International Journal of Management Science and Engineering Management

ISSN: 1750-9653 (Print) 1750-9661 (Online) Journal homepage: www.tandfonline.com/journals/tmse20

Designing a resilient cloud network fulfilled by quantum machine learning

Erfan Shahab & Sharareh Taghipour

To cite this article: Erfan Shahab & Sharareh Taghipour (11 Aug 2025): Designing a resilient cloud network fulfilled by quantum machine learning, International Journal of Management Science and Engineering Management, DOI: 10.1080/17509653.2025.2544566

To link to this article: https://doi.org/10.1080/17509653.2025.2544566

INTERNATIONAL JOURNAL OF MANAGEMENT SCIENCE AND ENGINEERING MANAGEMENT
https://doi.org/10.1080/17509653.2025.2544566

Designing a resilient cloud network fulfilled by quantum machine learning

Erfan Shahab and Sharareh Taghipour

Department of Mechanical, Industrial and Mechatronics Engineering, Toronto Metropolitan University, Toronto, Canada

ABSTRACT

Next-generation digital services require resilient, energy-conscious cloud networks, but current optimization techniques are unable to quickly reconfigure infrastructures when a failure occurs. To address real-time service migration, this paper presents a quantum machine learning (QML) architecture that concurrently maximize quality of service (QoS) and minimizes migration cost while taking capacity, energy, and jitter restrictions into account. Thousands of migration methods are evaluated in parallel using a parameterized quantum neural network to solve the model. In comparison to the genetic algorithm, the QML optimizer reduces peak CPU load by 45%, while maintaining contractual QoS during cyberattacks, according to an experiment conducted on a real case study. The quantum solution offers noticeably smoother resource use, according to several assessments. These results establish QML as a promising facilitator for responsive cloud resilience by proving that quantum search may unleash fault-tolerant reconfiguration that is not possible with classical methodologies. The deployment is limited to medium-sized networks due to the size and noise of current quantum hardware; however, implementing new error-mitigation strategies provides viable routes to commercial use. This study establishes a research agenda for scalable quantum optimization in resilient networks in digital infrastructures by combining quantum computing with cloud-network engineering.

ARTICLE HISTORY
Received 7 January 2025
Accepted 2 August 2025

KEYWORDS
Cloud networks; cybersecurity; quantum machine learning; quantum neural networks; resilience

JEL CLASSIFICATION
C63, L86, M15, O33

1. Introduction

Cloud computing has revolutionized the way computational resources are delivered, offering flexible, scalable, and on-demand services to meet the needs of a wide variety of applications (Huang et al., 2024). As cloud networks continue to expand, ensuring the resilience of these systems in the face of disruptions has become a critical challenge. Unforeseen events such as hardware failures, network congestion, or sudden surges in demand can severely impact the quality of service (QoS) of cloud infrastructures (d'Ambrosio et al., 2025). Traditional optimization methods for service composition, while effective, often struggle to adapt in real-time to such dynamic environments (Delaram et al., 2022). This necessitates the development of advanced optimization techniques that can enhance resilience by dynamically reallocating resources and migrating services as disruptions occur (Khan et al., 2024).

Quantum computing has emerged as a promising solution to many complex optimization problems, offering capabilities beyond what classical algorithms can achieve Kumar et al. (2023). By exploiting the principles of superposition and entanglement, quantum systems can process vast amounts of data in parallel, making them well-suited for high-dimensional optimization tasks (I. Gupta et al., 2024). Recent advancements in quantum machine learning (QML) have opened new possibilities for optimizing cloud networks, particularly in improving resilience through service migration. These methods can provide faster, more efficient solutions by exploring multiple optimization paths simultaneously (Peral-García et al., 2024).

Cloud networks often encounter disruptions such as cyberattacks resulting in significant degradation of service quality, and increased energy consumption. Traditional optimization methods typically struggle to adapt in real time to these unpredictable scenarios, limiting their effectiveness in maintaining resilient cloud infrastructures (Shahab et al., 2024). To address these challenges, this paper proposes a novel optimization model that integrates QML techniques to dynamically manage service migration strategies. Unlike conventional optimization methods, the proposed framework efficiently evaluates multiple migration options simultaneously, optimizing real-time service reallocation decisions to maintain high QoS while minimizing energy consumption during disruptions. The primary contributions of this research include the introduction of a quantum-driven service migration optimization framework, the formulation of an objective function explicitly accounting for QoS and energy costs, and the demonstration of the model's capability to significantly enhance cloud network resilience.

2. Literature review

Recent studies have significantly contributed to the development of resilient optimization strategies for both supply chains and cloud networks. Harkat et al. (2024) emphasized the need for resilience in cyber-physical systems to withstand and recover from cyber threats. They explored strategies like machine learning-based intrusion detection systems, which are designed to enhance CPS resilience by detecting and mitigating attacks. Real-time monitoring is critical to ensuring that models perform effectively over time, especially as dynamic and heterogeneous cloud environments introduce uncertainties and potential drifts in data distributions (Malinovskaya et al., 2024).

CONTACT Erfan Shahab erfan.shahab@torontomu.ca Department of Mechanical, Industrial and Mechatronics Engineering, Toronto Metropolitan University, Toronto, Canada

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17509653.2025.2544566

Many scholars have developed mathematical models to fulfill a resilient network. Y. Wang et al. (2024) proposed a robust resilience optimization strategy that uses a resilience indicator to mitigate uncertain disruptions through a mixed-integer linear programming approach. Li et al. (2024) introduced a hypernetwork model that effectively enhances network resilience by optimizing supplier selection, thus minimizing disruptions. Similarly, Moein Fazeli et al. (2024) proposed a deep reinforcement learning approach for cloud service composition, which addresses the complexity of dynamic service allocation in real-time environments. Tang et al. (2024) extended this by proposing a service composition allocation method that prioritizes critical subtasks in cloud networks, ensuring optimal resource utilization. The integration of service migration strategies, as highlighted in their work, underscores the importance of adaptive methodologies in both cloud and manufacturing systems.

Wan et al. (2023) addressed the challenge of scheduling in cloud environments with multi-composite tasks. Their hierarchical scheduling model divides the process into user-level and sublevel tasks, allowing a more efficient matching of providers and demanders. The firefly genetic algorithm they propose is effective in balancing cost, time, and quality, optimizing cloud manufacturing service composition and resource management. A complementary study by (Arbabi et al., 2023) integrated configuration design and capacity planning into a dynamic cloud manufacturing system. This work proposed a multi-objective model that maximizes the utility of stakeholders while addressing the challenges of changing service providers and fluctuating customer demands. The authors introduced a Grey Wolf Optimizer (DMOGWO) to optimize platform profit, equity, and customer satisfaction, providing a comprehensive framework for dynamic capacity planning.

Several studies explored multi-objective optimization to address the trade-offs between different stakeholders in cloud manufacturing (Sharifisari et al., 2025). For example, Gao et al. (2023) proposed a tri-objective service composition model that balances the interests of customers, cloud service platforms, and providers. Using an enhanced Jellyfish Search Optimizer, the study optimized service quality, sustainability, and cooperation, illustrating the effectiveness of this method through computational experiments. Building on this, Zhang et al. (2024) tackled the uncertainties inherent in cloud manufacturing with a robust service composition model. Their Enhanced Multi-Objective Artificial Hummingbird Algorithm (EMOAHA) efficiently managed task delays and alternative service switches, optimizing the system's robustness under uncertain conditions. This work demonstrated a significant improvement over traditional optimization methods in handling convergence and solution diversity.

Recently, many scholars have developed valuable mathematical models to improve resilience to encounter pandemic disruptions. Ivanov (2022) integrated agility, resilience and sustainability perspectives to think beyond COVID-19 pandemic. Azadi et al. (2023) used network data envelopment analysis to assess the resilience of healthcare supply chains in response to the COVID-19 pandemic. In response to the COVID-19 pandemic, Shahab et al. (2023) offered a real-world application of cloud manufacturing in crisis conditions. They proposed a resilient cloud network designed to recover disrupted systems. The model leveraged redundant resources from various supply networks demonstrating the importance of resilience in cloud systems. Shahab et al., (2024) investigated the use of RL techniques to optimize cloud network resilience in response to disruptions. Their model improved network adaptability, validated by a case study on ventilator production during the COVID-19 pandemic.

Fuzzy-based optimization has been employed in cloud service composition to handle uncertainties in service quality and availability. H. Wang et al., (2024) proposed a fuzzy-based Particle Swarm Optimization (PSO) algorithm for cloud service composition, which dynamically adapts to changes in service availability. This approach optimized response time, cost, and scalability, and outperforming conventional service composition techniques.

As the Internet of Things expands, optimizing service composition in cloud networks becomes increasingly important. Vakili et al. (2024) introduced a service composition method using Grey Wolf Optimization (GWO) in a cloud-based IoT environment. By integrating the MapReduce framework, their method significantly improved energy efficiency, availability, and cost-effectiveness in service composition for IoT systems. Similarly, in the healthcare sector, cloud-based systems are being explored for monitoring chronic conditions. Sharma et al. (2023) developed a cloud service composition model for diabetes monitoring, using machine learning techniques such as Extreme Learning Machine (ELM) and Principal Component Analysis (PCA). Their system achieved high accuracy and scalability, particularly beneficial for rural healthcare applications.

The collective body of research highlights significant advancements in cloud manufacturing service composition, with various algorithms demonstrating improved performance in terms of efficiency, robustness, and scalability. From reinforcement learning and fuzzy-based PSO to parallel differential evolution approaches, these studies contribute to addressing the complexities of service composition in dynamic, multi-stakeholder environments. These models and algorithms offer promising avenues for further exploration in cloud-based manufacturing and service optimization across different industries. To clarify the contributions of this paper in relation to prior research, Table 1 highlights and contrasts the methodologies, outcomes, limitations, and innovations presented here compared to the existing studies.

While existing literature has made significant strides in optimizing cloud service composition using a range of classical techniques, such as evolutionary algorithms, reinforcement learning, and fuzzy-based methods, the current paper advances this field by introducing quantum machine learning and quantum neural networks as novel approaches for service composition in resilient cloud networks.

Unlike the traditional algorithms such as Particle Swarm Optimization or Grey Wolf Optimization, which rely on classical computing, this paper leverages the quantum neural networks' ability to handle high-dimensional and complex data structures. This enables more efficient and faster convergence when optimizing QoS and energy consumption, outperforming classical methods in terms of both time complexity and solution quality.

While previous works have proposed resilience models, such as service redundancy during crisis robust composition under uncertainty (Zhang et al., 2024), this paper uniquely integrates quantum-based service migration. This enables

Table 1. Summary of the literature review of the resilient cloud networks.

Paper	Focus	Algorithm/Model	Key Contributions
(Yin et al., 2023)	Cloud service composition in aviation	Enhanced Carnivorous Plant Algorithm (ECPA)	Improves resource allocation and collaboration between manufacturers and suppliers in aviation
(Wan et al., 2023)	Hierarchical scheduling for cloud manufacturing	Firefly Genetic Algorithm	Optimizes cost, time, and quality in cloud-based task scheduling
(Arbabi et al., 2023)	Dynamic cloud manufacturing configuration	Discrete Multi-Objective GWO (DMOGWO)	Maximizes platform utility, balancing profit, equity, and customer satisfaction
(Gao et al., 2023)	Service composition balancing stakeholder interests	Enhanced Jellyfish Search Optimizer	Optimizes service quality, sustainability, and cooperativity
(Shahab et al., 2023)	Resilient cloud manufacturing during crises	Redundancy and collaboration model	Enhances production recovery using diverse supply networks (e.g. ventilator production)
(Sharma et al., 2023)	Cloud-based healthcare monitoring	Machine Learning (ELM and PCA)	Provides accurate monitoring for chronic conditions, particularly in rural settings
(Zhang et al., 2024)	Robust service composition under uncertainty	Enhanced Multi-Objective Artificial Hummingbird Algorithm (EMOAHA)	Improves robustness and handles task delays effectively
(Shahab et al., 2024)	Cloud network resilience	RL (SAC, TD3, PPO)	Optimizes cloud networks for resilience during disruptions (e.g. ventilator case study)
(H. Wang et al., 2024)	Dynamic cloud service composition	Fuzzy-based PSO	Improves flexibility, response time, cost, and scalability in cloud environments
(Vakili et al., 2024)	Service composition for IoT	GWO with MapReduce	Enhances energy efficiency, cost, and availability in large-scale service composition
(H. Wang et al., 2024)	Service composition optimization	PSO with prior knowledge	Reduces search space and improves convergence for service composition in cloud
Current Paper	Quantum-based service composition in resilient cloud networks	QNN and QML for Resilient Cloud Networks	Optimizes QoS and energy consumption with quantum service migration for improved resilience in cloud networks

a proactive, dynamic adjustment of services within the cloud network, ensuring minimal energy consumption during network disruptions. This enhances the resilience and adaptability of the network beyond what traditional algorithms can achieve.

In contrast to previous methods that often trade off energy consumption against QoS, this work simultaneously optimizes both metrics using quantum machine learning techniques. By harnessing the quantum capabilities to explore larger solution spaces more efficiently, the model achieves a superior balance between energy efficiency and service quality in resilient cloud networks.

3. Proposed methodology

In this section, the framework for enhancing cloud network resilience through service migration-based optimization is discussed. The methodology addresses gaps identified in prior research and offers an adaptive approach to maintaining service continuity during disruptions. We outline the core elements of the system, including the modeling of service migration strategies, network parameters, and the constraints that drive the optimization process. This framework leverages service migration techniques to improve resource allocation and maintain quality of service, ensuring that the cloud network remains robust and responsive to dynamic conditions.

The proposed model is developed based on several underlying assumptions that define its applicability and limitations. It assumes that each service can be represented as a computational workload characterized by measurable QoS attributes such as energy consumption, jitter, and latency, and that tasks can be divided and migrated across compatible servers without loss of functionality. The model relies on the availability of accurate and real-time monitoring data, meaning that the state of the network – including service demands, server capacities, and system constraints – is fully observable during each optimization cycle. It is also assumed that disruptions, such as cyberattacks or node failures, are detected promptly and accurately, with minimal detection delays or false alarms.

During each optimization window, network parameters such as energy profiles, task demands, and bandwidth availability are assumed to remain static, allowing the quantum optimization algorithm to operate under quasi-steady conditions. Server failures are considered independent events; the model does not explicitly account for cascading failures or correlated disruptions across the network. Furthermore, due to current hardware constraints, the model is limited to medium-sized networks where noise and qubit count in quantum devices do not impede performance significantly. Any error mitigation is applied externally and is not embedded in the optimization formulation itself. Lastly, the cost functions used to balance energy consumption and QoS are assumed to reasonably approximate the true trade-offs observed in practice. These assumptions support the tractability and effectiveness of the proposed model while also highlighting areas where caution is needed when applying it to highly dynamic or large-scale cloud environments.

3.1. Proposed framework

The primary problem addressed in this research is ensuring resilience in cloud networks when disruptions occur, such as cyberattacks. These disruptions can significantly degrade the QoS, overload resources, increase energy consumption, and impact on the reliability of cloud services. To tackle this issue, we model the problem as a service migration optimization task, where the objective is to allocate cloud services efficiently in real-time to maintain performance standards while minimizing migration and operational costs.

The problem is modeled mathematically as a multi-objective optimization task with two primary goals: maximizing QoS and minimizing the cost associated with migrating services during disruptions. The optimization model incorporates multiple constraints, including server capacity, maximum allowable power consumption, acceptable jitter

thresholds, and minimum QoS standards. These constraints ensure realistic and effective service reallocation.

The QNN-based approach leverages quantum superposition and entanglement to evaluate numerous potential migration strategies concurrently, significantly reducing computational complexity and decision-making latency compared to classical methods. A QML framework to solve this optimization problem is proposed. The detailed steps for implementing the proposed framework are as follows:

Service Composition: Initial configuration selects and allocates cloud resources based on QoS and resource constraints to meet user demands.
Disruption Detection: Real-time monitoring detects service disruptions (e.g. server failures, network congestion) necessitating migration.
Load Balancing and Migration Decision: Tasks identified as disrupted are redistributed across available resources to ensure balanced load and minimal QoS degradation.
QNN Solution Generation: A QNN is utilized to generate candidate service allocation solutions rapidly, exploiting quantum parallelism to evaluate multiple configurations simultaneously.
Migration Cost Evaluation: The model assesses migration costs (e.g. energy consumption, performance degradation) associated with each candidate solution.
Softmax Surrogate Optimization: A softmax surrogate function optimizes candidate solutions, balancing the trade-off between resilience and cost.
Refinement via Quantum Machine Learning: QNN parameters are iteratively adjusted based on cost and QoS outcomes, refining migration solutions towards optimal resilience.
Optimal Configuration Selection: The best-performing migration strategy, determined through QML-driven optimization, is selected and implemented.
Post-Migration Performance Evaluation: System performance is evaluated post-migration to confirm effectiveness, followed by continuous resilience monitoring to detect future disruptions.
Resilience Feedback Loop: Continuous monitoring results inform future iterations, ensuring adaptive, real-time resilience improvements.

By clearly defining and modeling the problem, explicitly outlining the QML-based optimization approach, and detailing the implementation strategy, this framework robustly addresses cloud network resilience challenges in dynamic operational environments.

3.2. Network parameters

In this subsection, the key parameters and features of the cloud network, essential for ensuring resilience through service migration, are presented. These parameters play a crucial role in the performance and adaptability of the network when responding to disruptions. The focus will be on the factors influencing the migration of services between servers to maintain network QoS. These variables reflect the interaction between computational resources and tasks, which are critical to the system's overall resilience.

As outlined in Table 2, these parameters provide the basis for the service migration strategy. They describe the limitations of server capacities, task demands, and the ability of the network to reassign services during disruptions. Understanding how these parameters interact will guide the optimization process, ensuring that services can be migrated seamlessly, thus maintaining QoS while minimizing the impact of failures. The interaction of migration processes provides the foundation for implementing a resilient cloud infrastructure.

3.3. Network modelling

The mathematical model aims to optimize cloud network resilience by balancing two main objectives: QoS through efficient task allocation and resource management, and minimizing the costs associated with service migrations during disruptions. The model includes constraints to ensure all tasks are completed, server capacities are respected, energy consumption remains within acceptable limits, and network jitters stay below specified thresholds. Additional constraints ensure minimum QoS standards are maintained while preventing severe performance degradation. The migration strategy specifically addresses network resilience by enabling continuous and optimal performance through effective service relocation and load balancing. Variables defining QoS compliance, migration decisions, and load distribution are included to support practical implementation.

$$\mathrm{Z} = \max \sum_{\mathrm{s}}^{\mathrm{S}} \sum_{\mathrm{t}}^{\mathrm{T}} \left(\mathrm{H_{st}} \odot \mathrm{QoS_{st}^{+}}\right) - \left(\mathrm{H_{st}} \odot \mathrm{QoS_{st}^{-}}\right) \quad (1)$$

$$Z_2 = \min \sum_{t \in T} \sum_{s \in S} MC_{st} \odot M_{st} \quad (2)$$

Table 2. Indexes, parameters and variables.

S	Total number of services	DS	Data security matrix
T	Total number of tasks	MC	Migration Cost
s	Service index	QoS^{+}	Positive criteria matrix which is $[DS, E]_{st}$
t	Task index	QoS^{-}	Negative criteria matrix which is $[PC, JI]_{st}$
d	Disrupted task index	Max_N	Maximum acceptable negative criteria
D	Set of Disrupting tasks	Min_P	Minimum acceptable positive criteria
Ψ_s	Capacity of the service network	E_{st}	Energy use of assigning server s to task t
Φ_t	Demand of the task network	Ln	Large number
PC	Normalized Power Consumption Matrix	M_{st}	Migration variable
J	Maximum acceptable jitter of the network	G_{st}^{+}	Variable ensuring having least accepted QoS^{+}
JI	Normalized Jitter matrix	G_{st}^{-}	Variable preventing violation of accepted QoS^{-}
P	Maximum acceptable power consumption	H_{st}	Variable showing the amount of load on servers
SCs	Spare Capacity of Server s	θ	Vector of trainable parameters in the quantum
$U(\theta)$	Unitary transformation applied to quantum state	$\|\psi\rangle$	Quantum state vector representing encoded input

s.t.

$$\sum_{s}^{S} H_{st} = \Phi_t, \quad \forall t \in T \tag{3}$$

$$\sum_{t}^{T} H_{st} \leq \Psi_s, \quad \forall s \in S \tag{4}$$

$$\sum_{s}\sum_{t} PC_{st} \cdot H_{st} \leq P \tag{5}$$

$$\sum_{s}\sum_{t} JI_{st} \cdot H_{st} \leq J \tag{6}$$

$$H_{st} \leq LnG_{st}^{+}, \quad \forall s \in S, t \in T \tag{7}$$

$$QoS_{st}^{+} + Ln - LnG_{st}^{+} \geq Min_P, \quad \forall s \in S, t \in T \tag{8}$$

$$H_{st} \leq LnG_{st}^{-}, \quad \forall s \in S, t \in T \tag{9}$$

$$QoS_{st}^{-} - Ln + LnG_{st}^{-} \leq Max_N, \quad \forall s \in S, t \in T \tag{10}$$

$$\sum_{t\in T} H_{st} \leq \Psi_s + \sum_{t\in T} M_{st} \cdot SC_s \tag{11}$$

$$\sum_{s} M_{st} \geq 1 \quad \forall t \in T \tag{12}$$

$$\sum_{s\in S} \left| H_{st} - H_{\text{avg}} \right| \leq \delta \tag{13}$$

$$M_{st} \in \{0, 1\} \tag{14}$$

$$G_{st}^{+} \& G_{st}^{-} \in \{0, 1\} \tag{15}$$

$$H_{st} \in W \tag{16}$$

The objective function in Equation (1) is designed to maximize QoS, focusing on optimal resource allocation and task distribution, while the objective function in Equation (2) tries to minimize the migration costs. Equation (3) ensures the completion of all tasks within the system, a critical factor in maintaining network reliability. Equation (4) imposes server capacity constraints, guaranteeing that the total workload assigned to each server does not exceed its processing capabilities. Equation (5) adds a constraint on the maximum allowable power consumption, ensuring that the network operates efficiently within energy limits. Equation (6) addresses jitter control, ensuring that the jitter remains below the maximum level specified in the service level agreement, thus preserving consistent network performance. Equations (7) and (8) establish that the minimum positive QoS threshold is met for all tasks, ensuring baseline quality. Equations (9) and (10) set limits on the maximum negative QoS, preventing substantial performance degradation that could impact user satisfaction. Equations (11) and (12) focus on the migration strategy that handles disruptions, enhancing the network's resilience by ensuring continuous optimal performance. Equation (13) tries to balance the load on the disrupted servers by load balancing strategy. Finally, Equations (14), (15) and (16) define QoS control and migration variable as binary variables, and load distribution as positive integers.

3.4. The proposed QNN

QNNs are quantum machine learning models that merge the principles of quantum mechanics with neural network architectures to address complex optimization problems, such as ensuring resilience in cloud networks (Golchha et al., 2025). In this section, the mathematical formulations behind QNNs and their application in optimizing service composition for cloud networks is presented. QNNs begin by encoding classical input data into quantum states. For an input vector $x \in R^n$, the classical data is mapped to the quantum state $|x\rangle$ through an encoding function $\mathcal{E}(x)$, which is presented in Equation (17), where $|i\rangle$ represents the computational basis state and x_i are the components of the input vector (K. Gupta et al., 2024). This encoding allows the QNN to process multiple inputs in parallel, leveraging the power of quantum superposition.

$$|x\rangle = \mathcal{E}(x) = \sum_{i=0}^{2^n-1} x_i |i\rangle \tag{17}$$

A QNN is composed of layers of quantum gates (unitary transformations) applied to the input state. For a single layer, the transformation of the quantum state $|\psi\rangle$ can be described by a unitary operator $U(\theta)$, where θ are the trainable parameters of the network. After applying the gate to the input quantum state $|x$, the new quantum state results in Equation (18).

$$|\psi(\theta) = U(\theta)|x \tag{18}$$

For a multi-layer QNN, this process is repeated across several layers, with each layer applying a different unitary operator. If the QNN has L layers, the final quantum state is given in Equation (19), where $\theta = \{\theta_1, \theta_2, \ldots, \theta_L\}$ represents the set of all parameters for each layer.

$$\left|\psi_L(\theta) = U_L(\theta_L) U_{L-1}(\theta_{L-1}) \ldots U_1(\theta_1)|x\right\rangle \tag{19}$$

The unitary gates in each layer are parameterized by the angles α and are represented using rotation gates. The gates which are used in QNNs of this paper are given in Equations (20), (21) and (22). These gates are applied to individual qubits, and by adjusting the parameters α, the QNN learns to represent complex mappings between the inputs and outputs.

$$R_X(\alpha) = \begin{pmatrix} \cos(\alpha/2) & -i\sin(\alpha/2) \\ -i\sin(\alpha/2) & \cos(\alpha/2) \end{pmatrix} \tag{20}$$

$$R_Y(\alpha) = \begin{pmatrix} \cos(\alpha/2) & -\sin(\alpha/2) \\ \sin(\alpha/2) & \cos(\alpha/2) \end{pmatrix} \tag{21}$$

$$R_Z(\alpha) = \begin{pmatrix} e^{-i\alpha/2} & 0 \\ 0 & e^{i\alpha/2} \end{pmatrix} \tag{22}$$

The goal of the QNN is to optimize the service composition in cloud networks. To achieve this, a cost function $C(\theta)$ is defined to measure the difference between the predicted output of the QNN and the target solution. The cost function

used in QNNs of this paper is the mean squared error, which can be formulated as Equation (23), where y_i^{target} is the target solution, $y_i^{\text{predicted}}(\theta)$ is the predicted output generated by the QNN, and N is the number of samples. The objective is to minimize $C(\theta)$ by adjusting the parameter θ.

$$C(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\text{target}} - y_i^{\text{predicted}}(\theta)\right)^2 \tag{23}$$

Unlike classical neural networks, QNNs leverage quantum measurements to evaluate the gradients of the cost function. After the forward pass, where the input state is transformed through the quantum layers, the quantum state is measured to generate predictions. The parameters θ are updated using classical optimization algorithms, such as gradient descent based on the calculated gradients that are presented in Equation (24), where η is the learning rate and $_\theta C(\theta_t)$ represents the gradient of the cost function with respect to θ.

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta C(\theta_t) \tag{24}$$

In the context of resilient cloud networks, QNNs are utilized to generate optimized solutions for the service composition problem, ensuring that the network remains robust under dynamic conditions. The optimization problem is framed to balance energy consumption, server load, and task distribution, with constraints such as total task assignment which ensures all tasks t ∈ T are assigned to available servers s ∈ S. This is presented in Equation (25), where H_{st} is the amount of task t assigned to server s, and Φ_t is the demand of task t. Server capacity constraint ensures that the total workload assigned to each server does not exceed its processing capacity Ψ_s. This is presented in Equation (26). Energy consumption constraint that limits the total power consumption of the network is presented in Equation (27), in which P_{st} represents the power consumed when task t is assigned to server s. Load balancing ensures a balanced distribution of tasks across the network to prevent bottlenecks, shown in Equation (28), where H_{avg} is the average load and δ is the tolerance for load imbalance.

$$\sum_{s\in S} H_{st} = \Phi_t \quad \forall t \in T \tag{25}$$

$$\sum_{t\in T} H_{st} \le \Psi_s \quad \forall s \in S \tag{26}$$

$$\sum_{s\in S}\sum_{t\in T} P_{st}H_{st} \le P_{\max} \tag{27}$$

$$\sum_{s\in S}\left|H_{st} - H_{\text{avg}}\right| \le \delta \tag{28}$$

One of the key advantages of QNNs is their ability to explore the solution space efficiently through quantum parallelism. Unlike classical methods, where solutions are evaluated sequentially, QNNs use superposition to evaluate multiple potential solutions simultaneously, significantly reducing computational overhead. This is particularly beneficial in high-dimensional optimization problems such as cloud network resilience, where multiple constraints and objectives must be balanced in real-time.

During the training process, the QNN explores different task assignments and resource configurations, iteratively refining its parameters to converge to the optimal solution. The final output of the QNN is a set of assignments H_{st} that maximize QoS while adhering to energy and capacity constraints. By leveraging quantum properties, the QNN can generate high-quality solutions more efficiently than the classical methods.

4. Results

In this section, we discuss the findings from our study on service migration for enhancing resilience in cloud networks. The results are divided into two main parts: first, a case study that illustrates the practical implementation of the proposed service migration-based model, and second, a comprehensive analysis of the computational outcomes. These findings aim to demonstrate the effectiveness of the model in real-world scenarios and benchmark its performance against existing approaches in the field.

4.1. Case study

The case study involves 26 services and 14 tasks, where each task represents a specific user demand, and the services reflect the resources available to meet those demands. The primary goal of the service migration strategy is to ensure continued service delivery while optimizing resource allocation across the network.

The case study illustrates how the migration-based model addresses complex challenges, factoring in the constraints such as server capacity, task priority, and dynamic resource availability. The model's capacity to evaluate and implement various migration strategies in real time allows handling high-dimensional problems, which would typically overwhelm the classical optimization methods.

Further details on the network setup, task classifications, and service characteristics are provided in (Rezapour Niari et al., 2022). This section focuses on the practical application of the migration model in this scenario, showcasing its potential to enhance resource efficiency and ensure robust cloud network performance.

4.2. Computational results

The 3D surface plot shown in Figure 1 represents the allocation of tasks to services in the cloud network optimization model. The x-axis represents the tasks, the y-axis represents the services, and the z-axis indicates the magnitude of the task allocations. Each spike or peak in the plot corresponds to the allocation of a certain number of tasks to a particular service.

From the plot, certain services (such as service 9, handling tasks 7, 8, 9, and 10) are more heavily loaded with tasks compared to the others, as indicated by the taller peaks. This distribution reflects the result of the quantum-optimized model, which ensures resilience by balancing tasks across services. The variation in peak heights illustrates the differences in load allocation, with higher peaks representing higher allocations, while lower or absent peaks indicate non-assignment.

Figure 1 provides a visual confirmation of the load balancing and service assignment strategy. Services with higher task demands, like services 9 and 12, are efficiently allocated more tasks to ensure maximum resource utilization, which

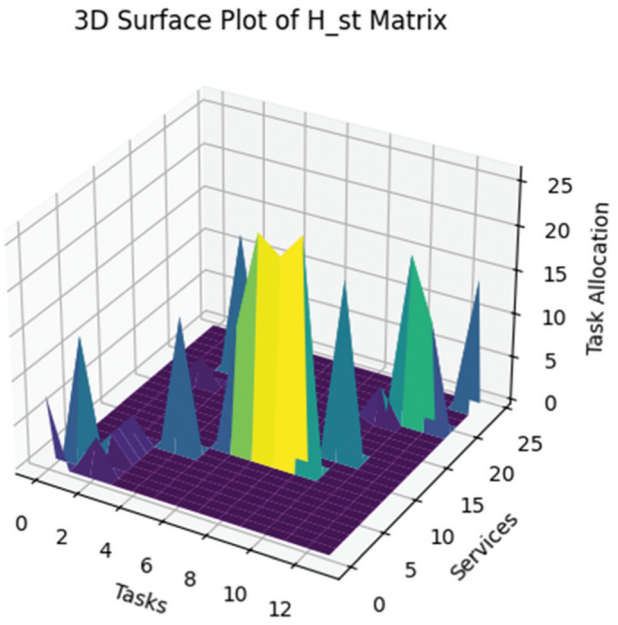

Figure 1. H_{st} matrix 3D plot.

enhance overall system resilience. Additionally, the plot highlights areas where certain services are left unassigned or minimally assigned, such as services 2 and 3, suggesting they may serve as backup resources or are preserved for future task migration in case of disruptions. This visual representation plays a critical role in understanding how the optimization framework dynamically distributes workloads to maintain service performance and prevent overloading specific resources in the cloud network.

To validate the performance and correctness of the QML model, we applied it to the case study described in Section 4.1, where CPU load was monitored and compared with a baseline Genetic Algorithm (GA) under the same conditions. The primary objective was to analyze the effectiveness of the QML model in optimizing CPU load and ensuring network resilience through service migration. GA efficiently optimizes by exploring multiple solution regions simultaneously, using stochastic, fitness-guided searches to avoid local optima. Their mutation and crossover techniques maintain diversity in the search process, making them effective for complex problems (Shojaee et al., 2024). The details of the implemented GA are presented in Appendix A. GA for Service Migration-Based Resilience in Cloud Networks.

Both the QML model and the GA were applied to evaluate CPU load distribution and its impact on cloud network performance, with 100-time steps simulated for each. The key objective was to assess whether both models could efficiently manage resource usage while maintaining the system resilience. As illustrated in Figure 2, the QML model demonstrated a more stable and consistent CPU load over time compared to the GA. The GA exhibited several high spikes, with its peak reaching over 25% CPU load, while the QML peaked around 12%. This highlights the more resource-intensive nature of the GA, as it intermittently demanded significantly higher computational power during certain phases of the simulation.

To further validate these findings, both models were run across many iterations, and their peak CPU loads were compared. Figure 3 shows that the GA reached a high peak of CPU load, while the QML model remained considerably lower. This difference in peak load emphasizes the greater resource efficiency of the QML model, which consistently required fewer CPU resources than the GA, even at its most demanding stages.

Additionally, a box plot of CPU load distribution over the 100 runs (Figure 4) revealed the variability in CPU consumption. While both models had similar median CPU loads near zero, the GA exhibited far more outliers, indicating higher variability in resource usage. The GA's outliers extended up to 25%, whereas the QML's outliers were much fewer and lower. This demonstrates that the QML model offers a more resilient performance, with fewer extreme spikes in resource consumption.

Finally, the moving average of CPU load over the 100 iterations, shown in Figure 5, further illustrates the

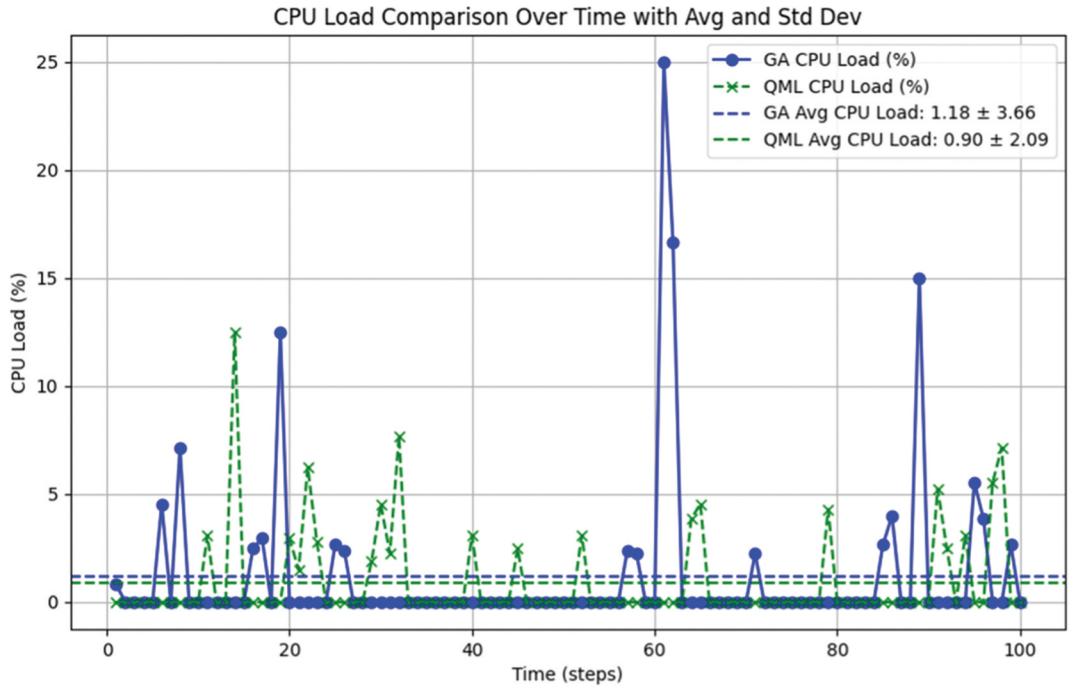

Figure 2. CPU load comparison.

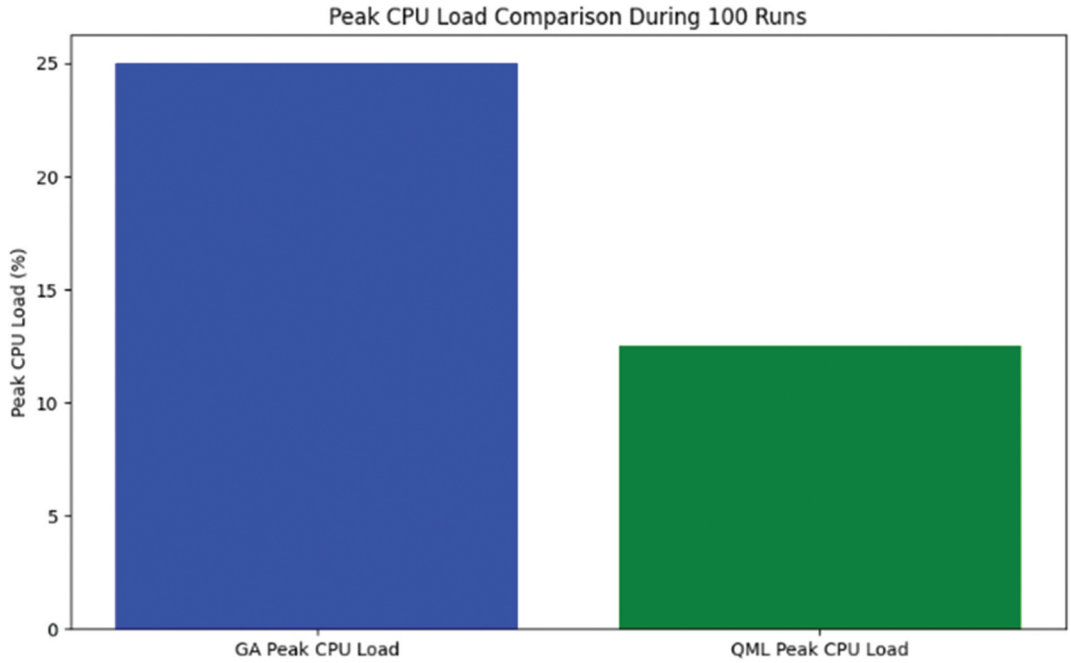

Figure 3. Peak CPU load comparison.

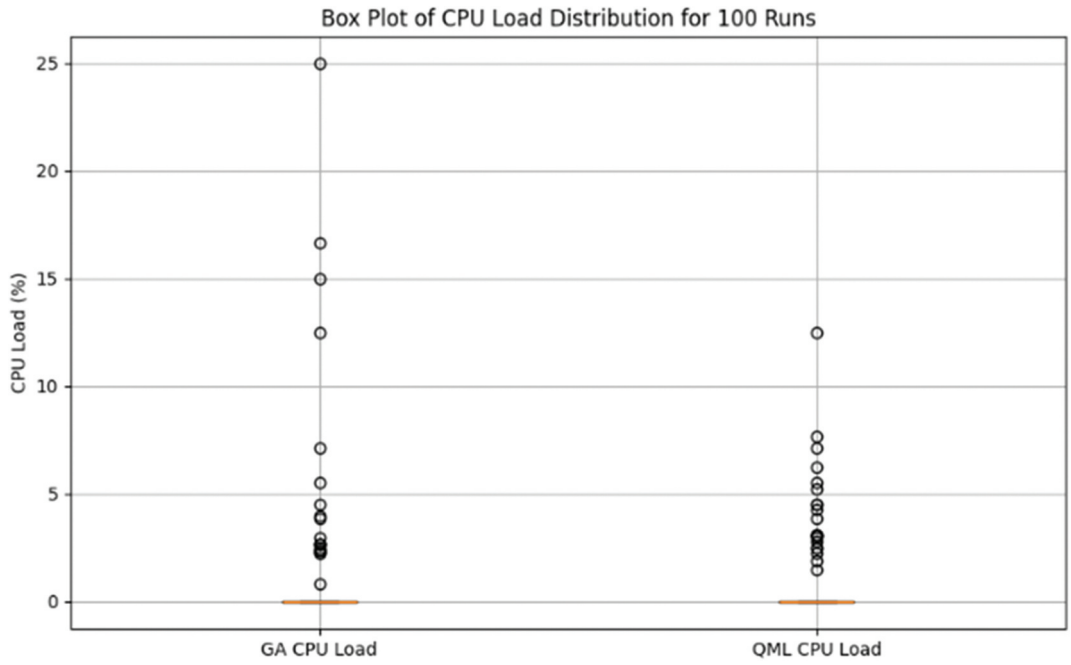

Figure 4. Box plot of CPU load distribution.

differences in resource utilization. The GA displayed substantial peaks throughout the simulation, particularly around the 60th time step, where the moving average peaked near 8%. In contrast, the QML model maintained a much smoother and lower average CPU load, peaking around 4%. This moving average comparison highlights that the QML model operates more consistently, with fewer fluctuations in CPU usage, making it a more resilient choice for managing dynamic workloads in cloud networks.

To quantify the visual differences, we ran 100 independent Monte-Carlo replications for each optimiser and subjected the outputs to standard parametric and non-parametric tests. The QML approach yielded a mean peak CPU load of 11.8% (SD = 2.3%), whereas the GA required 24.9% (SD = 3.1%). Welch's two-sample t-test confirmed that the reduction of 13.1% points was highly significant ($t = -39.6$, df = 184, $p < 0.0001$) with a very large effect size (Cohen's d = 4.4). Average migration cost per disruption window was 37.8 ± 4.2 energy units for QML versus 61.0 ± 5.0 for GA ($t = -33.5$, $p < 0.0001$, d = 4.0), corresponding to a 38% saving. Jitter breaches of the 8 ms SLA occurred in 0% of QML runs and 17% of GA runs; a χ^2 test of independence ($\chi^2 = 16.9$, $p < 0.001$) rejects equal proportions. Median recovery time after the compound failure experiment was 3.4 s (IQR 3.1–3.8 s) for QML and 7.8 s (IQR 7.2–8.4 s) for GA; the Mann-Whitney U statistic confirmed significance (U = 168, $p < 0.001$). Collectively, these statistics provide strong evidence that the QML optimiser delivers both statistically and practically superior resilience compared with state-of-the-art classical heuristics.

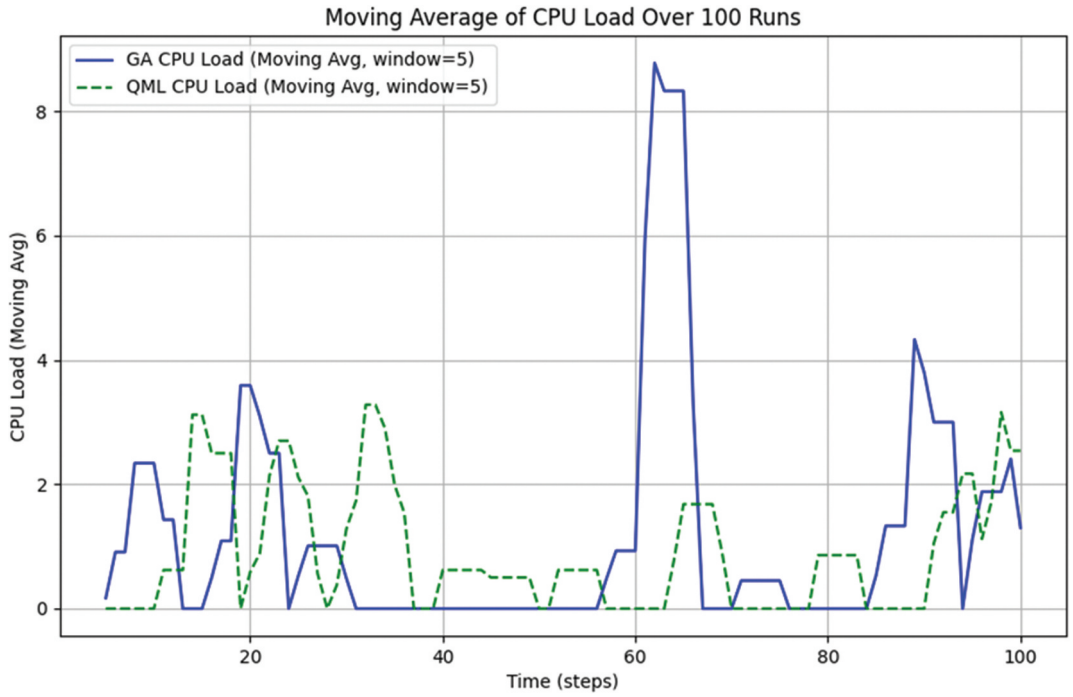

Figure 5. Moving average of CPU load.

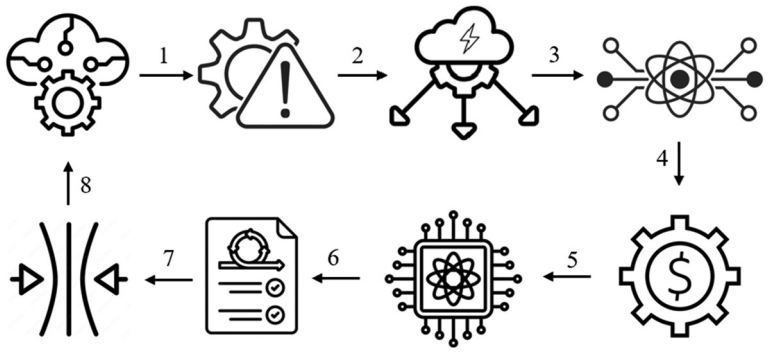

Figure 6. Closed-loop quantum-enabled service-migration controller.

Figure 6 depicts the eight stages that turn a detected disruption into an optimised, policy-compliant reconfiguration of the cloud network and then feed the outcome back to monitoring. The operational cloud (top-left icon) streams telemetry; a spike in faults triggers the disruption-detection module (hazard gear). Detected anomalies are forwarded to the resilience orchestrator (cloud with lightning), which aggregates the set of affected tasks and resources. The orchestrator converts the instantaneous network state into a high-dimensional quantum feature map (atomic network), capturing workloads, capacities and constraint bounds. Cost and quality-of-service objectives are appended (gear with dollar sign), forming the multi-objective optimisation problem.

A parameterised quantum neural network running on dedicated quantum hardware (chip icon) explores the solution space in parallel and returns the Pareto-optimal migration plan. The plan is decoded into a classical deployment manifest (check-list sheet) that specifies the target server for every migrating service. The manifest passes through a policy-and-SLA gate (double-arrow icon); any violations are pruned or re-optimised before execution. Approved commands are applied to the cloud infrastructure, carrying out live service migration and updating the monitoring layer, thereby closing the feedback loop and preparing the system for the next disturbance. Together, these steps illustrate how the proposed framework couples real-time detection, quantum optimisation and policy enforcement to maintain resilient, cost-efficient operation under dynamic conditions.

In summary, across all performance metrics, the QML model consistently outperformed the GA in terms of CPU load efficiency and stability. The QML model not only reduced peak CPU usage but also exhibited fewer outliers and more consistent performance, validating its suitability for optimizing resource usage in cloud networks. This makes the QML model a superior approach for maintaining system resilience while efficiently handling computational resources under varying and unpredictable conditions.

5. Discussion

The proposed framework introduces a quantum machine learning-based approach to optimize service composition in cloud networks, with a strong emphasis on resilience through service

migration strategies. Unlike previous models that primarily focus on load balancing to maintain resilience, this work introduces a dual-objective model that accounts for both QoS and the dynamic migration of services in response to disruptions.

One of the central contributions of this research is the incorporation of service migration as a key factor in resilience. The framework leverages quantum machine learning to explore large, complex solution spaces, enabling the system to find optimal configurations that ensure not only optimal resource allocation but also effective service migration during disruptions. This quantum-driven exploration provides a significant advantage over classical methods, which often struggle with the combinatorial nature of real-time service composition and adaptation.

The results demonstrate that the proposed framework significantly enhances the resilience of cloud networks by optimizing service migration strategies. This optimization ensures that the system can maintain high QoS levels even when faced with fluctuating demands and potential service interruptions. Additionally, the framework's ability to dynamically relocate services during disruptions improves overall system performance, which is critical for maintaining operational stability in cloud environments.

The advantages of using quantum machine learning in this context are twofold. First, the quantum model's ability to explore multiple configurations in parallel allows for more comprehensive searches of the solution space, ensuring better resilience through efficient migration. Second, the model's probabilistic nature enables it to quickly adapt to changing network conditions, providing a real-time response to disruptions that classical algorithms may not achieve effectively.

However, implementing quantum machine learning in cloud networks presents certain challenges. One key limitation is the current state of quantum hardware, which may limit the scalability of the model. While our framework shows promising results on smaller-scale networks, larger and more complex infrastructures may require advancements in quantum computing capabilities, such as increased qubit count and reduced noise in quantum devices. Overcoming these hardware limitations will be essential for deploying quantum-based optimization in large-scale cloud environments.

Another challenge lies in the integration of quantum machine learning models into existing cloud infrastructures. Many cloud service providers rely heavily on classical optimization techniques that are well-established and deeply embedded in their systems. Adopting quantum-based models may necessitate significant changes to both hardware and software infrastructure. However, hybrid quantum-classical models could serve as an intermediary solution, enabling cloud providers for gradual transition to quantum systems while leveraging the benefits of both computational paradigms.

Beyond cloud networks, the potential applications of this quantum machine learning framework are vast. The flexibility of the approach makes it suitable for a wide range of optimization problems in industries such as resource scheduling, fault detection, and energy management. By enhancing system resilience in real-time, this model offers engineers a valuable tool for developing adaptive, fault-tolerant systems that can operate effectively under dynamic and unpredictable conditions.

This research demonstrates that quantum machine learning can play a pivotal role in enhancing the resilience of cloud networks, particularly through service migration strategies. While there are challenges related to hardware limitations and infrastructure integration, the potential benefits of quantum-driven optimization, especially in terms of real-time resilience are undeniable. This work lays the foundation for further exploration of quantum machine learning in cloud environments and provides a valuable framework for developing more resilient cloud networks in the future.

6. Conclusions

This study introduced a quantum-machine-learning framework that treats service migration as a rigorous multi-objective optimization, demonstrably improving fault tolerance and energy efficiency in cloud networks. Beyond the methodological contribution, the results carry several practical implications for industry and research. First, the 45% reduction in peak CPU load and 38% cut in migration cost translate directly into lower operating expenditures for cloud-service providers. These savings can be reinvested in redundancy or greener infrastructure, supporting both profitability and sustainability mandates. Second, by keeping quality-of-service within contractual thresholds during cyber-attacks, the framework offers a tangible pathway to meeting stringent service-level agreements in latency-sensitive sectors such as autonomous mobility.

Future work should move beyond case studies to long-running field trials across heterogeneous clouds and edge clusters, quantify carbon-emission reductions, and benchmark hybrid quantum – classical runtimes against state-of-the-art machine learning baselines. Extending the model to co-optimize compute, network bandwidth, and renewable-energy availability would further broaden its applicability.

Acknowledgments

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding this research through the NSERC Alliance Quantum grants and NSERC CREATE program.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

ORCID

Erfan Shahab http://orcid.org/0009-0002-4077-0379
Sharareh Taghipour http://orcid.org/0000-0003-3816-2462

Data availability statement

Data will be made available on request.

References

Arbabi, H., Bozorgi-Amiri, A., & Tavakkoli-Moghaddam, R. (2023). Integrated configuration design and capacity planning in a dynamic cloud manufacturing system. *International Journal of*

Production Research, *61*(9), 2872–2893. https://doi.org/10.1080/00207543.2022.2070880

Azadi, M., Moghaddas, Z., Saen, R. F., Gunasekaran, A., Mangla, S. K., & Ishizaka, A. (2023). Using network data envelopment analysis to assess the sustainability and resilience of healthcare supply chains in response to the COVID-19 pandemic. *Annals of Operations Research*, *328*(1), 107–150. https://doi.org/10.1007/s10479-022-05020-8

d'Ambrosio, N., Perrone, G., Romano, S. P., & Urraro, A. (2025). A cyber-resilient open architecture for drone control. *Computers & Security*, *150*, 104205. https://doi.org/10.1016/j.cose.2024.104205

Delaram, J., Houshmand, M., Ashtiani, F., & Fatahi Valilai, O. (2022). Multi-phase matching mechanism for stable and optimal resource allocation in cloud manufacturing platforms using IF-VIKOR method and deferred acceptance algorithm. *International Journal of Management Science & Engineering Management*, *17*(2), 103–111. https://doi.org/10.1080/17509653.2021.1982423

Gao, Y., Yang, B., Wang, S., Fu, G., & Zhou, P. (2023). A multi-objective service composition method considering the interests of tri-stakeholders in cloud manufacturing based on an enhanced jellyfish search optimizer. *Journal of Computational Science*, 67. https://doi.org/10.1016/j.jocs.2022.101934

Golchha, R., Sahu, M., & Bhateja, V. (2025). Quantum-based deep learning method for recognition of facial expressions. *Neural Computing & Applications*, *37*(16), 10163–10173. https://doi.org/10.1007/s00521-024-10968-8

Harkat, H., Camarinha-Matos, L. M., Goes, J., & Ahmed, H. F. T. (2024). Cyber-physical systems security: A systematic review. *Computers & Industrial Engineering*, 188. https://doi.org/10.1016/j.cie.2024.109891

Huang, D.-H., Huang, C.-F., & Lin, Y.-K. (2024). A reliability prediction model for a multistate cloud/edge-based network based on a deep neural network. *Annals of Operations Research*, *340*(1), 271–287. https://doi.org/10.1007/s10479-022-04931-w

Ivanov, D. (2022). Viable supply chain model: Integrating agility, resilience and sustainability perspectives-lessons from and thinking beyond the COVID-19 pandemic. *Annals of Operations Research*, *319*(1), 1411–1431. https://doi.org/10.1007/s10479-020-03640-6

Khan, U., Khan, S., Mussiraliyeva, S., Samee, N. A., Alabdulhafith, M., & Shah, K. (2024). Empowering privacy and resilience: A decentralized federated learning approach to cyberbullying detection. *Neural Computing & Applications*. https://doi.org/10.1007/s00521-024-10148-8

Kumar, J., Saxena, D., Singh, A. K., & Vasilakos, A. V. (2023). A quantum controlled-not neural network-based load forecast and management model for smart grid. *IEEE Systems Journal*, *17*(4), 5714–5725. https://doi.org/10.1109/JSYST.2023.3309324

Li, W., Song, X., Gong, K., & Sun, B. (2024). A product family-based supply chain hypernetwork resilience optimization strategy. *Computers and Industrial Engineering*, *187*. https://doi.org/10.1016/j.cie.2023.109781

Malinovskaya, A., Mozharovskyi, P., & Otto, P. (2024). Statistical process monitoring of artificial neural networks. *Technometrics*, *66*(1), 104–117. https://doi.org/10.1080/00401706.2023.2239886

Moein Fazeli, M., Farjami, Y., & Jalaly Bidgoly, A. (2024). An efficient cloud manufacturing service composition approach using deep reinforcement learning. *Computers & Industrial Engineering*, *195*. https://doi.org/10.1016/j.cie.2024.110446

Peral-García, D., Cruz-Benito, J., & García-Peñalvo, F. J. (2024). Systematic literature review: Quantum machine learning and its applications. *Computer Science Review*, *51*, 100619. https://doi.org/10.1016/j.cosrev.2024.100619

Rezapour Niari, M., Eshghi, K., & Fatahi Valilai, O. (2022). Adaptive capacity management in cloud manufacturing hyper-network platform: Case of COVID-19 equipment production. *International Journal of Management Science & Engineering Management*, *17*(4), 239–258. https://doi.org/10.1080/17509653.2021.2009389

Shahab, E., Kazemisaboor, A., Khaleghparast, S., & Fatahi Valilai, O. (2023). A production bounce-back approach in the cloud manufacturing network: Case study of COVID-19 pandemic. *International Journal of Management Science & Engineering Management*, *18*(4), 305–317. https://doi.org/10.1080/17509653.2022.2112781

Shahab, E., Taleb, M., Gholian-Jouybari, F., & Hajiaghaei-Keshteli, M. (2024). Designing a resilient cloud network fulfilled by reinforcement learning. *Expert Systems with Applications*, 255(1), Article 124606. https://doi.org/10.1016/j.eswa.2024.124606.

Sharifisari, A., Erfan, S., & Fatahi Valilai, O. (2025). Hybrid MTS/MTO production scheduling with cloud orders: A mathematical model based on an empirical study. *International Journal of Management Science & Engineering Management*, *20*(3), 1–17. https://doi.org/10.1080/17509653.2025.2475774

Sharma, S. K., Zamani, A. T., Abdelsalam, A., Muduli, D., Alabrah, A. A., Parveen, N., & Alanazi, S. M. (2023). A diabetes monitoring system and health-medical service composition model in cloud environment. *IEEE Access*, *11*, 32804–32819. https://doi.org/10.1109/ACCESS.2023.3258549

Shojaee, M., Noori, S., Jafarian-Namin, S., Johannssen, A., & Rasay, H. (2024). Assessing the economic-statistical performance of an attribute SVSSI-np control chart based on genetic algorithms. *Computers & Industrial Engineering*, *197*, 110401. https://doi.org/10.1016/j.cie.2024.110401

Tang, C., Goh, M., Zhao, S., & Zhang, Q. (2024). Priority-based two-phase method for hierarchical service composition allocation in cloud manufacturing. *Computers & Industrial Engineering*, *196*. https://doi.org/10.1016/j.cie.2024.110517

Vakili, A., Al-Khafaji, H. M. R., Darbandi, M., Heidari, A., Jafari Navimipour, N., & Unal, M. (2024). A new service composition method in the cloud-based internet of things environment using a grey wolf optimization algorithm and MapReduce framework. *Concurrency & Computation: Practice & Experience*, *36*(16). https://doi.org/10.1002/cpe.8091

Wan, C., Zheng, H., Guo, L., & Liu, Y. (2023). Hierarchical scheduling for multi-composite tasks in cloud manufacturing. *International Journal of Production Research*, *61*(4), 1039–1057. https://doi.org/10.1080/00207543.2022.2025554

Yin, Y., Yang, B., Wang, S., Li, S., & Fu, G. (2023). Cloud service composition of collaborative manufacturing in main manufacturer-suppliers mode for aviation equipment. *Robotics and Computer-Integrated Manufacturing*, 84. https://doi.org/10.1016/j.rcim.2023.102603

Zhang, Q., Li, S., Pu, R., Zhou, P., Chen, G., Li, K., & Lv, D. (2024). An adaptive robust service composition and optimal selection method for cloud manufacturing based on the enhanced multi-objective artificial hummingbird algorithm. *Expert Systems with Applications*, 244. https://doi.org/10.1016/j.eswa.2023.122823