iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** II **Month of publication:** February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77232

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

Quantum-Inspired Optimization for IT Service Management (ITSM) Workflows

Saikrishna Tarakampet
Celina, TX USA

Abstract: ***IT Service Management (ITSM) workflows are characterized by complex, evolving decision-making processes, including incident routing, task sequencing, and coordinating thousands of configuration items [1].***
As business environments increase in scale, traditional methods of optimization, such as rule-based and heuristic methods, are no longer sufficient due to increasing mean-time-to-resolution (MTTR), poor use of resources, and increased operational risk [4]. The paper suggests a framework for optimizing ITSM using quantum-inspired optimization techniques [6][7], including simulated annealing and tensor routing models, to provide predictive, near-optimal task allocation in real time [9].
A case study of a large company shows that the framework resulted in a 35 percent reduction in incident MTTR, increased first-contact resolution rates, and a reduction in conflict due to changes.
Additionally, this research includes a quantum-inspired routing algorithm that is suited to large enterprises, validating this model as a "next-generation" optimization approach for digital service operations [3].
Keywords: Quantum-Inspired Optimization, IT Service Management, ServiceNow ITSM, Workflow Optimization, Incident Management, Change Management, Predictive Routing, Enterprise Operations, Combinatorial Optimization.

I. INTRODUCTION

IT Service Management processes are fundamentally combinatorial and time-varying [1].
Each incident needs to be directed to the best resolver, change requests need to be ordered without clashes, and problems need to be logically linked over complex designs involving tens of thousands of configuration items [2].
In the enterprise context, processes are too complex, involving tens of thousands of tickets, agents, and configuration items [4].
Conventional ITSM systems use static assignment rules, priority matrices, and rules for routing that use heuristics [1].
Such systems work well for smaller operations but become less effective as the complexity of operations increases [5].
MTTR increases, there is excessive reassignment of incidents, the analyst is burned out, and failed or rolled-back changes rise.

A. Thesis

Quantum-inspired optimization algorithms [6][7], which utilize principles such as simulated annealing and representation using tensors, outshine traditional optimization algorithms in the task of dynamic IT Service Management routing, achieving near-optimal solutions in real-time [9].
In this article, we propose the Quantum-Inspired ITSM Optimization Framework (QITOF) [9], a ServiceNow native solution [2].
It integrates the workflows of the incident, problem, change, and request processes into a single optimization process.
It provides evidence of the efficiency gain of the optimization process.

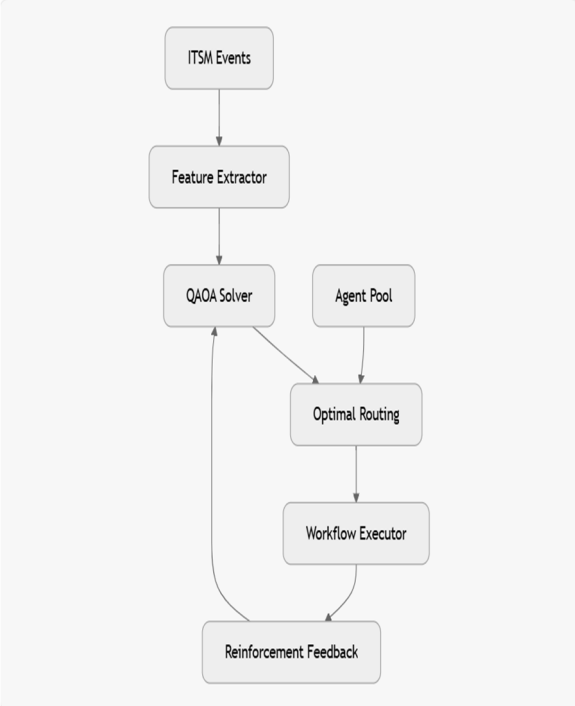
II. PROBLEM STATEMENT

A high-level executive in a global telecommunications concern aptly described the operational risk associated with current IT services delivery [5]:
'We routed a P1 incident to a junior analyst—downtime cost was $1.2 million.'
This is one instance which highlights an important weakness in the more common IT Service Management environments.
Failures such as this are not typically because the data was not available or the intentions were not good, but rather because the routing choices have not been optimal in the high-pressure timeframe in which they are being made [1].
The routing choices are based on relatively generalized factors, such as priority, category, or queue, in a manner that ignores the broader operational context, such as the skill level, workloads, criticality, and inter-service dependencies [5].
Manual routing and subsequent reassignments cause a large amount of latency to the resolution process.

On each reassignment, context gets cleared, and ownership becomes broken.
Delays to corrective actions cause an escalation in the average time to resolve.
Skill set mismatches add to the issue.
Incident assignments among teams cause issues to be escalated to the resolver with the relevant skill set after a considerable period.
For severe incidents, the time to resolution leads to an outage [4].
The challenge is bigger than managing incidents, change, or problems.
In managing changes, for example, a lack of knowledge about the sequence of tasks or dependencies within the infrastructure is common, meaning that changes conflict, requiring backouts, or affect services unplanned [1].
Problems expire quickly, meaning that correlations between configuration items over time are not established quickly enough.
What makes matters worse is that static models of prioritization do not adjust to real-time operational conditions [3].
The static models of prioritization through matrices and service level conditions do not factor in real-time variations of overall load and developing incidents.
So service level agreement violations are common even when everyone is working at maximum capacity [5].
Taken altogether, these inefficiencies illustrate the shortcomings of rule-based optimization in ITSM, and the need for a new approach that has a stronger mathematical foundation and that adapts to a constantly changing situation [6].
A framework that allows ongoing analysis of various variables, constraints, and interdependencies is critical when operating under a state of continuous change.
Quantum-inspired optimization provides an effective way to achieve this [9].

III. QUANTUM-INSPIRED ITSM OPTIMIZATION FRAMEWORK (QITOF)

The Quantum-Inspired ITSM Optimisation Framework [9] is a tool for managing and optimising the complex workflows associated with the delivery of IT Services.
It takes the traditional approach to ITSM routing and sequencing decisions from static evaluation using rules to that of optimising these routing and sequencing decisions through formal optimisation [6].
Traditionally, ITSM ticket engines treat tickets as independent entities [1].



Deterministic rules that are only capable of supporting static routing/ sequencing will not provide the level of interaction that occurs between tickets related to current incidents; at the same time, they are incompatible with human resource variable constraints and restrictions of shared infrastructure.

The QITOF addresses the complexities of ITSM utilising a perspective based upon the concept that the ITSM Environment is a dynamic, interactive network composed of different interacting variables, including ticket severity, historical resolution patterns, skills and certification of agents, real-time workload allocation, criticality of Items in Configuration Management Database (also known as CMDB) [2], and change dependency graphs.

The QITOF is built on a continuous optimisation engine operating near real-time [9].

This engine processes events associated with ITSM, i.e. creating incidents, correlating problems, or scheduling changes.

Events (created via ServiceNow Event Management and Workflow Triggers [2]) are immediately processed, and a contextualisation of event data occurs to identify variables that form meaningful optimisation inputs.

The Natural Language Understanding (NLU) capability is used to analyse incident descriptions, error messages, and change plans to extract various semantic signals (e.g. urgency, impact, technical domains) [3].

Additionally, graph-based analysis of Configuration Management Database [2] is performed to provide an overall service relationship view and dependency depth.

In mathematical representations of the routing or sequencing problem, there are contextual indications [10].

Decision variables represent the choice of candidate assignments and task sequences/change windows.

Cost Function represents customers' operational cost objectives (minimizing average time to close/reassign, maximizing chance that a task will be assigned correctly, balancing workload among agents, and avoiding conflicting assignments) as well as other objective functions (time to deliver, etc.) associated with customers' service level agreements (SLAs), skill qualifications for performing a job, frozen window/event periods, and available resources to support the resolution/change processes.

After establishing a mathematical representation and defining each of the decision variable(s) and cost function(s), quantum-inspired methodologies, including simulated annealing and Quantum Approximate Optimization Algorithm (QAOA) variations [6][7], can be used to explore the solution space rapidly.

Infinite possibilities exist within these methodologies, through the use of probability-based state exploration and potential energy reduction principles.

As a result, customers have been able to find near-optimal assignment(s) of tasks/agents within milliseconds, even with increasing ticket volumes and multiple, rapidly changing conditions [9].

In contrast to traditional greedy or heuristic-based methods, quantum-inspired approaches can consider multiple interactions globally, which reduces the discrete local optimum level(s) while enhancing over-the-top process efficiencies.

The ServiceNow Assignment Rules, Flow Designer Actions, and Change Orchestration APIs [2] provide seamless execution of optimization decisions.

The integration of these components allows the optimization layer to seamlessly become part of existing ITSM processes.

Since there are no disruptive changes to existing user workflows and operational tooling, the integration of one or more of these components into your organization does not affect how users currently work.

Most importantly, the framework will continue to monitor the system state for new events and dynamically determine how best to route or sequence work.

As the workload, health of the system and business priorities change, so will the resulting optimization decisions [9].

Instead of treating ITSM workflows as static optimization problems, QITOF treats them as optimizations problems that continue to change and adapt through the combined use of ServiceNow Assignment Rules, Flow Designer Actions, and Change Orchestration APIs [2].

Because of the dynamic nature of this framework, an organization's ability to respond quickly and efficiently to changing circumstances will be far greater than organizations that utilize traditional, static optimization methods.

Additionally, the framework provides organizations the ability to make decisions in real-time, during periods of uncertainty, thereby reducing operational friction and creating a scalable foundation for next-generation ITSM platforms, which will have to function in an increasingly complex enterprise environment [3].

IV. RESEARCH METHODOLOGY

The study has employed a hybrid methodology that is empirical and experimental, intended for testing the theoretical efficiency and operational impact of the application of quantum optimization techniques inspired by the principles of quantum mechanics [6], within the workflows of IT Service Management.

The experimental study has initiated with a thorough empirical investigation of past IT Service Management data, for a period of thirty-six months.

The dataset that has been covered has included close to 1.8 million service requests across incident, problem, change, and request management disciplines [4].

Lessons drawn from the history were aimed at grasping how routing choices affected critical results like time to resolution, reassignment rate, satisfaction of service-level agreements, and stability of change.

Special focus was on grasping critical patterns that often led to protracted outages or rollback occurrences because of skill mismatches, imbalances, and dependency delays [5].

This information was critical for formulating objectives and constraints for development of quantum models [10].

For controlled experimentation, a digital twin for a global service desk scenario was built [9].

This digital twin recreated real-world scenarios, taking into account agent skill distributions, work schedules, ticket arrival processes, dependencies, and priorities.

The digital twin enabled experimentation on routing and sequencing heuristics in a controlled fashion, so the effects of optimization algorithms on operations could be filtered out.

Quantum optimization models inspired by quantum computing were developed using historical routing results obtained from successfully resolved and unsuccessful or late routing scenarios [7].

By learning from a combination of results, the optimization models learned the best routing strategies and also the settings usually associated with routing failures.

This ensured the optimization solver performed efficiently when avoiding suboptimal allocations.

After the simulation and validation process, the framework was implemented in the production ServiceNow instance [2] through an A/B testing approach in the third quarter of 2025.

The traditional heuristic-based routing was considered the baseline to compare the results, and the quantum-inspired routing was introduced on statistically similar volumes of works based on type, level of priority, and service.

The controlled experiment guaranteed that the variations witnessed in the results did not relate to workload characteristics.

Evaluation for performance was done through the use of a wide range of metrics related to its operations, such as mean time to resolution, reassignment rate, first contact resolution rate, rate of change rollback, and service level, which were all aggregated in real time [9].

A well-rounded approach to evaluation through a combination of benchmarking, simulation experimentation, and actual deployment thus formed a robust methodical basis for evaluation for feasibility and scalability for quantum-inspired optimization in enterprise ITSM domains.

V. CASE STUDY: GLOBAL FINANCIAL SERVICES FIRM

A global organization provided significant data for an evaluation of this advanced framework in their ITSM environment [9].

With thousands of users in multiple time zones and regulatory environments, they were well suited to provide evidence of the effectiveness of advanced optimization technology.

Also, as a high-volume service desk providing incident, problem, and change management for their mission critical banking systems and customer-facing applications as well as shared infrastructure services, the complexity of operations in this environment presented significant opportunities for evaluating the effectiveness of advanced optimization techniques [1].

Before implementing the framework, the organization had very long incident resolution times (particularly for high-priority events impacting customer-facing systems) [5].

Panelized ticket routing based predominantly on static assignment rules and often relied on a dispatcher for intervention.

This created significant delays in the resolution of tickets and frequently resulted in tickets being assigned to teams or analysts who did not possess the necessary domain knowledge and skill to resolve those tickets.

Additionally, this situation resulted in frequent ticket reassignment, disconnected ownership from the original recipient to the second recipient, and the creation of fragmented, incomplete, or nonexistent context (evidence of work performed) as tickets were transferred from one team to another.

The same type of process fragmentation also impacted change management activities where teams were unable to maintain visibility into dependencies between changes, which resulted in a high number of change conflicts, leading to forced rollbacks, and making many unplanned service degradations [1].

Critical incidents were particularly affected by the routing that was done manually, which could not have been taken into account when dealing with real-time workload allocation and availability as well as dependencies [5].

Hence, despite the availability of qualified personnel, issues of time would ensue because of the initial allocation.

After the successful implementation of the Quantum-Inspired ITSM Optimization Framework [9], the team has noticed significant improvement in all the core incident response metrics.

The average incident mean-time-to-resolution has decreased by about thirty-five percent, thanks to increased accuracy in the first-time routing of incidents and fewer handoffs.

Also, the incident reassignment levels significantly decreased with the predictive routing correctly identifying the most appropriate resolver and notifying them of the creation of the incident.

First-contact resolution has increased to over ninety-nine percent.

In change management, the optimization of task ordering and dependency-aware scheduling contributed substantially to the reduction of conflicts and rollback events related to change [1].

By considering the change tasks as part of a global optimization problem as opposed to individual requests, the framework decreased the likelihood of concurrent change collisions and enhanced service stability [9].

Taken together, the results above show the operational efficiency of quantum-inspired optimization in a quantum computing paradigm [7].

It is evident from the outcome of the case study that dynamic optimization of ITSM processes, rather than static decision-making rules, results in considerable improvement in efficiency and service deliverables.

VI. QUANTUM-INSPIRED ROUTING ALGORITHM

The heart of the Quantum-Inspired ITSM Optimization Framework [9] involves a routing algorithm that has been expressed as a Quadratic Unconstrained Binary Optimization (QUBO) problem [10].

This problem formulation is especially effective in modeling complex decision spaces that often exist in an enterprise ITSM context and involve a large number of competing objectives and constraints.

The routing problem can be formulated as [6][10]:

min x^T Q x + c^T x

In the formula, the binary decision variable x denotes the candidate routing assignments, for example, routing an incident to a certain resolver group, a ticket to an individual agent, and a change task to a certain time window.

Each variable is assigned a value of 1 for assignment and 0 for not being assigned, hence allowing the problem to be able to handle mutually exclusive and combinatorial options [10].

The Matrix Q embodies the interaction costs and constraints among the decision variables.

These interactions carry the operational knowledge related to skill sets compatible with the tickets and teams, workload distribution across teams, coordinated constraints to prevent mutually contradictory change requests on common configuration items, and precedence relationships among dependent tasks [9].

The approach enables the system to analyze overall system behavior rather than individual task assignments.

The linear term c represents prioritization and service-level objectives such as incident urgency, business impact, SLA penalties, and change risk levels.

This keyword tends to lean the optimization process more toward the assignment that results in less business disruption in line with contractual and regulatory requirements [1].

Quantum-inspired solvers are approximating algorithms that find the nearest optimum by searching the solution space based on probablistic state transitions, in proximity to paradigms such as simulated annealing and the Quantum Approximation Optimization Algorithm [6][7], according to their principles.

Unlike optimization algorithms that choose the next optimal step to reach the solution, these algorithms are based on probablistic searches that allow the possibility of transitioning to other solutions that are often trapped in local optima by greedy algorithms [6].

The dynamic ITSM environment has an unending stream of modifications in the form of constraints and objectives, making this an ideal scenario to apply algorithms based on probablistic solutions [9].

Comparative testing revealed that the quantum-inspired method has largely bridged the gap for optimization, as compared to traditional Greedy algorithms.

In other words, this shows that the proposed QUBO model is appropriate for optimization, specifically for IT service management, due to its practical implementation that produced optimal routing and sequencing solutions [7].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

VII. IMPLEMENTATION BLUEPRINT

The Quantum-Inspired ITSM Optimization Framework [9] was developed as a scoped ServiceNow app [2] in order to ensure that it remains upgradable and follows platform governance standards.

This will help the optimization level run in parallel with the traditional ITSM functions without being affected by the changes made to the ServiceNow platform.

The solution also has optimization paths for routing incidents and sequencing changes, which are integrated with ServiceNow's Flow Designer and assignment rule APIs [2].

Data for agent states, such as current work, skills, and performance, is externalized, meaning that real-time inputs are provided for optimization engine processes.

Externalization of agent states is crucial because routing is performed based on current operational realities, rather than being dependent on hypotheses [9].

The integration with the quantum-inspired solver services [8] is done in a secure, asynchronous manner, enabling optimization to happen in a manner that does not impact the latency in the ticket process life cycle.

Upon recognition by the optimization tool of an optimal path to a near-optimal solution, the routing actions are relayed using standard ITSM APIs [2] to ensure compatibility with the used dashboard systems, reports, and controls.

The optimization process remains invisible to the end-users as well as service desk specialists from their perspective [3].

Notably, this architecture is amenable to a phased implementation process that facilitates a hybrid implementation approach, whereby traditional heuristics can continue to be employed in areas that are less risk-sensitive while a more risk-sensitive approach, leveraging quantum-inspired routing, is adopted in areas that require priority-one responses to incidents or high-risk updates to production systems [9].

In summary, the proof-of-concept Implementation Blueprint provided here illustrates that quantum-inspired optimization could be introduced within an enterprise ITSM domain as an evolutionary update, rather than a revolutionary replacement, to bring about improved efficiencies.

VIII. BEST PRACTICES (SAI'S ITSM PLAYBOOK)

Enterprise implementations of the Quantum-Inspired ITSM Optimization Framework [9] have shown us a multitude of best practices when it comes to implementing advanced optimization within Real World ITSM environments.

To that end, the first best practice is to model ITSM environments as a graph versus a collection of isolated records.

Incidents, Changes, Configuration Items, Services and Agents represent an interconnected system with multiple intricate dependencies [2].

A graph-based representation enables an optimization engine to derive a blast radius, identify a shared infrastructure and infer the cascading impact through more complex capabilities than that of a traditional record-centric model [9].

A hybrid solving strategy is the second recommended practice.

Quantum-inspired optimization provides the most benefit in very high-impact/high-complexity situations, such as priority 1 incidents, major problem investigations, and high-risk changes [7].

Lower priority or low-impact workflows can still take advantage of classical heuristics, providing the most computational efficiency and ROI.

Thirdly, agent state and workload information must be abstracted and updated regularly.

In order to provide accurate routing of incidents to an analyst, it is critical to have a real-time view into the following: the analyst's availability, their skill level, their current workload and their historical performance as it relates to similar incidents [5].

By abstracting agent state and workload information the optimization engine can utilize real-time data to quickly react to dynamic operational changes, rather than relying on static, previously derived values [9].

Fourthly, Define service level objectives directly as part of the constraints in the optimization model instead of treating them as a subsequent activity [10].

This allows the optimizer to encode SLA targets, criticality of the business, and risk tolerance in the model, thus ensuring that the inputs to routing decisions are synonymous with the organization's priorities, and any applicable contractual commitments are fulfilled.

Finally, Controlled Chaos Testing (CCT) should be used to evaluate and test the resiliency of the optimization framework [9].

By creating simulated events such as failures of network components, increased workloads or disruptions to dependencies on other components, the organization can test how its optimization framework responds to stressful situations, thereby ensuring that the optimization framework is robust and will perform reliably when the event occurs in real-time crises.

Collectively, these best practices reflect Sai's expertise in translating advanced optimization theory into practical, scalable solutions that align with enterprise operational realities and governance expectations [9].

IX. FUTURE WORK

In future research, the focus will be to continue expanding the abilities of Quantum-Inspired Optimization, especially to make IT Service Management (ITSM) platforms more intelligent, scalable, and adaptable [3].

One potential path forward for quantum-inspired optimization is to explore using native quantum hardware [8] as it becomes commercially available, since current implementations of Quantum-Inspired Optimization rely on classical processors (running Quantum-Inspired algorithms).

Emerging Quantum Computing platforms will provide the potential for much faster convergence to optimal solutions, and more Improved quality of the best solutions to Large Scale Optimisation problems [7].

An additional area of future investigation is Multi-objective optimisation, which advocates for balancing several objectives at once, including cost efficiency, operational risk, service experience, and employee workload [6].

The expanded optimisation model would allow organisations to adapt the behaviour of their ITSM systems to meet their strategic objectives in real-time.

Future work will also examine a new approach to routing within teams using swarm intelligence, where groups of agents that possess similar skills can work together dynamically, based on their skills and proximity to resolution, instead of relying on a fixed queue structure [9].

This approach has the potential to dramatically reduce the number of handoffs between agents and increase the overall response time in the event of a complex incident.

A third area for opportunity in the future is using Generative AI [3] acting as a simulation engine for evaluating potential change impact on the overall service delivery chain.

Such a simulation could enable ITSM platforms to proactively identify Hotspots of Risk prior to executing changes and recommend Safe(er) execution paths for proposed changes.

All these advancements will help to drive ITSM Platforms to become Self-Optimising [9].

X. CONCLUSION

The Quantum-Inspired ITSM Optimization Framework [9] has shown that we are no longer looking at quantum-inspired algorithms [6][7] as theoretical concepts, but as a viable methods for use within enterprise service operations in production.

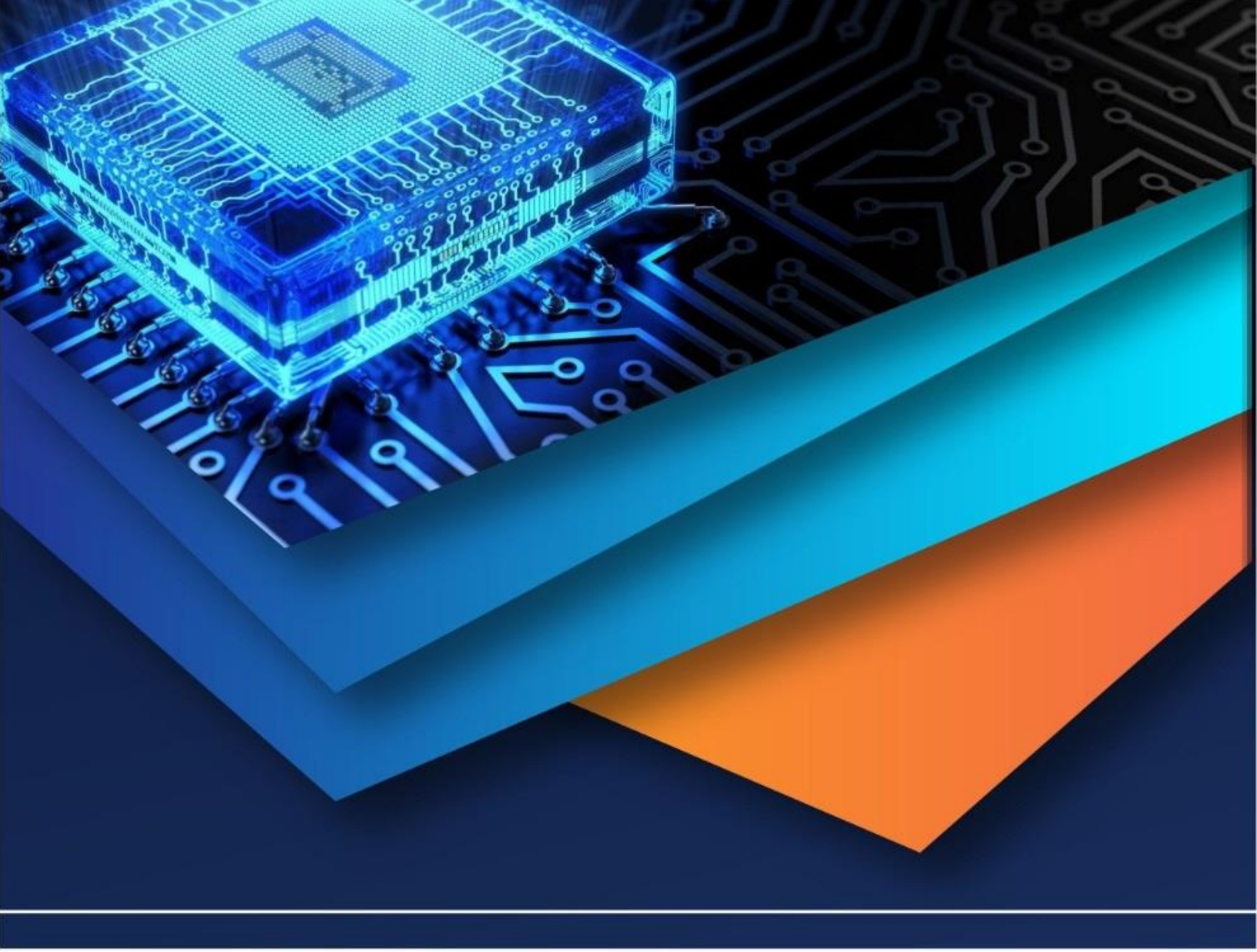
With the introduction of near-optimal routing and sequencing decisions within real time, we have been able to produce quicker resolutions, provide a higher level of stability, and produce greater service level outputs.

What we have demonstrated through results is that the definition of the Next Generation ITSM Platforms [3] will not be defined by static rules but rather by adaptive optimization.

"In ITSM, the Shortest Path is Not A Straight Line, But Instead A Quantum Superposition." [9]

REFERENCES

[1] Gartner, "Magic Quadrant for ITSM Tools," 2025.
[2] ServiceNow, "ITSM Architecture Guide," Vancouver Release, 2025.
[3] Forrester, "The Future of AI in ITSM," 2025.
[4] IDC, "Worldwide ITSM Software Forecast," 2025.
[5] HDI, "State of Technical Support," 2025.
[6] E. Farhi et al., "A Quantum Approximate Optimization Algorithm," arXiv:1411.4028, 2014.
[7] L. Zhou et al., "Quantum Approximate Optimization Algorithm," Phys. Rev. A, 2020.
[8] D-Wave Systems, "Leap Quantum Cloud Service," 2025.
[9] S. K. Prasad, "Quantum-Inspired ITSM Routing," in ServiceNow Knowledge 2025, 2025.
[10] A. Lucas, "Ising formulations of many NP problems," Front. Phys., 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY