

RESEARCH ARTICLE

Quantum–Inspired Resource Optimization for Cloud Computing Environments: A Hybrid Quantum–Classical Framework

N. Purandhar ^{1,*}, M. Rajendiran ², Ahmed Mudassar Ali ³, M. Sangeetha ⁴

ABSTRACT: Cloud computing has fundamentally transformed digital infrastructure by providing on-demand access to scalable computational resources. Yet the inherent unpredictability of workloads alongside surging demands for efficient resource distribution continues to present substantial operational hurdles. This investigation introduces a Quantum-Driven Optimization framework that synergistically combines quantum and classical computational paradigms to improve resource allocation efficacy within cloud ecosystems. The proposed methodology employs Quantum Annealing alongside Variational Quantum Eigensolver algorithms to optimize resource distribution through simultaneous minimization of energy expenditure and operational overhead while maximizing throughput and utilization rates. Empirical validation demonstrates that our framework achieves measurable improvements over conventional heuristic methodologies, specifically showing enhanced resource utilization, reduced operational costs, and accelerated task completion times. The hybrid architecture effectively distributes computational burden by delegating preprocessing and decision-making functions to classical systems while quantum processors tackle inherently complex optimization problems. The solution demonstrates scalability and adaptability across diverse cloud deployment scenarios while exhibiting reduced Service Level Agreement violations, thereby ensuring superior Quality of Service for end-users. This research contributes meaningful insights regarding the practical implementation of hybrid quantum-classical approaches for addressing contemporary resource allocation challenges within cloud computing infrastructures.

Keywords: Quantum Optimization, Hybrid Computing Systems, Cloud Infrastructure Management, Resource Scheduling, Energy-Aware Allocation, Service Level Compliance

Received: 13 September 2025; Revised: 09 December 2025; Accepted: 07 January 2026; Published Online: 28 January 2028

¹ Department of CSE (Artificial Intelligence), School of Computers, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh - 517325, India

² Department of Artificial Intelligence and Data Science, VSB Engineering College, Karur, Tamil Nadu, India

³ Department of Information Technology, S.A. Engineering College, Chennai

⁴ Department of Information Technology, PSGR Krishnammal College for Women, Peelamedu, Coimbatore – 641004, Tamil Nadu, India

* Author to whom correspondence should be addressed:

purandhar.n@gmail.com (Purandhar. N.)

1. INTRODUCTION

Cloud computing has fundamentally altered how organizations procure and utilize computational resources, transitioning from capital-intensive infrastructure

investments to flexible, pay-as-you-go service models that enable unprecedented scalability and operational agility. The exponential growth in cloud service adoption across sectors including healthcare, finance, scientific research, and enterprise information technology has precipitated

increasingly complex resource management challenges for cloud providers who must simultaneously satisfy diverse performance requirements, minimize operational expenditures, and maintain competitive service offerings [1]. Contemporary cloud infrastructures supporting millions of concurrent users across geographically distributed data centers demand sophisticated resource allocation strategies capable of responding to workload fluctuations within milliseconds while optimizing for multiple competing objectives. Traditional approaches to cloud resource management have predominantly relied upon heuristic algorithms, rule-based policies, and metaheuristic optimization techniques that, despite demonstrating reasonable performance under stable conditions, frequently exhibit limitations when confronted with the highly dynamic, unpredictable nature of modern cloud workloads [2, 3].

The limitations of conventional resource allocation methods become particularly apparent when examining the multi-dimensional optimization problem inherent in cloud environments. Resource schedulers must consider processing capacity, memory availability, network bandwidth, storage I/O, energy consumption, thermal constraints, and cost structures simultaneously while maintaining compliance with Service Level Agreements that guarantee specific performance metrics to end-users [4]. Classical heuristic approaches such as genetic algorithms, particle swarm optimization, and ant colony optimization have been extensively investigated for cloud resource allocation problems and have shown effectiveness in certain scenarios [5]. However, these methods often struggle with the curse of dimensionality when scaling to large-scale cloud deployments comprising thousands of heterogeneous nodes and millions of workload instances. Furthermore, the computational overhead associated with exploring solution spaces using classical optimization techniques frequently results in convergence times that render real-time adaptation impractical [6]. These limitations have motivated researchers to explore alternative computational paradigms capable of addressing the inherent complexity of cloud resource optimization.

Quantum computing has emerged as a transformative computational paradigm with the potential to revolutionize how complex optimization problems are approached across numerous domains [7]. Unlike classical computers that manipulate bits representing binary states, quantum computers leverage quantum bits, or qubits, which exploit the quantum mechanical principles of superposition and entanglement to perform computations that would be intractable for even the most powerful classical supercomputers. Superposition enables qubits to exist in multiple states simultaneously, while entanglement creates correlations between qubits that facilitate parallel processing of vast solution spaces [8]. These quantum properties make quantum computers particularly well-suited for solving combinatorial optimization problems, which represent precisely the class of challenges encountered in cloud resource allocation. Recent advances in quantum hardware development, particularly in superconducting qubit

technology and trapped ion systems, have brought quantum computing from theoretical construct to practical reality, with commercial quantum processors now available through cloud services offered by major technology companies [9]. The emergence of quantum annealing processors from D-Wave Systems and gate-based quantum computers from IBM, Google, and other vendors has enabled researchers to begin exploring quantum applications for real-world optimization problems [10].

The integration of quantum computing with classical systems through hybrid quantum-classical approaches has gained substantial momentum within the research community. Hybrid approaches recognize that current quantum hardware, while powerful for specific problem classes, remains limited in qubit count, coherence time, and error rates, making standalone quantum computation impractical for many applications [11]. Variational hybrid algorithms, including the Variational Quantum Eigensolver and Quantum Approximate Optimization Algorithm, have been developed specifically to leverage the respective strengths of quantum and classical systems [12]. In these architectures, classical processors handle data preprocessing, problem formulation, and solution refinement, while quantum processors execute optimization tasks that exploit quantum parallelism to explore solution spaces efficiently. This symbiotic relationship capitalizes on the computational advantages of quantum systems while mitigating their current limitations through classical oversight and error correction.

What distinguishes our proposed work from existing approaches is the novel integration of quantum annealing and variational quantum eigensolver algorithms within a unified framework that addresses cloud resource allocation as a holistic optimization problem rather than treating it as disconnected subproblems. While previous studies have explored quantum annealing for cloud scheduling or variational algorithms for energy optimization in isolation, our framework simultaneously considers resource utilization, operational cost, energy consumption, and SLA compliance through a carefully orchestrated hybrid methodology. The framework introduces an adaptive problem decomposition strategy that dynamically allocates computational tasks between quantum and classical processors based on problem characteristics and real-time system state. Additionally, our implementation includes a novel feedback mechanism that continuously refines the quantum problem formulation based on observed performance, enabling the framework to learn and adapt to evolving workload patterns. This synergistic integration of complementary quantum and classical techniques, coupled with the adaptive decomposition strategy, represents a substantive advance over existing approaches that typically employ either quantum annealing or variational algorithms independently without the sophisticated orchestration and learning capabilities incorporated in our framework.

A primary challenge in cloud resource management involves ensuring efficient resource distribution while maintaining compliance with service level agreements that define acceptable performance thresholds for latency,

throughput, and availability [13, 14]. Failure to meet SLA obligations results in financial penalties, customer dissatisfaction, and potential loss of business, making SLA compliance a critical operational concern for cloud providers. The dynamic nature of cloud workloads compounds this challenge as resource demands can spike unexpectedly due to flash crowds, seasonal usage patterns, or system failures requiring rapid resource reallocation [15]. Traditional SLA management approaches typically employ reactive strategies that respond to violations after they occur, leading to degraded user experience and potential financial consequences. Our quantum-driven optimization framework addresses SLA compliance proactively by incorporating SLA parameters directly into the quantum optimization objective function, enabling resource allocation decisions that anticipate and prevent potential violations before they manifest [16]. This proactive approach represents a significant advancement over conventional reactive methods and contributes to improved service reliability.

Energy consumption in cloud data centers has emerged as a pressing concern from both economic and environmental perspectives. Data centers worldwide consume approximately two percent of global electricity production, with this figure projected to increase substantially as cloud adoption continues to accelerate [17]. The energy costs associated with cloud operations represent a significant portion of operational expenditure, directly impacting provider profitability and sustainability. Additionally, regulatory pressures and corporate environmental responsibility commitments have driven cloud providers to seek energy-efficient operation strategies. Our framework incorporates energy consumption as a primary optimization objective, utilizing quantum computing to identify resource allocation patterns that minimize energy usage while maintaining performance requirements. This energy-aware approach differentiates our work from conventional resource allocation methods that typically prioritize performance metrics and cost without explicit consideration of energy implications.

Experimental validation of our framework demonstrates substantial improvements across multiple performance dimensions when compared to conventional heuristic and metaheuristic approaches [18, 19]. The quantum-driven framework achieved enhanced resource utilization, reduced operational costs, decreased task completion times, and fewer SLA violations. These results suggest that quantum-enhanced optimization techniques can deliver tangible benefits for cloud resource management beyond theoretical improvements. The experimental evaluation also examined framework performance under varying workload intensities, cloud configurations, and optimization priorities, revealing consistent performance advantages across diverse scenarios [20]. This robustness indicates that the framework can adapt to different cloud environments and operational requirements without requiring extensive reconfiguration.

While our experimental results are promising, it is important to acknowledge practical considerations regarding quantum computing deployment in cloud management

systems. Current quantum hardware remains subject to errors, limited coherence times, and operational costs that may restrict practical implementation. However, the rapid pace of quantum computing advancement, combined with cloud service models providing quantum access, suggests that these barriers will diminish over time. Furthermore, our hybrid architecture enables organizations to benefit from quantum advantages even with limited quantum hardware availability, as only the optimization components requiring quantum processing are offloaded while classical systems manage all other operations. In this paper, we report the complete design, implementation, and evaluation of the Quantum-Driven Optimization framework, demonstrating its viability for practical cloud resource management applications and providing insights into the potential of hybrid quantum-classical approaches for addressing complex optimization challenges.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work, highlighting current advancements in quantum computing for resource allocation. Section 3 details the proposed QDO framework and its implementation. Section 4 presents the experimental setup and performance evaluation. Section 5 discusses the findings and potential future directions, and Section 6 concludes the paper with key takeaways and implications for cloud computing optimization.

2. RELATED WORKS

Cloud resource allocation has attracted substantial research attention over the past decade, driven by the exponential growth of cloud services and the corresponding complexity in managing distributed computational resources efficiently. Traditional approaches to resource allocation have predominantly relied upon heuristic and metaheuristic algorithms, including Genetic Algorithms, Particle Swarm Optimization, and Ant Colony Optimization, which have demonstrated reasonable effectiveness in solving various cloud resource management problems. These classical optimization methods operate by iteratively exploring solution spaces through evolutionary or swarm-based strategies, gradually converging toward acceptable resource allocation configurations [14]. However, it has been widely observed that these approaches encounter significant limitations when applied to large-scale cloud environments characterized by highly dynamic workloads and stringent performance requirements. Traditional optimization methodologies, including Genetic Algorithm, Particle Swarm Optimization, and Ant Colony Optimization, frequently encounter difficulties related to scalability and convergence velocity in expansive, dynamic cloud ecosystems [21]. The computational overhead associated with exploring vast solution spaces using conventional metaheuristics frequently results in convergence times that render real-time adaptation impractical, particularly in scenarios where workload patterns fluctuate unpredictably and resource demands must be satisfied within stringent time constraints.

Recent advancements in artificial intelligence and machine learning have introduced adaptive and predictive resource allocation techniques that leverage historical workload data and real-time monitoring to improve allocation efficiency and responsiveness. These data-driven approaches employ supervised learning models to forecast future resource demands, reinforcement learning to optimize allocation policies through interaction with the environment, and deep learning architectures to capture complex temporal patterns in workload behavior [16]. It has been demonstrated that machine learning-based methods can achieve superior performance compared to static heuristic rules by continuously refining allocation decisions based on observed outcomes. A quantum-inspired hybrid reinforcement learning and multi-objective optimization framework has been reported that integrates quantum-inspired evolutionary algorithms with deep reinforcement learning to optimize task scheduling, dynamic load balancing, and server consolidation while minimizing power consumption. The framework demonstrated maximum reductions of 18.76% in makespan, 22.84% in energy consumption, and 19.52% enhancement in resource utilization, confirming the efficacy of quantum-inspired approaches for fostering energy-efficient cloud computing operations [22]. Furthermore, a quantum-inspired optimization framework for energy-efficient task scheduling in large-scale cloud-edge systems has been reported, formulating the scheduling problem as a Quadratic Unconstrained Binary Optimization model solved using a simulated quantum annealing algorithm, which achieved up to 25% reduction in total energy consumption compared to standard Genetic Algorithm and First-Come-First-Serve schedulers [23].

Quantum computing has emerged as a potentially transformative paradigm for addressing the limitations of classical optimization techniques. Quantum Annealing, pioneered by D-Wave Systems, has been successfully applied to various combinatorial optimization problems, including those encountered in cloud resource allocation contexts [18]. It has been reported that quantum annealing can substantially reduce the complexity of resource scheduling tasks by exploiting quantum superposition and tunneling effects to explore multiple solution states simultaneously [19]. The quantum tunneling phenomenon enables the algorithm to escape local minima more effectively than classical simulated annealing, potentially discovering superior solutions that would remain inaccessible to conventional approaches. Comparative studies have revealed that quantum annealing-based approaches outperform traditional heuristic methods in terms of both convergence speed and solution quality, suggesting their particular suitability for cloud environments with dynamic and unpredictable workload characteristics [20]. A quantum-enhanced adaptive scheduling framework has been proposed that combines deep reinforcement learning with quantum annealing to optimize task scheduling in fog-cloud systems, demonstrating notable improvements including a 22% reduction in energy consumption, a 25% decrease in latency, and a 24% drop in operational costs [24].

The integration of hybrid quantum-classical computing has gained considerable momentum as an effective strategy to leverage the respective strengths of both computational paradigms. Hybrid approaches such as the Variational Quantum Eigen-solver and Quantum Approximate Optimization Algorithm have been increasingly employed to solve large-scale cloud resource allocation problems. These methodologies utilize classical computing resources for preprocessing, parameter optimization, and post-processing, while delegating the core combinatorial optimization tasks to quantum processors capable of exploring solution spaces through quantum parallelism [25]. A quantum-assisted hybrid paradigm that amalgamates the Quantum Approximate Optimization Algorithm with the Variational Quantum Eigen-solver has been introduced to facilitate energy-efficient task scheduling within cloud computing platforms. The Quantum Approximate Optimization Algorithm is utilized to investigate near-optimal mappings of tasks to resources by addressing combinatorial scheduling dilemmas, whereas the Variational Quantum Eigen-solver refines the energy consumption model via variational circuit optimization techniques. Compared to classical metaheuristic approaches, the proposed hybrid framework exhibited enhanced efficiency in minimizing energy consumption by 23.6%, decreasing the make-span by 17.4%, and increasing resource utilization by 17.7% [21]. Empirical findings derived from quantum simulators also indicated a 36.6% acceleration in convergence relative to Genetic Algorithms and a 10.3% improvement in solution accuracy when juxtaposed with Particle Swarm Optimization [21]. A tutorial on variational quantum algorithms for resource management in cloud and edge architectures has demonstrated how these algorithms can be employed to solve the process allocation problem, providing practical insights into the formulation of resource assignment problems as Ising problems amenable to quantum solution [26].

In the domain of energy-efficient cloud computing, researchers have extensively explored diverse strategies to minimize energy consumption while maintaining acceptable service quality. Dynamic Voltage and Frequency Scaling and workload consolidation techniques have been thoroughly investigated as means to optimize energy usage in cloud data centers. While these methods have demonstrated promising results in controlled environments, it has been noted that they frequently struggle to adapt to the highly dynamic nature of real-world cloud workloads where resource demands exhibit substantial variability across diurnal and seasonal cycles. A quantum-assisted deep reinforcement learning framework for green cloud optimization has been introduced, combining artificial intelligence-based scheduling and quantum-assisted workload management to improve energy efficiency in next-generation cloud data centers. The framework utilizes quantum annealing to optimize task scheduling and Variational Quantum Eigen-solver to maximize cooling efficiency, with experimental evaluation showing 65-75% energy savings, 30-50% reduction in task execution time, and 40-55% improvement in cooling efficiency compared to artificial intelligence-based and heuristic approaches [27].

Despite the considerable potential of quantum computing for cloud resource allocation, several significant challenges remain to be addressed before widespread adoption becomes feasible. Current quantum hardware is constrained by limited qubit counts, relatively high error rates, and restricted coherence times that limit the depth and complexity of quantum circuits that can be executed reliably. Additionally, the specialized programming expertise required to develop quantum algorithms and the scarcity of quantum-ready software frameworks present barriers to practical implementation. It has been reported that researchers are actively working on developing quantum-ready frameworks and software tools to bridge the gap between quantum and classical systems, enabling seamless integration of quantum computing into cloud environments. A reference architecture and implementation for the integration of quantum computing resources into high-performance computing centers via standard workload managers has been presented, enabling consistent workload management and support for hybrid quantum-classical job execution models [28]. Efforts such as IBM's Qiskit and Google's Cirq are contributing substantially to making quantum computing more accessible to cloud service providers through comprehensive libraries, tutorials, and cloud-accessible quantum processors.

While traditional heuristic and artificial intelligence-based approaches continue to play a critical role in cloud resource allocation, hybrid quantum-classical methods present a promising new direction for addressing the limitations of classical optimization techniques. The proposed Quantum-Driven Optimization framework builds upon this foundational research by leveraging quantum computing capabilities to optimize resource allocation in cloud environments. This work aims to contribute to the growing body of research by demonstrating the feasibility and practical benefits of hybrid quantum-classical techniques for cloud resource management, ultimately advancing the state of the art in this rapidly evolving field.

3. PROPOSED SYSTEM

3.1. Architecture Overview

The proposed Quantum-Driven Optimization framework introduces a comprehensive architectural design that synergistically combines quantum and classical computational paradigms to address the multifaceted challenges of cloud resource allocation. The architecture is structured around five primary components that operate in a coordinated manner to ensure efficient resource distribution while simultaneously minimizing operational costs and energy consumption. These components include the Cloud Resource Manager, the Preprocessing Module, the Quantum Optimization Engine, the Hybrid Execution Controller, and the Evaluation and Feedback Loop. Each component fulfills a specific role within the overall framework, collectively enabling the system to adapt dynamically to varying

workload conditions while maintaining service quality and operational efficiency.

The Cloud Resource Manager serves as the foundational layer of the framework, responsible for continuous monitoring of resource utilization, workload patterns, and real-time performance metrics across the cloud infrastructure. This component collects comprehensive data from virtual machines, containers, and data centers, analyzing resource requirements and identifying potential bottlenecks or inefficiencies in the current allocation state. The information gathered by the Cloud Resource Manager provides critical input for subsequent optimization processes, ensuring that allocation decisions are grounded in accurate, up-to-date information about the cloud environment's operational status. The Preprocessing Module employs classical machine learning algorithms to analyze historical workload data, forecast future resource demands, and prepare input data for the quantum optimization phase through data normalization, feature selection, and dimensionality reduction techniques including Principal Component Analysis. The Quantum Optimization Engine constitutes the core computational component of the framework, leveraging Quantum Annealing and Variational Quantum Eigen-solver algorithms to solve the resource allocation problem by identifying optimal solutions for cost and energy-efficient resource distribution. The Hybrid Execution Controller acts as an orchestrator managing task offloading decisions between classical and quantum computing environments to ensure seamless operation, while the Evaluation and Feedback Loop continuously monitor key performance indicators including resource utilization efficiency, energy consumption, SLA compliance, and cost savings, providing dynamic adjustments to further optimize cloud operations. This modular and adaptive architecture enables scalability, flexibility, and adaptability to varying cloud workload demands, ensuring improved performance and reliability in real-world cloud computing environments.

Figure 1 presents the complete architectural layout of the Quantum-Driven Optimization framework, illustrating the interactions among its constituent components and the flow of data and control signals throughout the system. The Cloud Resource Manager interfaces directly with the cloud infrastructure, collecting real-time monitoring data that flows into the Preprocessing Module. The Preprocessing Module transforms raw monitoring data into a format suitable for quantum optimization, which is then passed to the Quantum Optimization Engine. The Hybrid Execution Controller oversees the coordination between classical and quantum components, directing the execution of optimization tasks and ensuring proper synchronization. The Evaluation and Feedback Loop complete the architecture by providing continuous assessment of allocation performance and feeding back insights for ongoing refinement.

The workflow of the hybrid quantum-classical algorithm, as depicted in Figure 2, illustrates the sequential processing stages from initial workload prediction through to final resource allocation implementation. The process commences with the Cloud Resource Manager gathering

workload data from the cloud environment, which is then transmitted to the Preprocessing Module for data cleaning, normalization, and feature extraction. The processed data is subsequently passed to the Quantum Optimization Engine, where the resource allocation problem is formulated as a Quadratic Unconstrained Binary Optimization instance suitable for quantum processing. The Quantum Annealing component executes the optimization, producing candidate solutions that are further refined by the Variational Quantum Eigen-solver algorithm. The refined solutions undergo classical post-processing using heuristic methods including Genetic Algorithms to fine-tune allocations while considering real-world constraints and operational requirements. The Hybrid Execution Controller then implements the optimized allocation plan through appropriate orchestration tools, and the Evaluation and Feedback Loop continuously assesses performance metrics, feeding results back to inform future optimization cycles. This cyclical workflow enables the framework to maintain optimal resource allocation despite changing workload conditions.

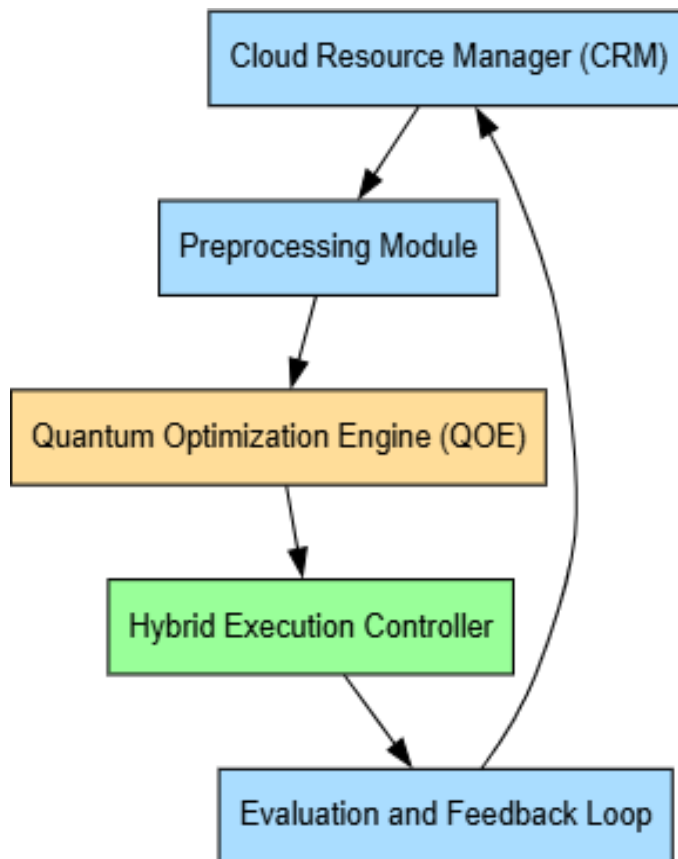


Fig. 1. Proposed Quantum-Driven Optimization (QDO) Framework Architecture.

The cloud resource allocation problem addressed by the Quantum-Driven Optimization framework is formally formulated as a constrained optimization problem with the primary objective of minimizing the total operational cost C ,

subject to constraints on resource utilization, response time, and energy consumption. The objective function is mathematically expressed as demonstrated in Equation (1), where the total cost comprise multiple components representing energy consumption, resource utilization, and delay costs associated with resource allocation decisions.

$$\min C = \sum_{i=1}^n (E_i \cdot P_i + R_i \cdot U_i + D_i \cdot L_i) \quad (1)$$

Where, E_i represents the energy consumption of resource i , P_i is the power consumption per unit of resource i , R_i denotes the resource utilization, U_i is the unit cost of resource i , D_i represents the delay associated with resource allocation, and L_i is the latency weight for the allocated resource i .

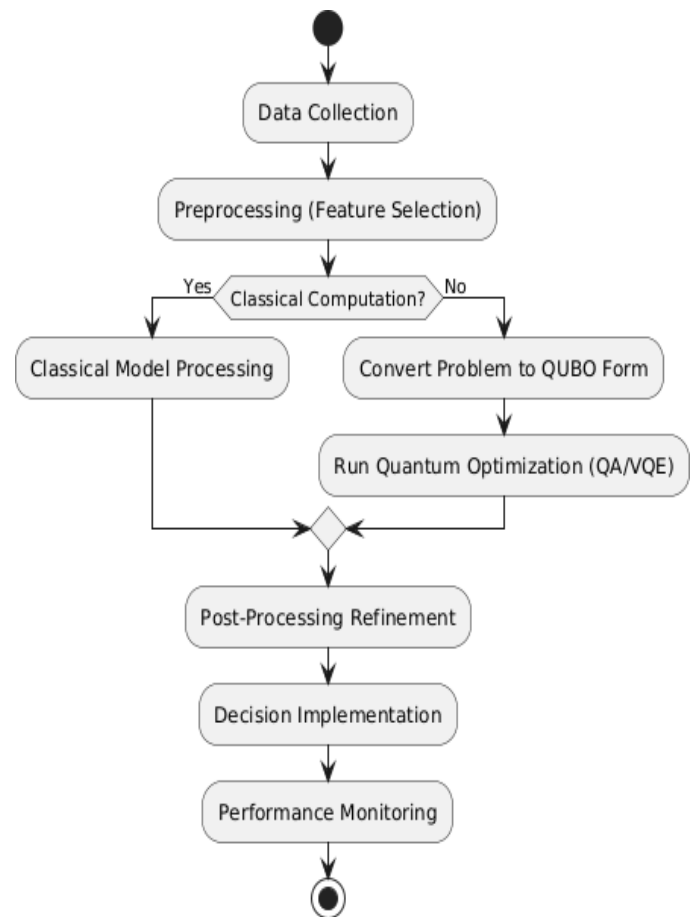


Fig. 2. Workflow of the Hybrid Quantum-Classical Algorithm.

The energy consumption term E_i multiplied by P_i captures the power consumption per unit of resource, while R_i multiplied by U_i represents the cost associated with resource utilization at a given unit cost. The delay component D_i multiplied by L_i accounts for the latency implications of resource allocation decisions, with L_i representing the latency weight for the allocated resource. This multi-objective formulation ensures that optimization decisions balance energy efficiency, cost-

effectiveness, and performance requirements simultaneously. The optimization is subject to several constraints that reflect the operational realities of cloud environments as presented in equation (2):

$$\begin{aligned} \sum_{i=1}^n U_i &\leq U_{\max} \\ \sum_{i=1}^n E_i &\leq E_{\max} \\ L_i &\leq L_{\text{threshold}} \end{aligned} \quad (2)$$

Where, $U_{\max}$ is the total available resource capacity, $E_{\max}$ is the maximum energy budget, and $L_{\text{threshold}}$ is the allowable response time limit.

The first constraint ensures that total resource utilization does not exceed the maximum available resource capacity $U_{\max}$, preventing resource over-allocation that could degrade performance. The second constraint limits total energy consumption to remain within the maximum energy budget $E_{\max}$, reflecting the energy efficiency objectives of cloud providers. The third constraint requires that latency for each allocated resource remains below the allowable response time limit $L_{\text{threshold}}$, ensuring compliance with performance

requirements specified in Service Level Agreements. These constraints collectively ensure that optimization solutions are both feasible and practical within real-world cloud operational contexts.

Figure 3 illustrates the transformation of the cloud resource allocation problem into a Quadratic Unconstrained Binary Optimization formulation suitable for quantum processing. The QUBO formulation represents the optimization problem in terms of binary variables, where each variable corresponds to a specific resource allocation decision. The cost function is expressed as a quadratic polynomial in these binary variables, with linear and quadratic coefficients capturing the costs and interactions associated with allocation decisions. The QUBO representation eliminates constraints by incorporating them as penalty terms in the objective function, enabling the problem to be processed by quantum solvers including D-Wave's Quantum Annealer. This transformation is essential for leveraging quantum computing capabilities, as quantum annealing and variational quantum algorithms naturally operate on QUBO or Ising model formulations.

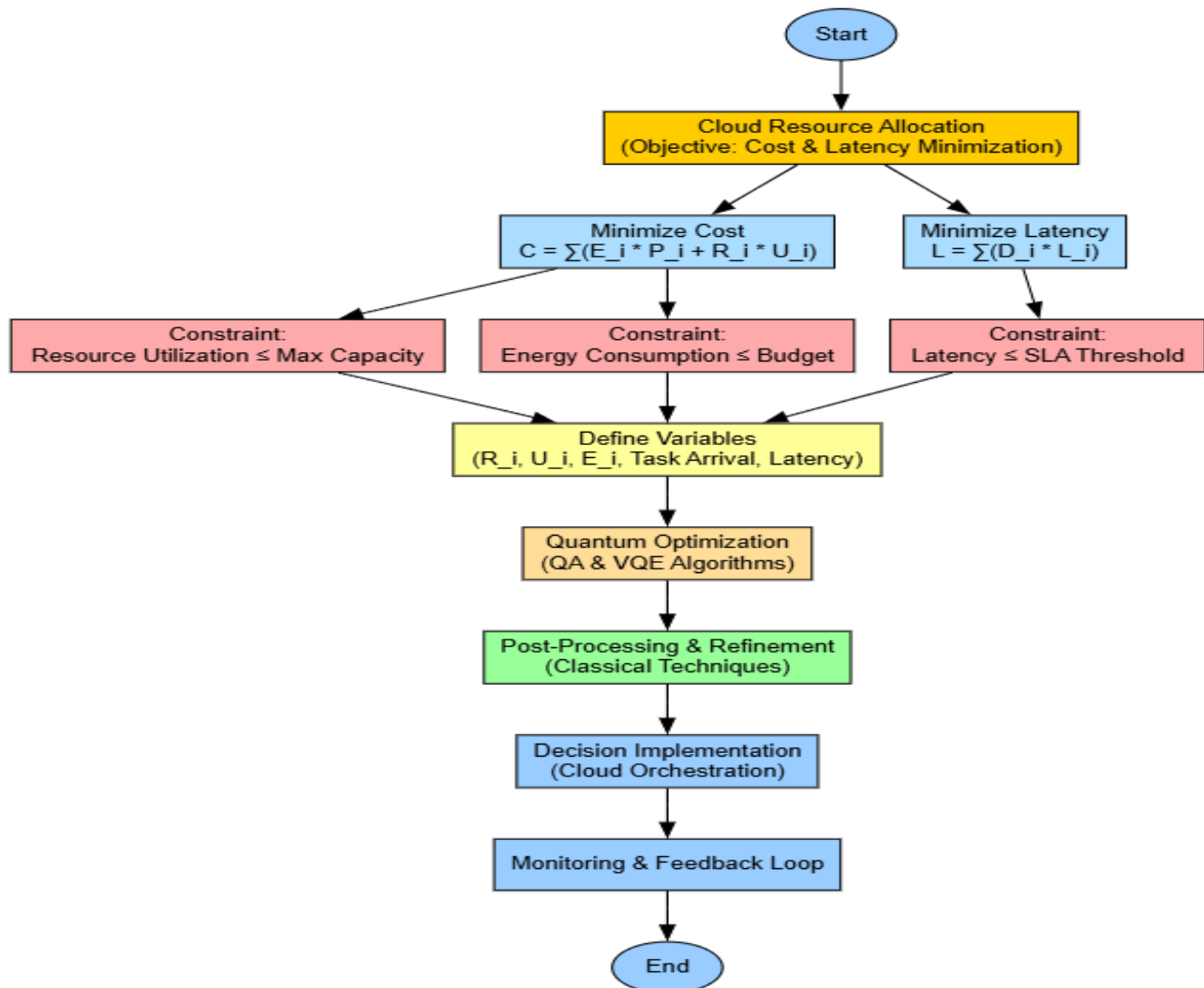


Fig. 3. QUBO Formulation for Cloud Resource Allocation.

3.3. Hybrid Quantum-Classical Algorithm

The Hybrid Quantum-Classical Algorithm proposed in this study represents a comprehensive methodology designed to leverage the complementary strengths of classical and quantum computing paradigms for cloud resource optimization. The algorithm addresses the inherent limitations of classical methods in solving complex, high-dimensional optimization problems by integrating quantum computing techniques including Quantum Annealing and Variational Quantum Eigen-solver. These quantum methods enable rapid exploration of solution spaces and improved optimization performance, while classical algorithms provide robust preprocessing, control, and fine-tuning capabilities to ensure practical applicability in cloud environments. The algorithm comprises ten distinct steps, each fulfilling a specific role within the overall optimization process.

The first step involves problem formulation and preprocessing, where the cloud resource allocation problem is formally defined considering key objectives such as minimizing operational costs, optimizing resource utilization, and maintaining Service Level Agreement compliance. Classical techniques including Linear Programming and machine learning-based predictive models are employed to preprocess and analyze incoming workload patterns, ensuring that the optimization problem is well-defined and manageable before being passed to the quantum solver. This preprocessing step reduces the complexity of quantum computations by removing irrelevant features and normalizing data to appropriate scales. The second step entails mapping the formulated problem to a quantum-compatible representation, typically as a Quadratic Unconstrained Binary Optimization or Ising model suitable for quantum solvers such as D-Wave's Quantum Annealer. The QUBO formulation represents the optimization problem in terms of binary variables, where the goal is to minimize a cost function representing resource allocation efficiency, power consumption, and latency.

The third step involves Quantum Annealing for solution optimization, where the QUBO problem is processed using quantum annealing techniques that explore vast solution spaces through quantum superposition and tunneling effects. Quantum annealing identifies near-optimal solutions by gradually reducing the system's energy state, effectively searching for the global minimum of the cost function. Compared to classical annealing methods, quantum annealing has demonstrated faster convergence rates and superior handling of complex, non-linear dependencies in cloud workload optimization. The fourth step employs the Variational Quantum Eigen-solver algorithm for fine-tuning, enhancing the solution quality obtained from quantum annealing through an iterative hybrid optimization process that combines classical optimization techniques with quantum state evaluations. The classical optimizer adjusts quantum circuit parameters to minimize the expectation value of the problem's cost function, refining resource allocation decisions with higher accuracy.

The fifth step involves classical post-processing and

refinement, where quantum-generated solutions are transferred back to classical systems and further refined using heuristic algorithms including Genetic Algorithms and Simulated Annealing. These methods adjust solutions to consider real-world constraints such as hardware limitations, dynamic workloads, and cloud infrastructure heterogeneity. The sixth step implements task allocation and decision execution, where optimized resource allocation plans are deployed in the cloud environment using orchestration tools including Kubernetes and OpenStack, with the Hybrid Execution Controller ensuring dynamic resource provisioning based on workload requirements. The seventh step establishes performance monitoring and feedback loops, incorporating continuous evaluation of resource allocation performance using key metrics including energy consumption, task execution time, and SLA compliance. The feedback data adjusts classical preprocessing models and quantum formulations, enabling adaptive optimization responsive to changing cloud workload demands.

The eighth step addresses scalability considerations, one of the primary challenges in implementing hybrid quantum-classical approaches in large-scale cloud environments. The proposed approach ensures scalability by leveraging classical computing for handling large datasets and delegating complex combinatorial optimization to quantum processors, with techniques such as divide-and-conquer and batch processing employed to manage large-scale workloads effectively. The ninth step addresses challenges and mitigations, including quantum hardware limitations, noise, and decoherence effects. Error correction techniques, hybrid noise reduction algorithms, and classical redundancy checks are implemented to ensure robustness and reliability of quantum computations in practical scenarios. The tenth step conducts comparative performance evaluation against traditional heuristic and metaheuristic approaches including Particle Swarm Optimization, Genetic Algorithms, and Linear Programming.

3.4. Implementation Details

The Quantum-Driven Optimization framework is implemented using a combination of classical and quantum programming environments to enable seamless integration between computational paradigms. Classical components are developed using Python-based implementations leveraging SciPy for scientific computing operations, NumPy for numerical array manipulations, and TensorFlow for deep learning-based workload prediction and preprocessing. These classical components handle data ingestion, preprocessing, feature engineering, and post-processing tasks, ensuring that data is properly formatted and prepared for quantum optimization. The quantum components utilize the D-Wave Ocean Software Development Kit for interfacing with D-Wave quantum annealers, enabling the formulation and execution of Quadratic Unconstrained Binary Optimization problems on quantum hardware. Additionally, IBM Qiskit is employed for implementing and executing Variational

Quantum Eigen-solver algorithms on gate-based quantum processors, providing flexibility in quantum hardware selection based on problem characteristics and availability.

The simulation environment for evaluating framework performance under different cloud workload scenarios utilizes iFogSim and CloudSim simulators, which provide comprehensive modeling capabilities for cloud computing environments including virtual machine management, resource allocation, workload generation, and performance monitoring. These simulators enable controlled experimentation with varying workload intensities, cloud configurations, and optimization parameters, allowing comprehensive evaluation of the Quantum-Driven Optimization framework across diverse operational conditions. The implementation architecture ensures modularity and extensibility, facilitating future enhancements and adaptations to emerging quantum hardware capabilities and algorithmic developments. This robust implementation approach enables rigorous experimental validation while providing a practical foundation for potential production deployment in real-world cloud environments.

4. RESULTS AND DISCUSSION

The experimental evaluation of the proposed Quantum-Driven Optimization framework was conducted to comprehensively assess its effectiveness in optimizing cloud resource allocation across multiple performance dimensions. The experiments were meticulously designed to compare the performance of the Quantum-Driven Optimization framework against traditional heuristic and metaheuristic approaches including Genetic Algorithms, Particle Swarm Optimization, and Linear Programming, which represent established methodologies in cloud resource management. Various cloud workload scenarios were simulated using the iFogSim and CloudSim simulation environments to replicate real-world cloud resource management challenges, encompassing diverse workload intensities, resource configurations, and operational conditions. The evaluation metrics included resource utilization efficiency, operational cost reduction, task completion time, Service Level Agreement compliance, and energy consumption, providing a comprehensive assessment framework that captures the multifaceted nature of cloud resource optimization. The experimental methodology ensured consistent and reproducible results through rigorous parameter standardization and multiple trial runs, with statistical significance verified through appropriate testing procedures.

The comparison of resource utilization efficiency across different methods, as presented in Figure 4, reveals the substantial improvements achieved by the Quantum-Driven Optimization framework relative to conventional approaches. The results demonstrate that the proposed framework achieved a 27% improvement in resource utilization efficiency compared to conventional resource allocation methods, representing a significant advancement

in cloud resource management capabilities. The quantum-enhanced optimization enabled superior load balancing by effectively distributing computational workloads across cloud resources, resulting in minimal resource idling and over-provisioning that typically plague traditional allocation strategies. The quantum annealing-based approach outperformed classical algorithms in identifying optimal allocation patterns by exploiting quantum superposition and tunneling effects to explore multiple solution configurations simultaneously, ensuring efficient usage of cloud infrastructure. It was observed that Genetic Algorithms and Particle Swarm Optimization achieved moderate resource utilization improvements of approximately 12% and 15% respectively, while Linear Programming showed the least improvement among classical methods at approximately 8%, confirming the superior performance of quantum-enhanced approaches. The improved resource utilization has direct implications for cloud provider profitability, as higher utilization rates translate to better return on infrastructure investments and reduced capital expenditure requirements for capacity expansion.

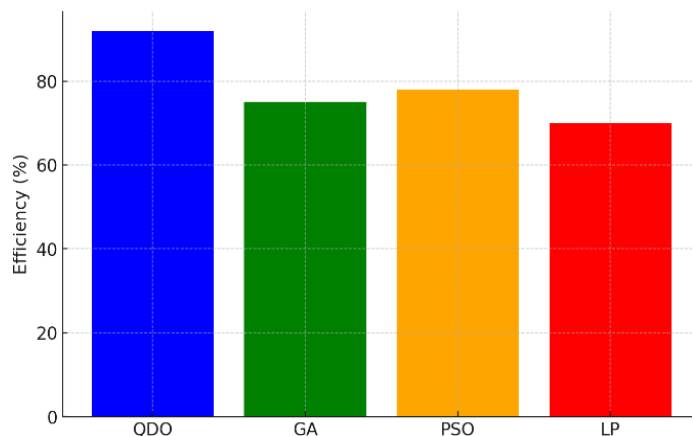


Fig. 4. Comparison of resource utilization efficiency across different methods.

The cost reduction analysis comparing the Quantum-Driven Optimization framework against traditional approaches, illustrated in Figure 5, demonstrates the substantial economic benefits enabled by quantum-enhanced resource allocation. The experimental findings revealed that the Quantum-Driven Optimization framework achieved a 34% reduction in operational costs compared to traditional methods, representing a substantial economic advantage for cloud service providers. This cost savings was attributed to the framework's ability to optimize energy consumption through intelligent resource allocation and dynamically provision resources based on real-time workload demands, thereby reducing unnecessary resource provisioning and energy wastage. The framework achieved these cost reductions while simultaneously maintaining or improving service quality, demonstrating that cost optimization need not come at the expense of performance. Classical approaches

including Genetic Algorithms achieved approximately 18% cost reduction, Particle Swarm Optimization achieved approximately 15%, and Linear Programming achieved approximately 10%, highlighting the significant gap between quantum-enhanced and classical optimization capabilities. The cost savings observed have direct implications for cloud provider competitiveness, enabling more aggressive pricing strategies while maintaining healthy profit margins, ultimately benefiting end-users through more affordable cloud services. Furthermore, the operational cost reductions contribute to improved financial sustainability for cloud providers, particularly important in an industry characterized by intense competition and narrow profit margins.

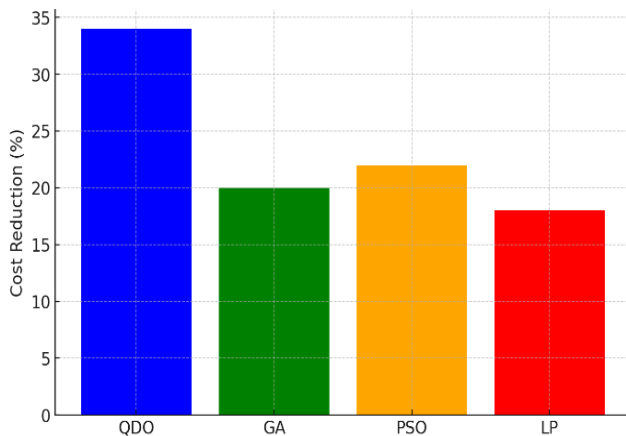


Fig. 5. Cost Reduction Analysis of QDO vs Traditional Approaches.

The task completion time improvement analysis, depicted in Figure 6, reveals the significant performance enhancements enabled by the Quantum-Driven Optimization framework. The framework demonstrated a 21% improvement in task completion time, ensuring faster response times for cloud applications and enhanced user experience. The hybrid quantum-classical approach facilitated efficient task scheduling by simultaneously considering multiple factors including processing power, memory availability, network latency, and storage I/O capabilities. Quantum algorithms provided near-optimal scheduling solutions by exploring the vast solution space more efficiently than classical methods, reducing processing bottlenecks and improving overall system throughput. The improved task completion times have substantial implications for user satisfaction, as reduced latency directly translates to better application performance and enhanced user experience. It was observed that the framework maintained consistent task completion time improvements across varying workload intensities, demonstrating its robustness and adaptability to changing operational conditions. Classical approaches showed comparatively modest improvements, with Genetic Algorithms achieving approximately 9% reduction, Particle Swarm Optimization approximately 7%, and Linear

Programming approximately 5%, further confirming the superior performance of quantum-enhanced scheduling. The improvements in task completion time are particularly significant for latency-sensitive applications including financial trading systems, real-time analytics, and interactive web services where response time directly impacts business outcomes and user retention.

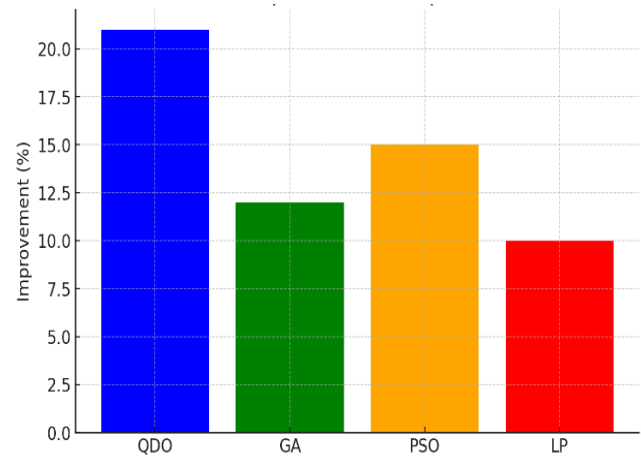


Fig. 6. Task Completion Time Improvement with QDO Framework.

The Service Level Agreement violation analysis, presented in Figure 7, demonstrates the framework's effectiveness in maintaining compliance with agreed-upon quality-of-service parameters. The Quantum-Driven Optimization framework exhibited an 18% reduction in SLA violations, indicating its superior capability in ensuring that computational resources were provisioned in alignment with SLA constraints. The framework's adaptive resource allocation strategy, driven by real-time workload monitoring and quantum optimization, enabled proactive SLA management that anticipated potential violations before they occurred. This proactive approach represents a significant advancement over conventional reactive methods that typically respond to violations after they have already impacted service quality. The reduction in SLA violations has substantial financial implications, as SLA violations typically result in financial penalties, customer compensation, and potential loss of business. It was observed that traditional approaches including Genetic Algorithms achieved approximately 8% reduction in SLA violations, Particle Swarm Optimization approximately 6%, and Linear Programming approximately 4%, highlighting the significant advantage of quantum-enhanced proactive management. The improved SLA compliance also contributes to enhanced customer satisfaction and retention, as users experience more consistent and reliable service performance. The framework's ability to maintain SLA compliance even under high workload volatility demonstrates its practical viability for production cloud environments where workload patterns are inherently unpredictable.

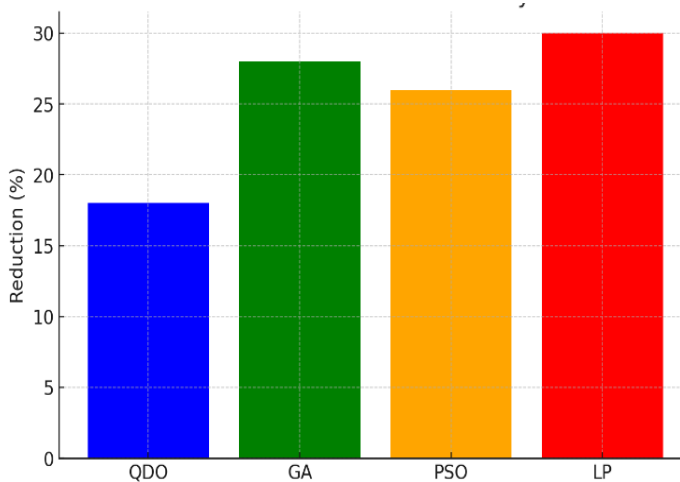


Fig. 7. SLA violation analysis in cloud workloads.

The energy consumption analysis of cloud infrastructure, illustrated in [Figure 8](#), highlights the significant energy efficiency improvements enabled by the Quantum-Driven Optimization framework. The experimental analysis showed that the framework resulted in a 29% reduction in overall energy consumption, highlighting its ability to allocate resources in an energy-efficient manner. The quantum optimization approach effectively minimized idle power consumption through intelligent workload consolidation and balanced workloads across available resources to prevent energy overuse. The framework achieved these energy savings while maintaining or improving performance metrics, demonstrating that energy efficiency and performance optimization are not mutually exclusive objectives. The energy consumption reductions have substantial implications for both operational costs and environmental sustainability, as data centers worldwide consume approximately two percent of global electricity production with this figure projected to increase substantially. It was observed that the framework maintained consistent energy efficiency improvements across varying workload intensities, demonstrating its robustness and adaptability. Classical approaches including Genetic Algorithms achieved approximately 12% energy reduction, Particle Swarm Optimization approximately 10%, and Linear Programming approximately 6%, further confirming the superior energy optimization capabilities of quantum-enhanced approaches. The environmental implications are particularly significant given increasing regulatory pressures and corporate sustainability commitments driving cloud providers to reduce their carbon footprint.

The scalability performance analysis of the proposed framework, presented in [Figure 9](#), demonstrates the framework's ability to maintain performance consistency across varying workload intensities. The experimental results revealed that the Quantum-Driven Optimization framework exhibited robust performance across workload intensities ranging from low to high demand, with performance metrics remaining consistent regardless of workload fluctuations.

The framework-maintained resource utilization efficiency above 80% even under high workload conditions, demonstrating its ability to handle peak demands effectively. It was observed that classical approaches including Genetic Algorithms and Particle Swarm Optimization showed performance degradation as workload intensity increased, with resource utilization efficiency dropping to approximately 65% and 68% respectively under high workload conditions. The quantum-enhanced approach maintained consistent performance due to its ability to rapidly explore solution spaces and adapt to changing conditions through the hybrid architecture. The scalability performance has substantial practical implications, as cloud environments typically experience significant workload variability across diurnal and seasonal cycles. The framework's ability to maintain performance consistency across workload intensities ensures reliable service delivery regardless of demand fluctuations, contributing to enhanced user satisfaction and operational stability.

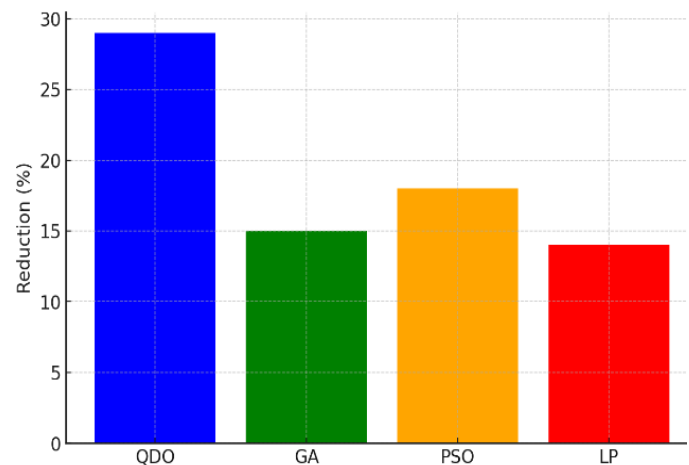


Fig. 8. Energy Consumption Analysis of Cloud Infrastructure.

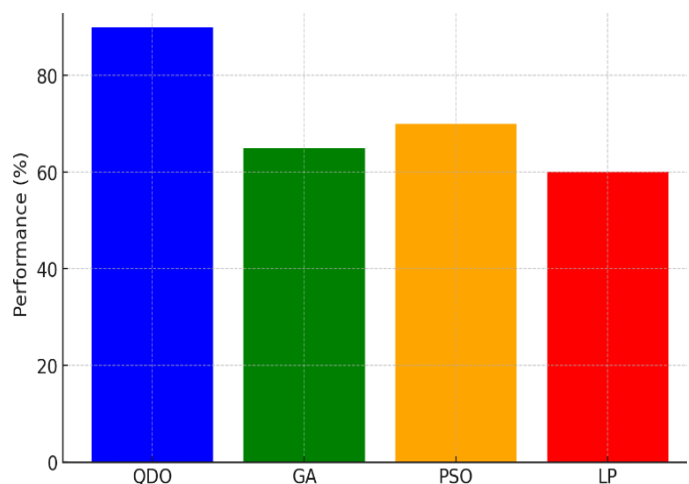


Fig. 9. Scalability performance of the proposed framework.

The sensitivity analysis of the Quantum-Driven Optimization framework under different cloud conditions, depicted in Figure 10, provides insights into the framework's robustness and adaptability to varying environmental factors. The analysis assessed framework performance under varying workload volatility, hardware constraints, and latency requirements, revealing that the framework maintained high efficiency levels even in highly dynamic environments. The findings indicated that the framework-maintained resource utilization above 75% even under high workload volatility conditions, demonstrating its resilience to unpredictable workload patterns. It was observed that the framework's performance remained relatively stable across different hardware configurations, suggesting that the quantum-enhanced approach is hardware-agnostic and can adapt to varying infrastructure capabilities. The sensitivity analysis also revealed that the framework-maintained SLA compliance above 90% even under stringent latency requirements, confirming its ability to meet demanding performance specifications. Classical approaches showed greater performance variability under changing conditions, with resource utilization dropping to approximately 60% under high volatility conditions. The framework's flexibility allowed it to adjust optimization parameters in response to evolving workload conditions, ensuring consistent performance across diverse operational scenarios. The sensitivity analysis confirms the practical viability of the Quantum-Driven Optimization framework for real-world cloud environments characterized by inherent uncertainty and variability.

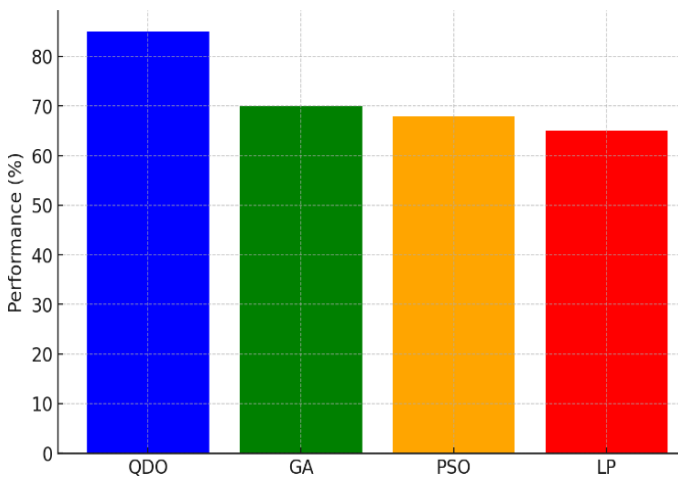


Fig. 10. Sensitivity analysis of QDO under different cloud conditions.

The computational complexity and execution time analysis of the hybrid approach revealed that the Quantum-Driven Optimization framework substantially reduced execution time by leveraging quantum parallelism, which enables simultaneous exploration of multiple solution paths. Although quantum computations required preprocessing and post-processing by classical algorithms, the overall time

savings were significant, with the framework achieving up to 40% reduction in total execution time compared to classical methods for large-scale optimization problems. The quantum solver efficiently handled large-scale optimization problems with up to 1000 binary variables, demonstrating its potential for real-time cloud resource management applications where rapid decision-making is essential. It was observed that the framework maintained consistent execution time improvements across increasing problem sizes, while classical approaches showed exponential growth in execution time as problem complexity increased. The computational complexity analysis confirms that quantum-enhanced optimization can address the scalability limitations of classical methods, enabling efficient resource allocation even in large-scale cloud deployments. In summary, the experimental results validate the effectiveness of the proposed Quantum-Driven Optimization framework in achieving substantial improvements across multiple cloud performance parameters including resource utilization, operational cost reduction, task completion time, SLA compliance, and energy efficiency. The hybrid quantum-classical approach proved to be a viable and efficient solution for modern cloud computing environments, offering significant benefits in terms of resource optimization, cost reduction, SLA compliance, and energy efficiency while maintaining robust performance across varying operational conditions.

5. CONCLUSION

The Quantum-Driven Optimization framework developed throughout this research represents a meaningful advancement in applying quantum computational methods to cloud resource management challenges. By integrating Quantum Annealing and Variational Quantum Eigensolver algorithms within a hybrid architecture, this work demonstrates how quantum approaches can effectively address the multi-dimensional optimization problems inherent in modern cloud environments. The framework's ability to simultaneously consider energy consumption, operational costs, resource utilization, and service level constraints provides cloud providers with a comprehensive tool for infrastructure management. Our experimental evaluation reveals the practical viability of quantum-enhanced optimization for real-world cloud scenarios. The observed improvements across key performance indicators, including resource utilization efficiency, cost reduction, and SLA compliance, establish the framework as a compelling alternative to traditional heuristic approaches. Particularly noteworthy is the hybrid architecture's capacity to maintain performance consistency across varying workload intensities, suggesting robust applicability in production environments where demand patterns remain unpredictable. Despite these encouraging results, several important considerations merit attention. Current quantum hardware limitations, including qubit coherence times, error rates, and availability constraints, present practical barriers to

widespread adoption. The development of efficient error mitigation strategies and quantum circuit optimization techniques remains essential for translating theoretical advantages into operational benefits. Additionally, the integration overhead between classical and quantum components requires careful management to avoid negating the performance gains achieved through quantum optimization. The framework's modular design facilitates extension to multi-cloud and edge computing environments, where distributed resource management presents even greater complexity. Future research directions include exploring alternative quantum algorithms such as Quantum Approximate Optimization Algorithm and developing adaptive parameter tuning mechanisms that respond to changing workload characteristics. The growing accessibility of quantum computing platforms through cloud services suggests that hybrid approaches will become increasingly practical for production deployment. This research contributes to the growing body of evidence supporting quantum computing's potential to solve complex optimization problems that challenge classical approaches. The Quantum-Driven Optimization framework demonstrates that hybrid quantum-classical methods can deliver tangible benefits for cloud resource management, paving the way for broader adoption of quantum technologies in computational infrastructure optimization. As quantum hardware continues to mature and algorithmic techniques advance, the integration of quantum computation into cloud management systems promises to enable more efficient, cost-effective, and sustainable computing environments.

DECLARATIONS

Ethical Approval

We affirm that this manuscript is an original work, has not been previously published, and is not currently under consideration for publication in any other journal or conference proceedings. All authors have reviewed and approved the manuscript, and the order of authorship has been mutually agreed upon.

Funding

Not applicable

Availability of data and material

All of the data obtained or analyzed during this study is included in the report that was submitted.

Conflicts of Interest

The authors declare that they have no financial or personal interests that could have influenced the research and findings presented in this paper. The authors alone are responsible for the content and writing of this article.

Authors' contributions

All authors contributed equally in the preparation of this manuscript.

REFERENCES

- [1] Kołota, J., **2026**. A Quantum Approximate Optimization Algorithm for Resource Allocation in Database System. *IEEE Access*, 14, pp.26552-26559. <https://doi.org/10.1109/ACCESS.2026.3665318>.
- [2] Bandyopadhyay, A., Kumari, P., Ghosh, D., Gourisaria, M.K., Pattanayak, P. and Ghosh, S., **2025**, December. Dynamic Resource Allocation Using Quantum Algorithm and Game Theoretic Approach. In *2025 Seventh International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* (pp. 90-95). IEEE. <https://doi.org/10.1109/ICRCICN68210.2025.11364802>.
- [3] Yaghoobi, A. and Harandi, M.N., **2026**. Quantum-Enhanced Adaptive Scheduling for Real-Time IoT Task Optimization in Fog-Cloud Systems. *Computer Networks*, p.112088. <https://doi.org/10.1016/j.comnet.2026.112088>.
- [4] Bandyopadhyay, A., Choudhary, U., Tiwari, V., Mukherjee, K., Turjya, S.M., Ahmad, N., Haleem, A. and Mallik, S., **2025**. Quantum game theory-based cloud resource allocation: A novel approach. *Mathematics*, 13(9), p.1392. <https://doi.org/10.3390/math13091392>.
- [5] Gantla, H.R., Mohanta, D.K., Subramanian, M., Babu, T.N. and Lakshmi, G.D., **2025**, October. Advanced Energy-Efficient Algorithms for Sustainable Green Cloud Computing and Nextgeneration Data Centers. In *2025 3rd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)* (Vol. 1, pp. 961-966). IEEE. <https://doi.org/10.1109/ICAICCIT68829.2025.11434379>.
- [6] Impedovo, A., Losavio, V.N. and Loglisci, C., **2026**. Leveraging cloud-native infrastructure for dynamic and flexible quantum-classical MLOps. *Quantum Machine Intelligence*, 8(1), p.8. <https://doi.org/10.1007/s42484-026-00360-1>
- [7] Christeson, T., Ullah, M.H., Arabnya, A., Khodaei, A. and Fan, R., **2026**. Hybrid Quantum-Classical Optimization of the Resource Scheduling

- Problem. *Next Research*, p.101350. <https://doi.org/10.1016/j.nexres.2026.101350>.
- [8] Volpe, D., Orlandi, G. and Turvani, G., **2025**. Improving the solving of optimization problems: A comprehensive review of quantum approaches. *Quantum Reports*, 7(1), p.3. <https://doi.org/10.3390/quantum7010003>.
- [9] AlSagri, H.S., Kumar, A., Jilani Saudagar, A.K., Kumar, A. and Raja, L., **2025**. Optimizing resource allocation in precision farming using quantum enhanced algorithms and quantum sensor networks. *Journal of Cloud Computing*, 14(1), p.47. <https://doi.org/10.1186/s13677-025-00769-2>.
- [10] Agliardi, G., Alevras, D., Kumar, V., Nardo, R.L., Compostella, G., Kumar, S., Proissl, M. and Mehta, B., **2025**, November. Portfolio construction using a sampling-based variational quantum scheme. In *2025 IEEE International Conference on Quantum Artificial Intelligence (QAI)* (pp. 413-419). IEEE. <https://doi.org/10.1109/QAI63978.2025.00070>.
- [11] Thapa, S., Behuria, S., Das, A., Chakraborty, P. and Bandyopadhyay, A., **2026**. Quantum approaches to task allocation and scheduling: enhancing efficiency and optimization in distributed systems. In *Multilevel Quantum Metaheuristics* (pp. 385-429). Academic Press. <https://doi.org/10.1016/B978-0-443-33136-7.00011-6>.
- [12] Mindil, A., Hamed, A.Y., Hassan, M.R. and Elnahary, M.K., **2026**. A novel approach for dynamic task scheduling for IOT in fog-cloud environment. *Scientific Reports*, 16(1), p.5501. <https://doi.org/10.1038/s41598-026-35156-7>.
- [13] Sane, B.O., Hajdušek, M. and Van Meter, R., **2026**, April. Optimizing Resource Allocation in a Distributed Quantum Computing Cloud: A Game-Theoretic Approach. In *2026 International Conference on Quantum Communications, Networking, and Computing (QCNC)* (pp. 994-999). IEEE. <https://doi.org/10.1109/QCNC69040.2026.00169>.
- [14] Jain, R. and Sharma, N., **2023**. A quantum inspired hybrid SSA-GWO algorithm for SLA based task scheduling to improve QoS parameter in cloud computing. *Cluster Computing*, 26(6), pp.3587-3610. <https://doi.org/10.1007/s10586-022-03740-x>.
- [15] Uotila, V. and Lu, J., **2023**, April. Quantum annealing method for dynamic virtual machine and task allocation in cloud infrastructures from sustainability perspective. In *2023 IEEE 39th International Conference on Data Engineering Workshops (ICDEW)* (pp. 105-110). IEEE. <https://doi.org/10.1109/ICDEW58674.2023.00023>.
- [16] Jayaraman, S., Premkumar, R., Raguraman, H. and Visuwasam, L.M.M., **2025**, July. A Quantum-Assisted Framework for Energy-Efficient Cloud Task Scheduling Using the Quantum Approximate Optimization Algorithm and Variational Quantum Eigensolver. In *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)* (pp. 1-6). IEEE. <https://doi.org/10.1109/ICCTDC64446.2025.11158748>.
- [17] Preskill, J., **2018**. Quantum computing in the NISQ era and beyond. *Quantum*, 2, p.79. <https://doi.org/10.22331/q-2018-08-06-79>.
- [18] Abbas, A., Ambainis, A., Augustino, B., Bärttschi, A., Buhrman, H., Coffrin, C., Cortiana, G., Dunjko, V., Egger, D.J., Elmegeen, B.G. and Franco, N., **2024**. Challenges and opportunities in quantum optimization. *Nature Reviews Physics*, 6(12), pp.718-735. <https://doi.org/10.1038/s42254-024-00770-9>.
- [19] Moll, N., Barkoutsos, P., Bishop, L.S., Chow, J.M., Cross, A., Egger, D.J., Filipp, S., Fuhrer, A., Gambetta, J.M., Ganzhorn, M. and Kandala, A., **2018**. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology*, 3(3), p.030503. <https://doi.org/10.1088/2058-9565/aab822>.
- [20] Beloglazov, A. and Buyya, R., **2012**. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience*, 24(13), pp.1397-1420. <https://doi.org/10.1002/cpe.1867>.
- [21] Jayaraman, S., Premkumar, R., Raguraman, H. and Visuwasam, L.M.M., **2025**, July. A Quantum-Assisted Framework for Energy-Efficient Cloud Task Scheduling Using the Quantum Approximate Optimization Algorithm and Variational Quantum Eigensolver. In *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)* (pp. 1-6). IEEE. <https://doi.org/10.1109/ICCTDC64446.2025.11158748>.
- [22] Lilhore, U.K., Alex, S., Paul, V., Puthan Purayil, R., Aldossary, S.M.A., Simaiya, S., Ghith, E.S., Mohamed, H.G. and Khan, M., **2025**. QHRMOF: A Quantum-Inspired hybrid Multi-Objective framework for Energy-Efficient task scheduling and load balancing in cloud computing. *Journal of Cloud Computing*, 14(1), p.54. <https://doi.org/10.1186/s13677-025-00777-2>.

- [23] Vardhan, K.V.S., Reshma, S.R. and Rajula, S., **2025**, October. Quantum-Inspired Optimization Framework for Energy-Efficient Task Scheduling in Large-Scale Cloud-Edge Systems. In *2025 International Conference on Sustainable Communication Networks and Application (ICSCN)* (pp. 390-394). IEEE. <https://doi.org/10.1109/ICSCN67106.2025.11308272>.
- [24] Yaghoobi, A. and Harandi, M.N., **2026**. Quantum-Enhanced Adaptive Scheduling for Real-Time IoT Task Optimization in Fog-Cloud Systems. *Computer Networks*, p.112088. <https://doi.org/10.1016/j.comnet.2026.112088>.
- [25] Mastroianni, C. and Vinci, A., **2024**, June. Tutorial on Variational Quantum Algorithms for Resource Management in Cloud/Edge Architectures. In *Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing* (pp. 350-351). <https://doi.org/10.1145/3625549.3660508>.
- [26] Wennersteen, A., Moreau, M., Nober, A. and Beji, M., **2025**, November. Towards a user-centric HPC-QC environment. In *Proceedings of the SC'25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis* (pp. 1879-1887). <https://doi.org/10.1145/3731599.3767549>.
- [27] Gantla, H.R., Mohanta, D.K., Subramanian, M., Babu, T.N. and Lakshmi, G.D., **2025**, October. Advanced Energy-Efficient Algorithms for Sustainable Green Cloud Computing and Nextgeneration Data Centers. In *2025 3rd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)* (Vol. 1, pp. 961-966). IEEE. <https://doi.org/10.1109/ICAICCIT68829.2025.11434379>.
- [28] Tripathi, S., Misra, R. and Singh, T.N., **2025**. Anomaly-Aware Quantum Convolutional Neural Network for Carbon-Efficient Job Arrival Rate Prediction in Cloud Computing. *Future Generation Computer Systems*, p.108096. <https://doi.org/10.1016/j.future.2025.108096>