelity and minimizing latency, selecting links that provide the best compromise for both performance metrics.

In all three cases, the routing choices reflect the use of the Q_{FiLa} metric to ensure the best overall performance, as evidenced by the optimal combination of fidelity and latency across the network topology.

3.4 Summary

This chapter tackles the challenge of balancing fidelity and latency in quantum communication networks. Quantum networks face an inherent trade-off, where improving fidelity often results in higher latency and vice versa. The chapter emphasizes the need for a routing strategy that accounts for both parameters simultaneously to optimize quantum network performance.

The issue is illustrated with the fidelity-latency trade-off, where enhancing

fidelity can lead to increased latency, and minimizing latency may compromise the required fidelity for quantum operations. Existing algorithms primarily focus on either fidelity or latency, but not both. To address this, a novel composite metric, Q_{FiLa} , is introduced, combining fidelity and latency to guide routing decisions in dynamic quantum networks.

The chapter also introduces the Nested Purification Protocol (NPP) for entanglement distribution and its impact on latency. A detailed analysis of the entanglement generation time, transmission delays, and quantum memory storage time reveals how these factors influence overall network latency. Numerical studies show how entanglement success probability and memory coherence time affect the achievable rate and latency.

Chapter 4

Evaluation Studies and Discussions

4.1 Evaluation Overview

This chapter presents the evaluation of the proposed Q_{FiLa} link metric integrated into the Q-LEAP routing algorithm. We compare its performance against the baseline Q-LEAP under various fidelity-latency trade-off settings. The evaluation spans real-world and synthetic topologies, enabling a comprehensive assessment of performance trends, trade-offs, and scalability.

4.1.1 Simulation Setup and Parameters

To thoroughly assess the effectiveness of the proposed Q_{FiLa} metric, we conducted extensive simulations on a machine equipped with an Intel i7-4790 CPU (3.6 GHz), 32 GB of RAM, and running 64-bit Windows 10.

We considered two categories of network topologies for evaluation: To comprehensively evaluate the proposed routing strategies, this study considers two types of quantum network topologies: the **Japan Photonic Network Model (JPNM)** and synthetically generated **random topologies**. This dual approach allows benchmarking against real-world constraints while testing scalability and adaptability in abstract, generalized networks.

4.1.1.1 Random Topology Generation

We generate random topologies to generalize findings and test routing performance across varying network sizes. Random topologies are constructed for node counts of $\{25, 50, 75, 100\}$, with each node assigned a unique position within a 100×100 2D coordinate grid. The Euclidean distance between nodes i and j is computed as:

$$\text{dist}_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (4.1)$$

To emulate physical distance in kilometers, a linear scale factor s is applied:

$$d_{ij} = s \cdot \text{dist}_{ij} \quad (4.2)$$

The default setting is $s = 5$, chosen to align the average link distance in the random 50-node topology with that of JPNM for comparison. Additionally, multiple density configurations are tested with $s \in \{1, 2, 3, 4, 5\}$ for the 50-node case to analyze the impact of link density on routing behavior.

Connectivity is ensured by linking each node to its $k = 3$ nearest neighbors, forming bidirectional links. This k -nearest neighbor approach yields sparse yet connected networks, which approximate quantum repeater constraints where long-distance links are both costly and lossy.

This random generation framework enables controlled experimentation across topological scales and densities, providing a valuable counterpart to the fixed structure of the JPNM.

4.1.1.2 Japan Photonic Network Model (JPNM)

The Japan Photonic Network Model (JPNM) [1] is a logical network topology that represents inter-prefecture connectivity across Japan. It was designed to support research, development, and commercialization efforts in photonic and quantum networks. The construction of JPNM is based on publicly available regional datasets—including station coordinates, railway networks, routing paths, and population data—combined with standard communication network design principles. As a result, JPNM exhibits structural characteristics that closely resemble real-world optical backbone networks deployed by major Japanese telecommunications carriers. It specifies concrete parameters such as node locations, physical routes, and link distances, making it a practical and realistic benchmark for quantum network simulations.

4.1.2 Experiment Scenarios

To comprehensively evaluate the effectiveness of the proposed Q_{FiLa} link metric integrated into the Q-LEAP routing framework, we define a series of experiment scenarios. These scenarios are designed to measure key performance metrics under diverse network conditions, routing preferences, and topology configurations. Each scenario targets a specific aspect of the fidelity-latency tradeoff and its impact on overall routing performance.

- **Scenario 1 – Link Density Sensitivity Analysis (Random Topology, 50 nodes):** First and foremost, we investigate how differ-

ent network densities affect routing performance. Using a fixed network size of 50 nodes, we vary the distance scaling factor $s \in \{1, 2, 3, 4, 5\}$, which controls link density. This helps evaluate the sensitivity of Q_{FiLa} to topology sparsity and connectivity.

- **Scenario 2 – Performance of Q-LEAP with and without Q_{FiLa} in Random Topologies (25 to 100 Nodes):** Next, we will evaluate the generalizability of the proposed Q_{FiLa} metric beyond real-world topologies by simulation random topology of sizes 25, 50, 75, and 100 nodes. This scenario assesses how Q_{FiLa} performs under varying network scales and structural randomness.
- **Scenario 3 – Fidelity-Latency Tradeoff Visualization:** Consequently, we visualize how fidelity and latency evolve across different link distances using the Japan Photonic Network Model (JPNM). This experiment aims to highlight the inherent tradeoff and motivates the necessity of a composite metric such as Q_{FiLa} .
- **Scenario 4 – Performance of Q-LEAP with and without Q_{FiLa} in JPNM:** Finally, this scenario replicates the comparison of scenario 2 on a realistic topology which is JPNM. We evaluate performance in terms of entanglement throughput, routing latency, and computation time under three weighting configurations:
 - ($w_f = 1, w_l = 0$) — **Fidelity-Focused Configuration:** Prioritizes the quality of entangled links over speed. This setting is suitable for applications requiring extremely high quantum state fidelity, such as *quantum teleportation*, *distributed quantum computing*, and *device-independent quantum key distribution (DI-QKD)*, where even small fidelity losses can render results invalid or insecure.
 - ($w_f = 0.5, w_l = 0.5$) — **Balanced Configuration:** Strikes a trade-off between entanglement quality and delivery latency. This setting is appropriate for *multi-user quantum networks*, where moderate fidelity and responsiveness are both desirable, such as in *quantum conferencing*, *entanglement-based synchronization*, or when network resources are constrained and shared.
 - ($w_f = 0, w_l = 1$) — **Latency-Focused Configuration:** Optimizes for minimum delay in entanglement delivery, even if fidelity is moderate. This is ideal for *time-critical quantum applications* such as *real-time QKD in dynamic environments*, *quantum sensing with low coherence time*, or systems with limited quantum memory where long delays risk decoherence.

These scenarios provide a well-rounded assessment of both fidelity-aware and latency-sensitive routing strategies across realistic and synthetic quantum network settings.

4.2 Experiment Results

4.2.1 Scenario 1: Link Density Sensitivity Analysis

To evaluate the impact of physical link density on routing performance, we conduct a sensitivity analysis by varying the scale factor in a 50-node random topology. The topology is constructed using a k -nearest neighbor model with $k = 3$, and the scale factor $s \in \{1, 2, 3, 4, 5\}$ determines the spatial spread of the nodes in a 100×100 unit grid. As described in Section 4.1.1.1, the physical distance between two nodes i and j is calculated using a scaled Euclidean formula as above.

This experiment uses the Q-LEAP routing protocol enhanced with the proposed Q_{FiLa} metric under the balanced setting ($w_f = 0.5$, $w_l = 0.5$). This configuration jointly considers fidelity and latency in path selection and is designed to maintain an optimal trade-off between entanglement quality and distribution speed.

Figure 4.1 illustrates how the average link distance grows with the scale factor. Table 4.1 summarizes key statistics including the mean, minimum, maximum, and standard deviation of distances at each scale.

Table 4.1: Distance Statistics Across Scale Factors

Scale Factor	Mean (km)	Min (km)	Max (km)	Std Dev (km)
1	51.43	1.62	121.68	24.74
2	102.87	3.26	243.35	49.48
3	154.30	4.88	365.03	74.21
4	205.74	6.52	486.68	98.95
5	257.17	8.16	608.36	123.69

As expected, the average link distance increases linearly with the scale factor, nearly doubling with each increment. At $s = 1$, most links are relatively short (average ≈ 51.4 km), but by $s = 5$, the average link stretches to over 250 km. This has several implications:

- **Quantum Fidelity Impact:** Longer links are more prone to fidelity degradation due to exponential attenuation. As link lengths grow, the probability of successful entanglement generation sharply decreases, potentially invalidating many long-distance connections.

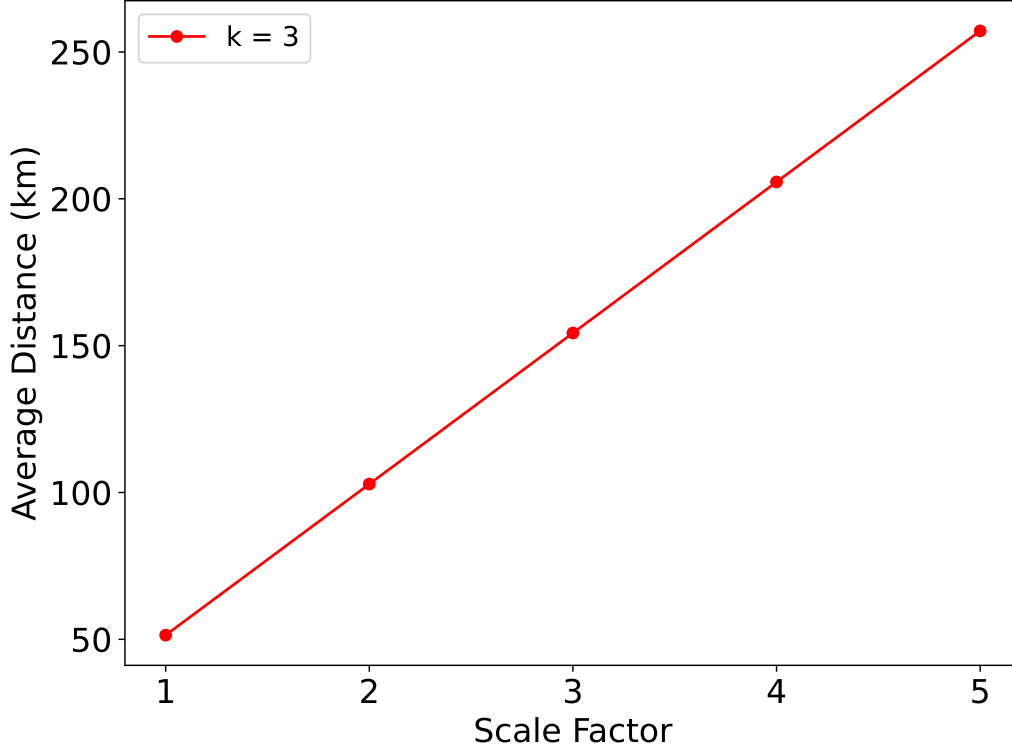


Figure 4.1: Average link distance vs. scale factor in a 50-node k -nearest neighbor topology with $k = 3$ under Q-LEAP with Q_{FiLa} ($w_f = 0.5$, $w_l = 0.5$).

- **Routing Constraints:** At higher scale factors, fewer links may satisfy the fidelity threshold, reducing the available routing paths. This can lead to increased multihop routes or routing failure under strict fidelity requirements.
- **Latency Tradeoff:** While more spread-out topologies increase physical latency, they may also force the use of additional purification stages, compounding the end-to-end delay.
- **Topology Sparsity and Density:** At smaller scale factors, the topology becomes denser, with many short links ensuring multiple valid routing options and greater robustness. At large scale factors, the network becomes sparse and thin, resembling more of a long-haul structure, which is more challenging for quantum networking protocols.

Overall, these results confirm that scale factor plays a significant role in shaping routing performance. Denser configurations (e.g., $s = 1$ or $s = 2$) are advantageous for high-fidelity, low-latency routing. In contrast, sparser

networks (e.g., $s = 4$ or $s = 5$) introduce harsher trade-offs and stricter path feasibility conditions. The analysis validates the robustness of the Q_{FiLa} -based Q-LEAP design in adapting to such topological diversity while maintaining efficiency.

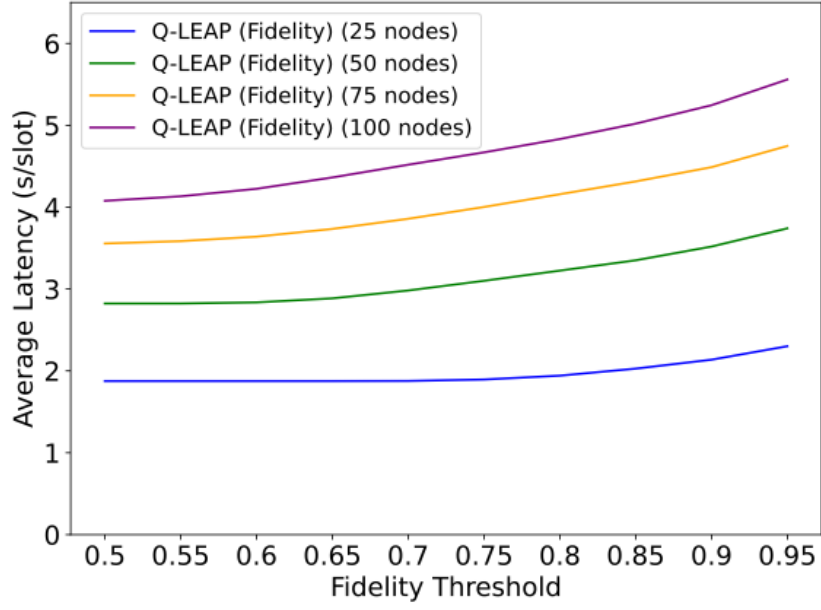
4.2.2 Scenario 2: Performance of Q-LEAP with and without Q_{FiLa} in Random Topologies

To assess the generalizability of Q_{FiLa} beyond structured real-world networks, we simulate our evaluation to synthetic random topologies with varying sizes of $\{25, 50, 75, 100\}$ nodes. These topologies are generated using a k -nearest neighbor construction with $k = 3$, where each node connects to its three nearest neighbors based on scaled Euclidean distance. A uniform scale factor of $s = 5$ is applied to preserve average link lengths similar to those found in JPNM, thereby enabling fair comparisons. Each simulation setup mirrors that of the previous scenarios, with routing evaluated under the balanced configuration of Q_{FiLa} using weights $w_f = 0.5$ and $w_l = 0.5$.

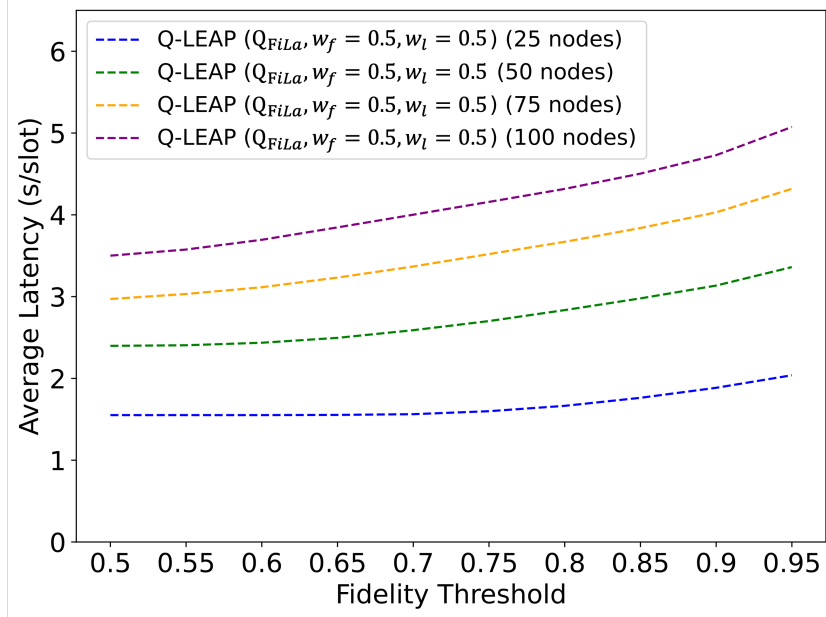
The detailed simulation parameters in the random topologies are summarized in Table 4.2.

Table 4.2: Simulation Parameters for Random Topologies

Parameter	Random Topology
Number of nodes	25, 50, 75, 100
Number of links	Varies by topology and k
Distance between nodes	Scaled Euclidean distance
Quantum link capacity	50 qubits/slot
Number of requests	50
Attenuation length (L_{att})	100 km
Decoherence length (L_{dec})	500 km
Speed of light in fiber	2×10^8 m/s
Entanglement success probability	0.25
Quantum storage time	500 ms
Fidelity threshold	0.55–0.95
Number of trials	100,000
Nearest neighbors (k)	3
Distance scale factor (s)	1, 2, 3, 4, 5



(a)



(b)

Figure 4.2: Average latency vs. fidelity threshold in random topologies (25–100 nodes): (a) Original Q-LEAP; (b) Q-LEAP with Q_{FiLa} , $w_f = 0.5$, $w_l = 0.5$.

Average Latency in Random Topologies: Figures 4.2a and 4.2b present the average latency across varying fidelity thresholds for both the original Q-LEAP and the Q-LEAP enhanced with the Q_{FiLa} metric. As expected, latency increases consistently with higher fidelity thresholds in all network configurations. This trend reflects the impact of stringent fidelity requirements, which limit the set of viable links and force the routing algorithm to select longer, multihop paths involving more purification operations.

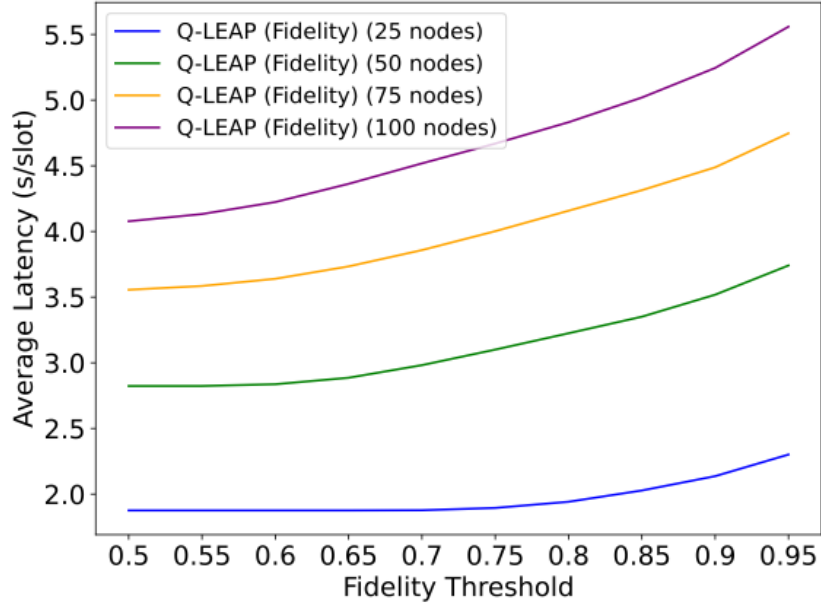
Among the topologies considered, the 25-node networks exhibit the lowest latency, benefiting from a small network diameter and limited path diversity. In contrast, the 100-node networks consistently incur the highest latency due to their larger diameter and more extensive routing paths. Notably, the Q-LEAP configuration augmented with Q_{FiLa} outperforms the original Q-LEAP in latency across all network sizes. This advantage becomes more pronounced as the network size increases, indicating Q_{FiLa} 's superior capacity to efficiently balance fidelity and latency even in sparse and irregular topologies.

Path Computation Time in Random Topologies: The path computation time for both routing configurations is shown in Figure 4.3, evaluated across varying topology sizes and fidelity thresholds. As the fidelity threshold increases, both Q-LEAP and Q-LEAP with Q_{FiLa} exhibit longer computation times. This trend is attributed to the shrinking set of feasible links under stricter fidelity constraints, which increases the search complexity for viable end-to-end paths.

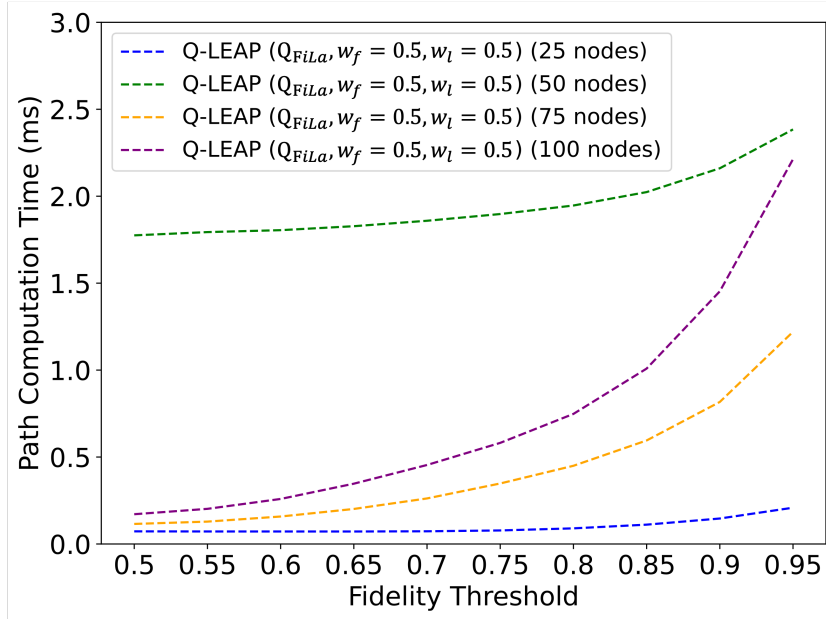
For the original Q-LEAP configuration (Figure 4.3a), the computation time scales predictably with network size. The 25-node topology incurs the lowest computation time, followed sequentially by 50, 75, and 100 nodes. This linear trend reflects the fact that larger topologies inherently introduce more routing candidates and increase Dijkstra-style search complexity.

In contrast, Q-LEAP using Q_{FiLa} (Figure 4.3b) displays a non-linear pattern: while the 25-node case still achieves the shortest computation time, the order becomes $25 < 75 < 100 < 50$ nodes. This deviation from linear scaling arises due to Q_{FiLa} 's fidelity-latency tradeoff mechanism, which actively prunes the topology based on threshold constraints and link quality. In some topologies (e.g., 100 nodes), this pruning results in sparser effective graphs, reducing the number of candidates and accelerating pathfinding.

Conversely, in the 50-node case, higher link density and less aggressive pruning may enlarge the search space, leading to unexpectedly higher computation times.

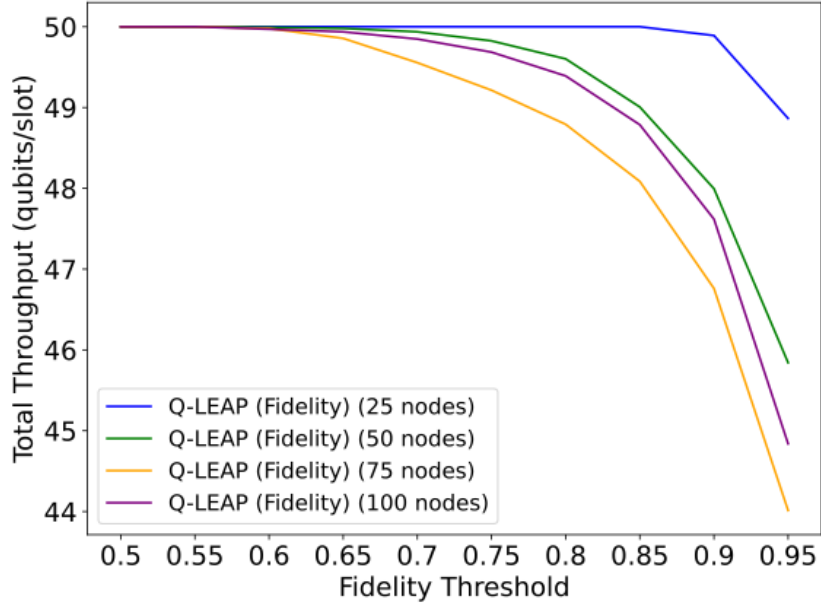


(a)

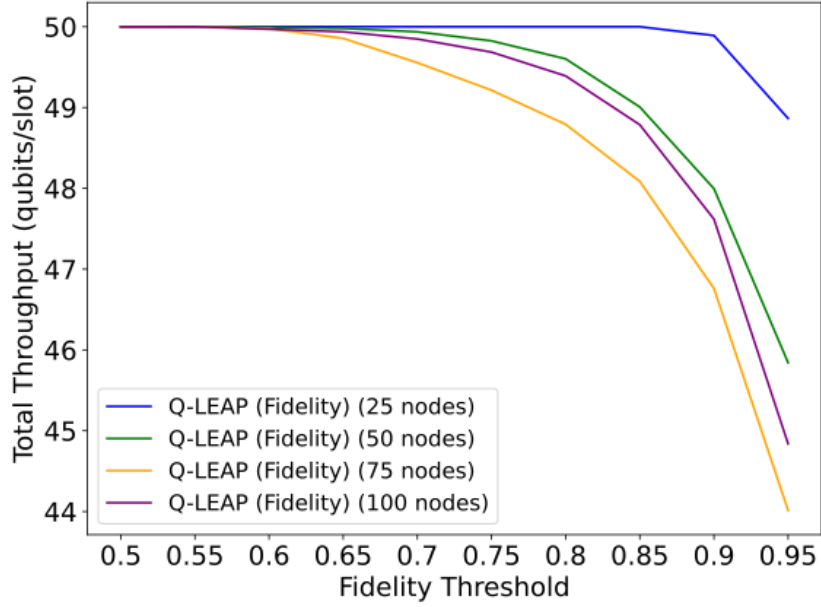


(b)

Figure 4.3: Path computation time across fidelity thresholds in random topologies of 25–100 nodes: (a) Original Q-LEAP; (b) Q-LEAP with Q_{FiLa} ($w_f = 0.5, w_l = 0.5$).



(a)



(b)

Figure 4.4: Total throughput vs. fidelity threshold in random topologies (25–100 nodes): (a) Original Q-LEAP; (b) Q-LEAP with Q_{FiLa} , $w_f = 0.5$, $w_l = 0.5$.

Throughput Analysis in Random Topologies: Figure 4.4 illustrate the total throughput (in qubits per time slot) under varying fidelity thresholds, across all topology sizes. In both cases, throughput remains high and nearly saturated when fidelity requirements are moderate (e.g., $f_{th} \leq 0.75$), but begins to degrade more noticeably at higher thresholds. This degradation occurs as stricter fidelity constraints filter out longer or noisier links, shrinking the number of feasible paths and thus limiting end-to-end entanglement delivery.

The 25-node topology consistently yields the highest throughput, followed by the 50-node and 100-node topologies, with the 75-node case lagging behind. This non-linear ordering stems from differences in link density caused by random placement and the fixed scale factor. Specifically, the 75-node configuration exhibits relatively sparse connectivity, leading to fewer redundant paths and ultimately lower throughput.

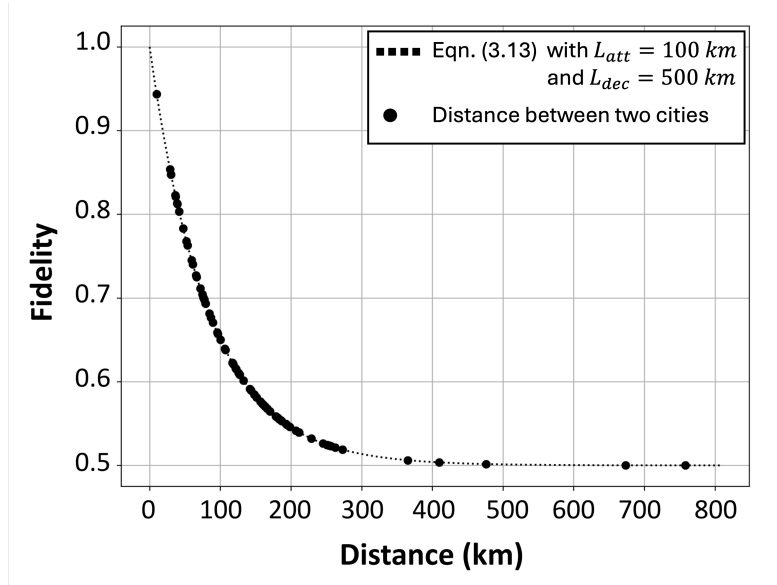
When comparing Q-LEAP with and without Q_{FiLa} , both routing strategies exhibit nearly identical throughput profiles across all network sizes and fidelity thresholds. This observation confirms that the latency and computation improvements achieved by Q_{FiLa} do not compromise its ability to sustain high throughput.

The similarity of throughput curves reinforces Q_{FiLa} 's suitability for quantum applications requiring high entanglement generation rates while also benefiting from reduced delay and improved scalability.

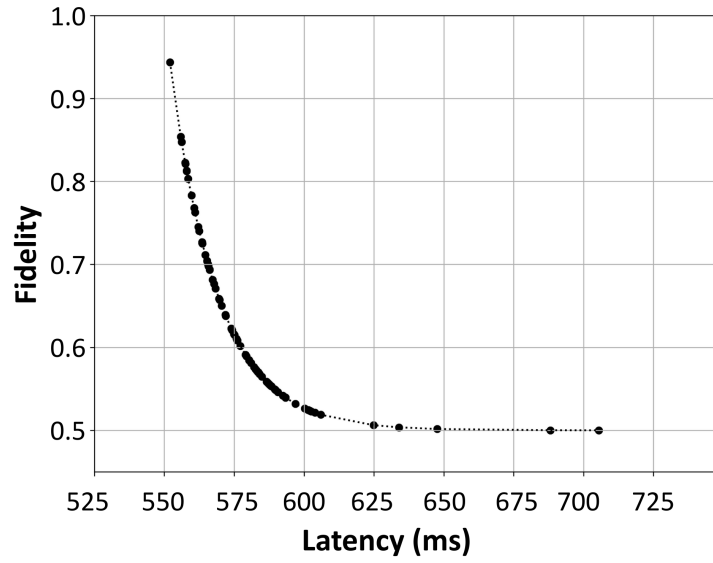
4.2.3 Scenario 3: Fidelity-Latency Trade-off Visualization

The fundamental trade-off between fidelity and latency is visualized in Figure 4.5b, based on simulation results from the Japan Photonic Network Model (JPNM). As the link distance increases, fidelity declines sharply due to attenuation and decoherence, while paths that achieve higher fidelity often suffer greater delays due to operational overheads and route complexity (Figure 4.5a(a)). In contrast, achieving higher fidelity typically leads to increased latency due to additional purification rounds and the use of longer multihop paths (Figure 4.5b(b)).

These observations motivate the development of the Q_{FiLa} metric, which aims to balance these competing objectives by combining them into a tunable weighted score.



(a) Fidelity vs. distance



(b) Fidelity vs. latency

Figure 4.5: Visualization based on JPNM: (a) Fidelity vs. distance and (b) Fidelity vs. latency.

4.2.4 Scenario 4: Performance of Q-LEAP with and without Q_{FiLa} in JPNM

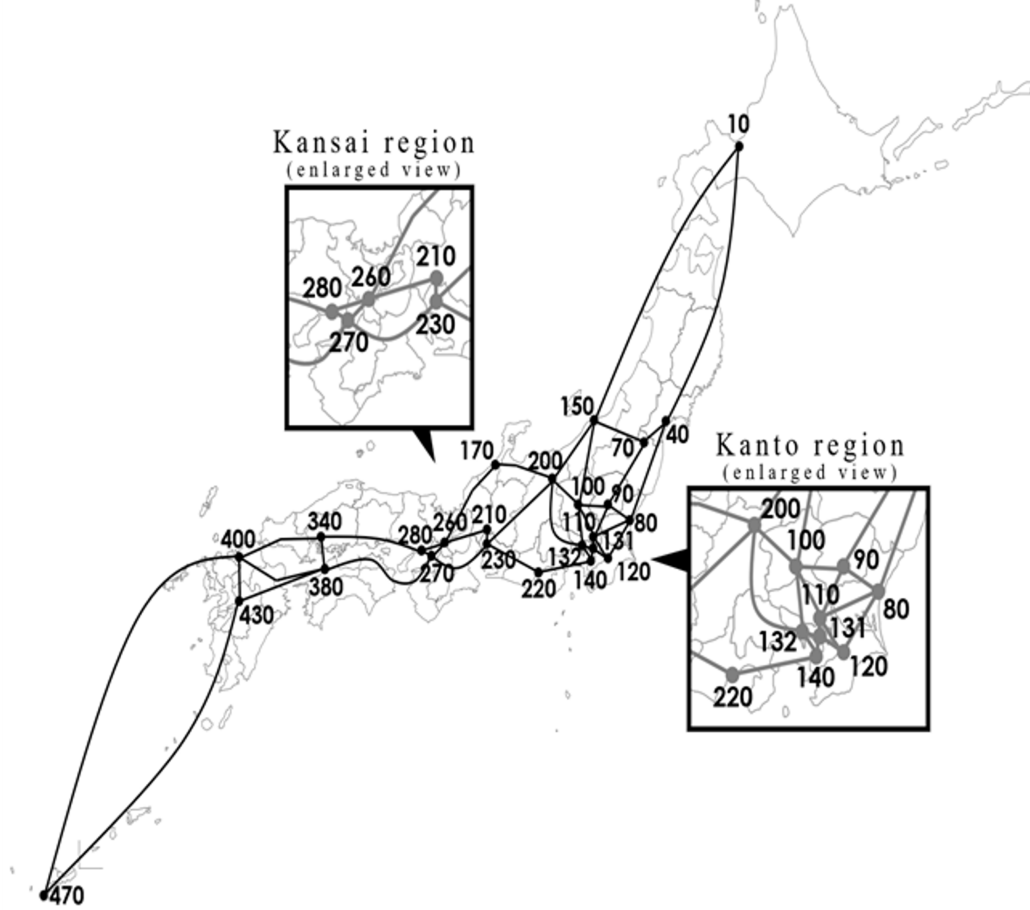


Figure 4.6: Illustration of the Japan Photonic Network Model (JPNM) [1].

After the simulation of the random topology, we evaluate by using a more smaller and realistic topology by conducting simulations on the Japan Photonic Network Model (JPNM), which is illustrated in Figure 4.6 represents a geographically informed network with 48 nodes and 82 links, each link distances range from 10 km to 673 km, showing a realistic showcase of a real-world topology. As shown in the setting in Table 4.3, the fidelity threshold f_{th} was varied from 0.55 to 0.95 to examine the effect of stricter entanglement quality requirements. The comparison includes four routing configurations: the original Q-LEAP algorithm, and Q-LEAP enhanced with the proposed Q_{FiLa} metric under three weighting modes—fidelity-focused

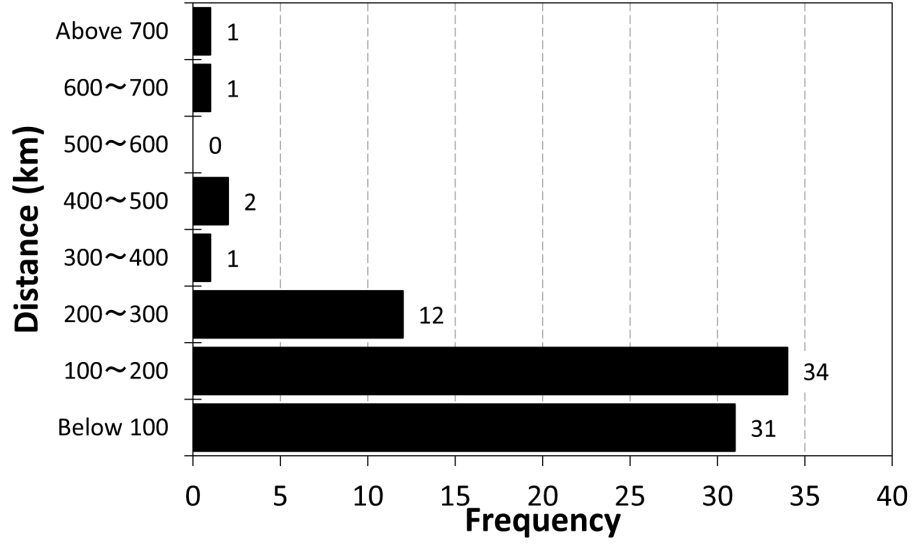
($w_f = 1.0$, $w_l = 0.0$), balanced ($w_f = 0.5$, $w_l = 0.5$), and latency-focused ($w_f = 0.0$, $w_l = 1.0$).

Table 4.3: Simulation Parameters for JPNM Topology

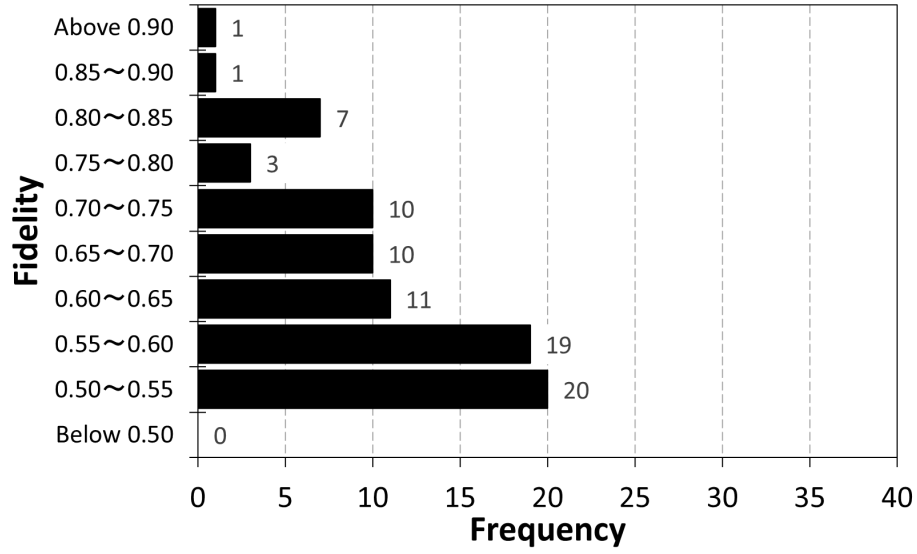
Parameter	Value (JPNM)
Number of nodes	48
Number of links	82
Distance between nodes	10–673 km
Quantum link capacity	50 qubits/slot
Number of requests	50
Attenuation length (L_{att})	100 km
Decoherence length (L_{dec})	500 km
Speed of light in fiber	2×10^8 m/s
Entanglement success probability	0.25
Quantum storage time	500 ms
Fidelity threshold	0.55–0.95
Number of trials	100,000

Figure 4.7 and Figure 4.8 shows the distribution of 82 simulated links across a 48-node network in the JPNM. Most links are below 200 km, with few exceeding 500 km, resulting in a majority of fidelity values ranging between 0.5 and 0.75, and latencies mostly below 600 ms. These empirical observations reinforce the inverse relationship between fidelity and distance, and the direct correlation between latency and both distance and purification complexity.

Figure 4.9a presents the total throughput performance across varying fidelity thresholds. While all configurations achieve comparable throughput at low thresholds, the balanced Q_{FiLa} configuration closely tracks the original Q-LEAP, slightly outperforming the fidelity-only and latency-only settings across most conditions. Notably, both extreme configurations (fidelity- or latency-focused) exhibit decreased throughput under stricter fidelity thresholds due to their one-sided optimization strategies. This underscores the importance of balancing fidelity and latency to maintain usable entanglement rates in heterogeneous network environments.

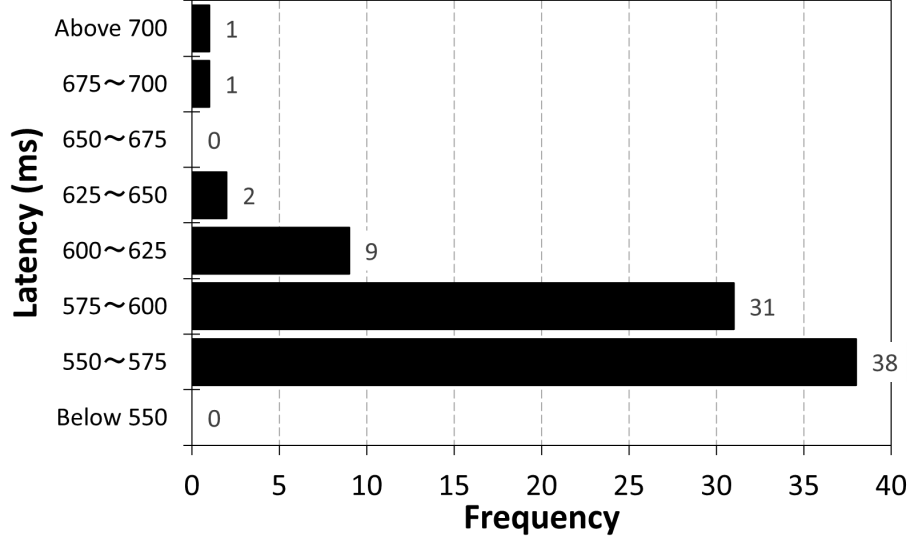


(a) Link distance distribution



(b) Fidelity distribution

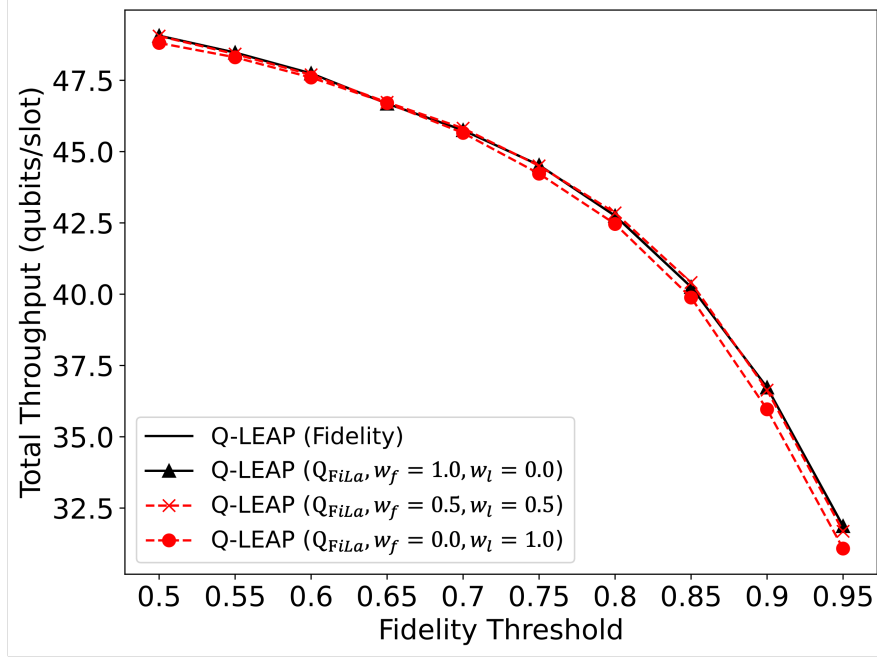
Figure 4.7: Distribution of the 82 quantum links in JPNM: (a) Link distances and (b) Fidelity.



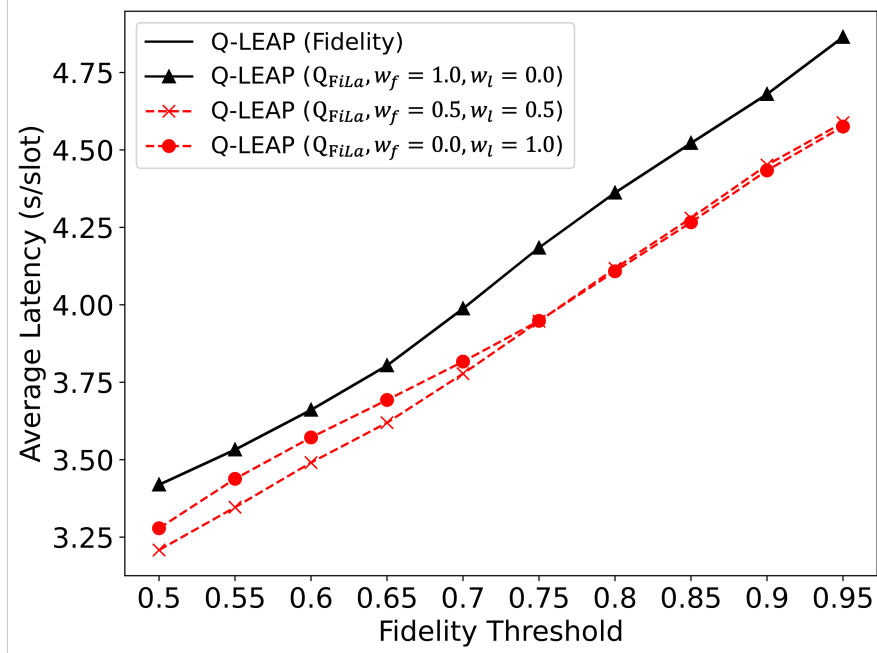
(a) Latency Distribution

Figure 4.8: Latency Distribution of the 82 quantum links in JPNM.

Figure 4.9b depicts the average latency across the same threshold range. The latency-focused configuration ($w_f = 0.0$, $w_l = 1.0$) consistently achieves the lowest latency, especially in the range $0.55 \leq f_{th} \leq 0.75$, by preferring shorter, low-hop paths even if fidelity is modest. In contrast, the balanced configuration trades slightly higher latency for fidelity resilience, avoiding low-quality links that require excessive purification. This balance results in performance close to the optimal case across the full fidelity range. The original Q-LEAP and the fidelity-focused Q_{FiLa} configuration incur noticeably higher latency, particularly at high fidelity thresholds, due to longer paths and more intensive purification. These results demonstrate the sensitivity of latency to path quality, purification overhead, and hop count—all factors modulated effectively by Q_{FiLa} .

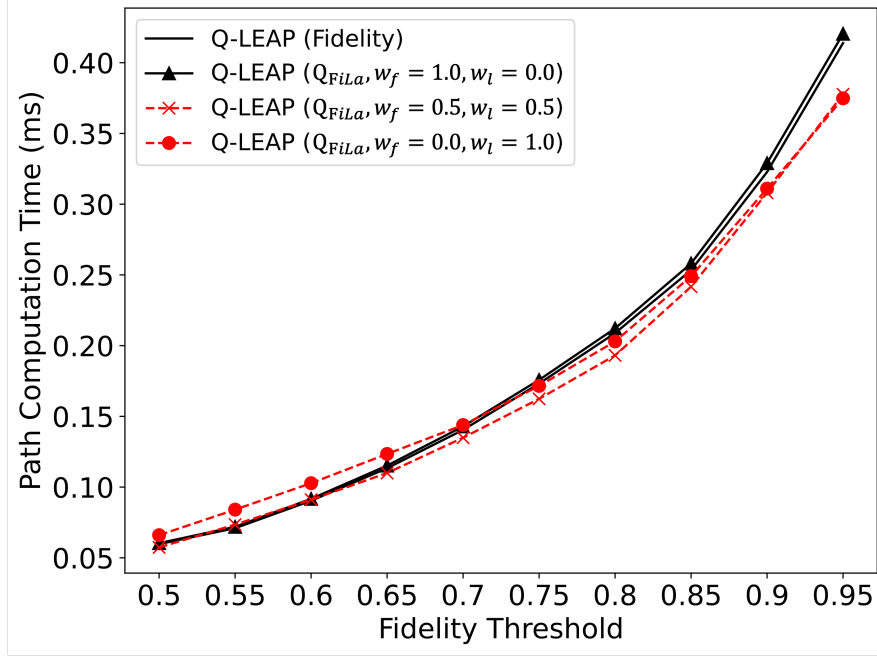


(a) Total throughput



(b) Average latency

Figure 4.9: Performance comparison on JPNM under different fidelity thresholds: (a) Total throughput and (b) Average latency.



(a) Path computation time

Figure 4.10: Path computation time performance comparison on JPNM under different fidelity thresholds.

Figure 4.10a shows the average path computation time. Here, both the latency-focused and balanced Q_{FiLa} variants achieve lower computation times compared to the original Q-LEAP. This is because their metric design allows for early pruning of unqualified paths during Dijkstra traversal, reducing search complexity. In contrast, the fidelity-focused mode incurs the highest computation cost due to the algorithm’s effort to identify rare, high-fidelity paths over long distances. This highlights a key advantage of the Q_{FiLa} formulation—by explicitly incorporating both latency and fidelity, it simplifies route evaluation and improves computational efficiency.

In summary, the results on JPNM validate the efficacy of the Q_{FiLa} metric. While throughput remains comparable across configurations, significant gains in latency and computation time are achieved using the balanced and latency-aware variants. These improvements make Q_{FiLa} -enhanced routing a practical choice for quantum network applications where delay sensitivity and scalability are critical, such as in entanglement-based sensing, real-time QKD, or distributed quantum computing platforms.

4.3 Summary

This chapter evaluates the performance of the proposed Q_{FiLa} link metric integrated with the Q-LEAP routing algorithm, focusing on its effectiveness in balancing the fidelity-latency trade-off. The evaluation is conducted on two network topologies: the Japan Photonic Network Model (JPNM) and random topologies with varying node counts. The results demonstrate how Q_{FiLa} performs across different network sizes and fidelity thresholds, compared the Q-LEAP with and without the using of Q_{FiLa} .

The evaluation reveals that Q_{FiLa} significantly improves routing efficiency by balancing fidelity and latency. Scenarios highlight the trade-offs between throughput, latency, and path computation time. The Q_{FiLa} metric achieves lower latency and faster computation times in most configurations, especially under stringent fidelity conditions. The experimental results confirm that Q_{FiLa} can be effectively applied to real-world and synthetic networks, offering scalable, efficient routing for quantum applications.

Furthermore, the sensitivity analysis of link density in random topologies shows that denser configurations lead to better performance in terms of fidelity and latency, validating the robustness of the Q_{FiLa} -based Q-LEAP design across different network densities.

Chapter 5

Conclusions and Outlook

This chapter presents a concise summary of the research contributions and key findings from this thesis. It reflects on the performance and scalability of the proposed Q_{FiLa} metric across diverse quantum network scenarios. In addition, it outlines the current limitations of the study and discusses promising directions for future research aimed at improving quantum routing efficiency and adaptability in real-world deployments.

5.1 Summary of Findings

In this thesis, we proposed Q_{FiLa} —a fidelity-latency-aware link metric designed to enhance quantum routing performance by balancing the tradeoff between entanglement fidelity and communication latency. Integrated into the Q-LEAP routing framework, Q_{FiLa} allows for tunable path selection via weighted scoring, enabling more flexible and efficient decision-making under varying network conditions.

To validate the effectiveness of this metric, we conducted extensive simulations across both realistic and synthetic quantum topologies. Specifically, we utilized the randomly generated topologies with sizes ranging from 25 to 100 nodes, in the addition of Japan Photonic Network Model (JPNM), which emulates the physical infrastructure of Japanese topology networks. These topologies were evaluated under diverse scenarios and fidelity thresholds (0.55–0.95), with performance measured in terms of throughput, latency, and path computation time.

Our findings reveal that Q_{FiLa} delivers consistently improved latency and path computation performance compared to the baseline Q-LEAP, while preserving comparable throughput. In high-fidelity regimes, Q_{FiLa} significantly reduces latency and computational overhead by prioritizing near-optimal links and reducing search complexity through early pruning. This is particularly beneficial in large-scale networks, where the routing overhead can become a major bottleneck.

Moreover, Q_{FiLa} demonstrates robust adaptability across varying network

scales and densities. Through link density sensitivity analysis in 50-node random topologies with scale factors from 1 to 5, we observed consistent behavior of Q_{FiLa} in response to average physical link lengths and connectivity. The metric performs well even in irregular and sparse topologies, reinforcing its generalizability.

Overall, this research highlights the practical benefits of fidelity-latency joint optimization in quantum networks and establishes Q_{FiLa} as a promising direction for scalable, quality-aware quantum routing protocols. The metric’s tunability and efficiency make it particularly suited for real-world applications such as quantum key distribution (QKD), quantum internet backbone design, and distributed quantum computing environments.

5.2 Study Limitations

Despite the promising results achieved by the proposed Q_{FiLa} -based routing strategy, several limitations remain in the current study:

- **Static Topology Assumption:** All simulations were conducted assuming a fixed network topology. Dynamic changes in node/link status—common in real-world quantum networks—were not considered.
- **Simplified Link Model:** The model assumes homogeneity in quantum link characteristics, such as fixed capacity and fidelity decay functions. In practical scenarios, link quality may vary due to hardware heterogeneity, environmental interference, or hardware failure.
- **Ideal Hardware Assumptions:** Parameters such as entanglement success probability and storage time were chosen based on optimistic assumptions or literature defaults. These may not reflect real system limitations such as decoherence, loss, or memory errors.
- **No Entanglement Purification Overhead:** While latency metrics account for purification-related delays, the impact of purification failure rates, resource contention, and scheduling were not explicitly modeled in the simulations.
- **Focus on Unicast Communication:** The evaluation focused solely on single source-destination pairs per request. Group communication scenarios such as multicast or broadcast entanglement were not explored.
- **Limited Experimental Topologies:** Although both realistic (JPNM) and synthetic (random) topologies were examined, other network types (e.g., grid, ring, scale-free) and real deployment traces were not included.

- **Evaluation by Simulation Only:** The results are validated purely through simulation. Practical implementation and deployment of Q_{FiLa} on quantum hardware or quantum simulators are left for future work.

5.3 Directions for Future Work

While this thesis lays a strong foundation for fidelity- and latency-aware routing in quantum networks through the Q_{FiLa} metric, several promising avenues remain for future exploration and enhancement:

- **Integration of Quantum Error Management:** Future work will focus on incorporating key quantum functionalities such as quantum error correction, entanglement purification, and entanglement swapping. These mechanisms are essential to enhance reliability and extend communication distances in practical quantum networks, and their integration will provide a more holistic view of end-to-end performance.
- **Extension to Diverse Network Scenarios:** The current study evaluates performance on the JPNM and random topologies with up to 100 nodes. To improve generalizability, future experiments will include other structured topologies (e.g., grid, tree, mesh, scale-free) as well as realistic deployment traces and dynamic environments with node churn.
- **Dynamic Link Characteristics:** The present evaluation assumes constant values for entanglement success probability and entanglement storage time $T(N)$. Future work will involve dynamically computing these parameters based on physical-layer conditions, routing distance, and real-time channel feedback, leading to a more realistic performance model.
- **Multi-flow and Concurrent Routing:** This study focuses on single unicast requests per trial. A natural extension is to evaluate the performance of Q_{FiLa} under concurrent flows and background traffic, analyzing fairness, contention resolution, and throughput sustainability under load.
- **Hybrid Classical-Quantum Routing Architectures:** A potential direction is to investigate the integration of classical control plane coordination (e.g., SDN-based controllers) with quantum entanglement management. This could enable more adaptive and programmable routing decisions in quantum networks.
- **Hardware-in-the-Loop Simulation:** Moving beyond software simulations, future research can involve hybrid simulations or hardware-in-the-loop platforms that combine quantum network simulators (e.g.,

NetSquid) with actual quantum hardware or emulators to better assess feasibility and hardware bottlenecks.

- **Machine Learning for Adaptive Weight Tuning:** Finally, the static assignment of weights (w_f, w_l) in Q_{FiLa} may be enhanced using reinforcement learning or metaheuristic techniques to adaptively learn optimal trade-offs based on network state, user demand, or service level agreements.

These directions aim to bridge the gap between theoretical models and real-world deployment, further advancing the robustness, adaptability, and scalability of quantum network routing.

List of Publications

- [1] Thu Trang Nguyen, Yuto Lim, and Ruidong Li, “Study of latency-distance analysis in long-distance quantum networks,” *IEICE Technical Committee on Information Networks (IN)*, vol. 124, no. 420, IN2024-95, pp. 92–97, 7-8 March 2025.
- [2] Yuto Lim, Zhaowei Zhong, Jianwen Sun, Thu Trang Nguyen, and Ruidong Li, “QFide: Quantum teleportation fidelity simulator for developing quantum networks,” *International Conference on Quantum Communications, Networking, and Computing (QCNC)*, Nara, Japan, 31 March-2 April 2025, pp. 693–695, doi: 10.1109/QCNC64685.2025.00116.
- [3] Thu Trang Nguyen, Yuto Lim, and Ruidong Li, “Novel quantum fidelity and latency link metric for entanglement routing design in quantum networks,” *The 52nd Quantum Information Technology Symposium (QIT52)*, Shizuoka University, 28-30 May 2025.
- [4] Thu Trang Nguyen, Yuto Lim, and Ruidong Li, “Novel quantum fidelity and latency link metric for entanglement routing design in quantum networks,” *Quantum Innovation 2025 (QI2025)*, Osaka, 29 July-2 August 2025. (to be presented)
- [5] Thu Trang Nguyen and Yuto Lim, “QFiLa: Entanglement-assured routing metric design in long-distance quantum networks,” *IEEE Network, Quantum Communications and Networking: Series 4*, 2025. (under review)

References

- [1] N. NEWS, “White paper on quantum network,” *NICT NEWS*, vol. 491, no. 1, 2022.
- [2] A. I. Nurhadi and N. R. Syambas, “Quantum key distribution (qkd) protocols: A survey,” in *2018 4th International Conference on Wireless and Telematics (ICWT)*, 2018, pp. 1–5.
- [3] T.-S. Lin, I.-M. Tsai, H.-W. Wang, and S.-Y. Kuo, “Quantum authentication and secure communication protocols,” in *2006 Sixth IEEE Conference on Nanotechnology*, vol. 2, 2006, pp. 863–866.
- [4] J. Shi and S. Shen, “A clock synchronization method based on quantum entanglement,” *Scientific Reports*, vol. 12, no. 1, p. 10185, 2022.
- [5] S. P. Neumann, A. Buchner, L. Bulla, M. Bohmann, and R. Ursin, “Continuous entanglement distribution over a transnational 248 km fiber link,” *Nature Communications*, vol. 13, no. 1, p. 6134, 2022.
- [6] R. Van Meter, *Quantum networking*. John Wiley & Sons, 2014.
- [7] D. Bruss and G. Leuchs, *Quantum information: from foundations to quantum technology applications*. John Wiley & Sons, 2019.
- [8] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, “Quantum repeaters: The role of imperfect local operations in quantum communication,” *Phys. Rev. Lett.*, vol. 81, pp. 5932–5935, Dec 1998. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.81.5932>
- [9] J. Li, M. Wang, K. Xue, R. Li, N. Yu, Q. Sun, and J. Lu, “Fidelity-guaranteed entanglement routing in quantum networks,” *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 6748–6763, 2022.
- [10] Z. Wang and J. Crowcroft, “Analysis of shortest-path routing algorithms in a dynamic network environment,” *ACM SIGCOMM Computer Communication Review*, vol. 22, no. 2, pp. 63–71, 1992.
- [11] S. M. Kumari and N. Geethanjali, “A survey on shortest path routing algorithms for public transport travel,” *Global Journal of Computer Science and Technology*, vol. 9, no. 5, pp. 73–76, 2010.

- [12] S. Lai and B. Ravindran, “Least-latency routing over time-dependent wireless sensor networks,” *IEEE transactions on computers*, vol. 62, no. 5, pp. 969–983, 2012.
- [13] W. K. Wootters, W. K. Wootters, and W. H. Zurek, “A single quantum cannot be cloned,” *Nature*, vol. 299, pp. 802–803, 1982. [Online]. Available: <https://api.semanticscholar.org/CorpusID:4339227>
- [14] B. M. Terhal, “The fragility of quantum information?” 2013. [Online]. Available: <https://arxiv.org/abs/1305.4004>
- [15] L. K. Grover, “The advantages of superposition,” *Science*, vol. 280, no. 5361, pp. 228–228, 1998.
- [16] Z. Li, K. Xue, J. Li, L. Chen, R. Li, Z. Wang, N. Yu, D. S. Wei, Q. Sun, and J. Lu, “Entanglement-assisted quantum networks: Mechanics, enabling technologies, challenges, and research directions,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2133–2189, 2023.
- [17] A. Abane, M. Cubeddu, V. S. Mai, and A. Battou, “Entanglement routing in quantum networks: A comprehensive survey,” *IEEE Transactions on Quantum Engineering*, 2025.
- [18] N. Zou, “Quantum entanglement and its application in quantum communication,” in *Journal of Physics: Conference Series*, vol. 1827, no. 1. IOP Publishing, 2021, p. 012120.
- [19] C. H. Bennett and G. Brassard, “Quantum cryptography: Public key distribution and coin tossing,” *Theoretical Computer Science*, vol. 560, p. 7–11, Dec. 2014. [Online]. Available: <http://dx.doi.org/10.1016/j.tcs.2014.05.025>
- [20] A. K. Ekert, “Quantum cryptography based on bell’s theorem,” *Phys. Rev. Lett.*, vol. 67, pp. 661–663, Aug 1991. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.67.661>
- [21] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, “Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels,” *Phys. Rev. Lett.*, vol. 70, pp. 1895–1899, Mar 1993. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.70.1895>

- [22] R. Van Meter, “Quantum networking and internetworking,” *IEEE Network*, vol. 26, no. 4, pp. 59–64, 2012.
- [23] P. Kómár, E. Kessler, M. Bishof *et al.*, “A quantum network of clocks,” *Nature Phys.*, vol. 10, pp. 582–587, Aug 2014. [Online]. Available: <https://doi.org/10.1038/nphys3000>
- [24] C. Hughes, J. Isaacson, A. Perry, R. F. Sun, and J. Turner, “What is a qubit?” in *Quantum computing for the quantum curious*. Springer, 2021, pp. 7–16.
- [25] J.-W. Pan, D. Bouwmeester, H. Weinfurter, and A. Zeilinger, “Experimental entanglement swapping: entangling photons that never interacted,” *Physical review letters*, vol. 80, no. 18, p. 3891, 1998.
- [26] J.-W. Pan, C. Simon, Č. Brukner, and A. Zeilinger, “Entanglement purification for quantum communication,” *Nature*, vol. 410, no. 6832, pp. 1067–1070, 2001.
- [27] W. J. Munro, K. Azuma, K. Tamaki, and K. Nemoto, “Inside quantum repeaters,” *IEEE Journal of Selected topics in quantum electronics*, vol. 21, no. 3, pp. 78–90, 2015.
- [28] K. Azuma, S. E. Economou, D. Elkouss, P. Hilaire, L. Jiang, H.-K. Lo, and I. Tzitrin, “Quantum repeaters: From quantum networks to the quantum internet,” *Rev. Mod. Phys.*, vol. 95, p. 045006, Dec 2023. [Online]. Available: <https://link.aps.org/doi/10.1103/RevModPhys.95.045006>
- [29] P. Kok, C. P. Williams, and J. P. Dowling, “Construction of a quantum repeater with linear optics,” *Phys. Rev. A*, vol. 68, p. 022301, Aug 2003. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevA.68.022301>
- [30] L. S. Martin and K. B. Whaley, “Single-shot deterministic entanglement between non-interacting systems with linear optics,” 2019. [Online]. Available: <https://arxiv.org/abs/1912.00067>
- [31] L. H. Pedersen, N. M. Møller, and K. Mølmer, “Fidelity of quantum operations,” *Physics Letters A*, vol. 367, no. 1-2, pp. 47–51, 2007.
- [32] F. Verstraete, K. Audenaert, and B. De Moor, “Maximally entangled mixed states of two qubits,” *Physical Review A*, vol. 64, no. 1, p. 012316, 2001.

- [33] A. L. Grimsmo, “Time-delayed quantum feedback control,” *Physical review letters*, vol. 115, no. 6, p. 060402, 2015.
- [34] S. Muralidharan, L. Li, J. Kim, N. Lütkenhaus, M. D. Lukin, and L. Jiang, “Optimal architectures for long distance quantum communication,” *Scientific reports*, vol. 6, p. 20463, 2016. [Online]. Available: <https://doi.org/10.1038/srep20463>
- [35] W. Dai, T. Peng, and M. Z. Win, “Optimal remote entanglement distribution,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 3, pp. 540–556, 2020.
- [36] Y. Zhao and C. Qiao, “Redundant entanglement provisioning and selection for throughput maximization in quantum networks,” in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.
- [37] R. Zhou, Y. Gan, Y. Liu, K. Obraczka, and C. Qian, “Towards qos-aware quantum networks,” in *2024 International Conference on Quantum Communications, Networking, and Computing (Q CNC)*. IEEE, 2024, pp. 288–296.
- [38] K. S. Elsayed, W. R. KhudaBukhsh, and A. Rizk, “On the trade-off between fidelity and latency for the quantum link layer with few memories and entanglement purification,” in *2024 International Conference on Quantum Communications, Networking, and Computing (Q CNC)*. IEEE, 2024, pp. 17–24.
- [39] A. Dahlberg and S. Wehner, “Simulaqron—a simulator for developing quantum internet software,” *Quantum Science and Technology*, vol. 4, no. 1, p. 015001, 2018.
- [40] M. Razavi, M. Piani, and N. Lütkenhaus, “Quantum repeaters with imperfect memories: Cost and scalability,” *Phys. Rev. A*, vol. 80, p. 032301, Sep 2009.
- [41] M. Pant, H. Krovi, D. Englund, and S. Guha, “Rate-distance tradeoff and resource costs for all-optical quantum repeaters,” *Phys. Rev. A*, vol. 95, p. 012304, Jan 2017.
- [42] K. Azuma, K. Tamaki, and H.-K. Lo, “All photonic quantum repeaters,” *Nature Commun.*, vol. 6, 09 2013.