Journal of Mathematical Problems, Equations and Statistics

E-ISSN: 2709-9407
P-ISSN: 2709-9393
Impact Factor (RJIF): 5.94
JMPES 2026; 7(1): 19-31

www.mathematicaljournal.com
Received: 20-11-2025
Accepted: 25-12-2025

Nasim Akhtar
School Teacher (11-12),
Department of Mathematics,
UMV Jhauwan Dighwara,
Saran, Bihar, India

Corresponding Author:
Nasim Akhtar
School Teacher (11-12),
Department of Mathematics,
UMV Jhauwan Dighwara,
Saran, Bihar, India

An integrated study of fuzzy mathematical programming and queueing models with applications to real-world operational systems

Nasim Akhtar

DOI: https://www.doi.org/10.22271/math.2026.v7.i1a.283

Abstract
Service and production systems such as banking counters, healthcare registration units, and call centers operate under substantial uncertainty in arrivals, service rates, and cost parameters, while being required to meet explicit service-level targets. Classical queueingbased optimization typically assumes precisely known parameters, which limits its applicability in data-scarce or expert-driven environments. This paper addresses this gap by developing an integrated framework that combines fuzzy mathematical programming with queueing models, allowing congestion-aware decisions to be made under imprecise informatins.

Fuzzy parameters are modeled using membership functions, and system performance measures derived from queueing theory are embedded directly into the optimization layer through either α-cut decomposition or a satisfaction-level (λ) maximization approach. The resulting model captures trade-offs between operational cost and service quality while preserving interpretability of ambiguity. A real-world service-system case study, motivated by a single-station multi-server operational setting, demonstrates the practical implementation of the proposed framework.

Numerical results indicate that the integrated fuzzy-queueing approach yields solutions that are more robust and managerially transparent than crisp benchmarks, particularly when service-level constraints are critical. The framework supports informed staffing and capacity decisions and offers actionable insights for practitioners managing uncertainty in operational systems.

Keywords: Fuzzy mathematical programming, queueing theory, α-cut method, service-level optimization, operational uncertainty, service systems, decision analytics

Introduction

Modern operational systems are increasingly characterized by high demand variability, tight service-level expectations, and limited tolerance for congestion. Service-oriented environments such as bank branches, hospital outpatient registration units, call centers, transportation hubs, and public service facilities routinely face the challenge of balancing operational efficiency with customer satisfaction. In these systems, decision makers must determine appropriate staffing levels, service capacities, and scheduling policies while accounting for uncertain arrivals, fluctuating service times, and ambiguous cost structures. Queueing theory has long served as a fundamental analytical tool for modeling congestion and delay phenomena in such settings, offering explicit performance measures such as expected waiting time, queue length, and system utilization [5, 9].

Despite its analytical strength, classical queueing-based decision models typically assume that key parameters arrival rates, service rates, and cost coefficients are precisely known. In practice, this assumption is rarely satisfied. For example, a bank branch may experience seasonal demand patterns, walk-in variability, and behavioral uncertainty that cannot be accurately captured by a single arrival-rate estimate. Similarly, hospital registration desks are influenced by physician schedules, patient mix, and emergency interruptions, while call centers face stochastic call volumes driven by marketing campaigns, system outages, or external events. In such contexts, historical data may be sparse, outdated, or nonstationary, and expert judgments often play a central role in operational planning. As a result, parameter uncertainty becomes an inherent feature of real-world service systems rather than a secondary modeling inconvenience.

To address uncertainty, several modeling paradigms have been proposed, including stochastic programming, robust optimization, and simulation-based approaches. While these frameworks are powerful, they frequently require strong distributional assumptions, large data samples, or complex scenario generation procedures, which may not be feasible for day-to-day operational decision making. Moreover, such approaches can obscure the interpretability of uncertainty, making it difficult for managers to understand how subjective assessments and operational preferences influence final decisions. Fuzzy set theory offers an alternative and complementary perspective by explicitly modeling ambiguity through membership functions that capture linguistic assessments such as "approximately high arrival rate" or "acceptable waiting time" [10, 3]. Instead of forcing uncertain parameters into precise probabilistic forms, fuzzy modeling allows decision makers to represent imprecision in a transparent and flexible manner.

Fuzzy mathematical programming has evolved as a structured methodology for decision problems involving vague objectives and constraints. Since the seminal work on fuzzy linear programming, a wide range of models has been developed to handle fuzzy goals, fuzzy right-hand sides, and multi-objective trade-offs using satisfaction levels or α-cut decompositions [11, 1]. These models are particularly attractive in operational contexts where performance targets are often expressed in imprecise terms, such as "waiting time should be short" or "service quality should be high." However, in many existing applications, the performance measures embedded in fuzzy optimization models are either static or simplified proxies, with limited connection to the dynamic congestion behavior captured by queueing theory.

This disconnect highlights a critical methodological gap. Queueing models provide rigorous relationships between system design variables (such as number of servers) and congestion outcomes, but they struggle to accommodate imprecise inputs. Fuzzy optimization models handle ambiguity effectively, but they often lack realistic performance mappings when congestion effects are central to the system. Treating these two paradigms separately can lead to suboptimal or misleading decisions. For instance, optimizing staffing levels based solely on fuzzy cost considerations without embedding queueing-based waiting-time relations may yield solutions that violate service-level expectations. Conversely, designing queueing systems using crisp parameter values may underestimate congestion risks under uncertainty. An integrated approach that combines fuzzy mathematical programming with queueing models is therefore essential to support robust and interpretable operational decisions.

The need for such integration is particularly evident in service systems where congestion costs and service quality are tightly coupled. In a hospital registration unit, excessive waiting times can lead to patient dissatisfaction and downstream delays in clinical workflows, while overstaffing increases operational costs. In a call center, meeting contractual service-level agreements (Such as answering a given percentage of calls within a target time) is crucial, yet call volumes and handling times are inherently uncertain. Bank branches face similar trade-offs between teller staffing costs and customer waiting experiences. In all these examples, decision makers operate under partial information and rely on expert judgment alongside limited data. An integrated fuzzy-queueing framework enables these uncertainties to be explicitly represented while preserving the analytical structure needed for performance evaluation.

From a methodological standpoint, integration can be achieved by embedding queueingperformance relations directly into fuzzy mathematical programming formulations. Expected waiting times, queue lengths, or delay probabilities derived from queueing theory can be treated as fuzzy-valued functions when arrival and service parameters are fuzzy. These fuzzy performance measures can then appear in fuzzy constraints or objectives, handled through α- cut decomposition or satisfaction-level (λ) maximization techniques [11, 3]. Such an approach allows the optimization model to account for congestion effects in a manner that is consistent with the representation of uncertainty. Importantly, it also facilitates sensitivity analysis with respect to ambiguity levels, enabling decision makers to explore conservative and optimistic planning scenarios.

This paper develops a unified framework that integrates fuzzy mathematical programming and queueing models for real-world operational systems. The proposed approach is designed to be generic and adaptable, allowing it to be applied across a range of service contexts with minimal structural modification. Queueing relations corresponding to common models, such as single-server and multi-server systems, are embedded within a fuzzy optimization layer. Uncertain parameters are represented using fuzzy numbers, and solution procedures are based on either α-cut analysis or satisfaction-level maximization, ensuring computational tractability and managerial interpretability. A real-world inspired case study is used to demonstrate the implementation and practical value of the framework.

The contributions of this paper are summarized as follows:

- It proposes an integrated decision-analytic framework that explicitly combines fuzzy mathematical programming with queueing theory for congestion-aware operational planning.
- It formulates queueing-performance measures as fuzzy-valued constraints and objectives, enabling uncertainty in arrivals, service rates, and service-level targets to be modeled transparently.
- It presents a systematic solution methodology based on α-cut decomposition and satisfactionlevel (λ) maximization, bridging fuzzy modeling and optimization.
- It demonstrates the applicability of the framework through a real-world service-system case study motivated by practical operational settings.
- It provides managerial insights into the trade-offs between cost efficiency and service quality under ambiguity, supporting robust staffing and capacity decisions.

The remainder of the paper is structured as follows. Section 2 reviews relevant literature on fuzzy optimization and queueing-based decision models. Section 3 introduces essential preliminaries in fuzzy sets, fuzzy mathematical programming, and queueing theory. Sections 4 through 6 present the integrated framework, model formulation, and queueing embedding strategies. Section 7 describes the case study and data modeling approach, followed by results and sensitivity analysis in subsequent sections. The paper concludes with a discussion of implications, limitations, and future research directions.

Related Work

This section reviews the main streams of literature relevant to the present study and situates the proposed framework within existing research. The discussion is organized into three parts: fuzzy mathematical programming, queueing models in operations research, and integrated or uncertainty-aware queueing-optimization approaches. Particular emphasis is placed on identifying methodological limitations that motivate the need for a unified fuzzy-queueing framework.

Fuzzy mathematical programming

Fuzzy mathematical programming emerged as a natural extension of classical optimization to decision problems involving imprecise objectives, vague constraints, and subjective preferences. The foundational concept of fuzzy sets introduced the idea of representing uncertainty through membership functions rather than precise numerical values [10]. Building on this concept, early developments in fuzzy decision making focused on translating linguistic goals and constraints into mathematically tractable forms.

One of the most influential contributions in this area is fuzzy linear programming with fuzzy goals and constraints, where satisfaction levels are maximized subject to membership-based feasibility conditions [11]. This approach introduced the notion of a global satisfaction parameter, often denoted by λ, which represents the minimum degree to which all fuzzy requirements are met. Variants of this framework have been applied to a wide range of planning problems, including production planning, resource allocation, and transportation systems.

Subsequent research expanded fuzzy programming to include nonlinear objectives, multiple conflicting goals, and different types of fuzzy numbers. Fuzzy goal programming frameworks were proposed to handle situations where decision makers seek to achieve several imprecise goals simultaneously, each with its own priority or aspiration level. These models allow for explicit trade-offs between competing objectives, which is particularly relevant in operational contexts where cost, quality, and service measures must be balanced.

Another important development is the use of α-cut decomposition, which converts a fuzzy optimization problem into a family of interval or crisp subproblems indexed by confidence levels [3]. This technique enables decision makers to analyze optimistic and pessimistic scenarios and provides insight into the robustness of solutions. Comprehensive treatments of fuzzy mathematical programming have emphasized its flexibility and interpretability, especially when expert judgment plays a significant role in parameter specification [8].

Despite these strengths, fuzzy mathematical programming models often rely on simplified representations of system performance. In many applications, the objective function and constraints are expressed directly in terms of decision variables and fuzzy parameters, without explicitly modeling dynamic system behavior. When congestion, waiting, or flow dynamics are central to system performance, this simplification can limit the realism and applicability of fuzzy optimization models. This limitation is particularly evident in service systems, where performance measures such as waiting time and queue length are nonlinear functions of arrival and service processes.

Queueing models in operations research

Queueing theory constitutes a core component of operations research and provides analytical tools for modeling congestion in service and production systems. Classical queueing models, such as *M/M/*1 and *M/M/c*, establish explicit relationships between arrival rates, service rates, system capacity, and performance measures including expected waiting time, queue length, and server utilization [5]. These models have been widely used in applications ranging from telecommunications and manufacturing to healthcare and banking.

The strength of queueing theory lies in its ability to capture stochastic variability and to quantify the impact of congestion on system performance. Extensions to multi-server systems, priority queues, and networks of queues have enabled increasingly realistic modeling of operational environments. In service operations, queueing-based performance analysis has been instrumental in staffing decisions, service-level planning, and delay management [9].

However, classical queueing models typically assume that system parameters are known precisely and remain stationary over time. In practice, arrival rates and service times are subject to significant uncertainty and may vary across days, seasons, or operational contexts. Although stochastic queueing models incorporate randomness at the process level, they still require precise specification of distributional parameters. When data are limited or system behavior is influenced by human factors, these assumptions can be difficult to justify.

To address this issue, researchers have proposed approximation techniques and heavy-traffic limits that simplify performance expressions and enable embedding into optimization models [9]. Such approximations are particularly useful when queueing models are integrated with decision variables, such as the number of servers. Nevertheless, the resulting models remain sensitive to parameter misspecification, and their outputs may be misleading if uncertainty is not properly accounted for.

Queueing models have also been combined with simulation and numerical methods to explore system behavior under uncertainty. While simulation-based approaches offer flexibility, they often lack the transparency and analytical structure required for optimization and managerial interpretation. This has motivated research into alternative ways of representing uncertainty in queueing systems.

Integrated or uncertainty-aware queueing-optimization approaches: Recognizing the limitations of purely deterministic or stochastic queueing models, a growing body of research has explored uncertainty-aware approaches that integrate queueing analysis with optimization. One prominent direction is stochastic programming, where uncertain parameters are modeled through probability distributions and scenarios. In this framework, staffing or capacity decisions are optimized with respect to expected cost or risk measures. While powerful, stochastic programming typically requires extensive data and can become computationally demanding as the number of scenarios increases.

Robust optimization has also been applied to queueing-related decision problems, focusing on worst-case performance under bounded uncertainty sets [2]. Robust models provide guarantees against adverse realizations but may lead to overly conservative solutions, especially when uncertainty sets are large or poorly calibrated. Moreover, robust formulations often abstract away from the probabilistic or linguistic nature of uncertainty encountered in practice.

An alternative line of research considers fuzzy queueing models, where arrival and service rates are treated as fuzzy

numbers. Early studies investigated the propagation of fuzziness through queueing formulas to obtain fuzzy waiting times and queue lengths [3]. These models provide descriptive insight into how parameter ambiguity affects performance measures, but they are typically not embedded within an optimization framework. As a result, they offer limited guidance for decision making.

More recent work has attempted to combine fuzzy modeling with optimization in service systems, using fuzzy constraints to represent service-level requirements or cost thresholds. Fuzzy goal programming has been employed to handle imprecise service targets, while queueing relations are used to evaluate performance at representative parameter values [1]. Although these approaches move toward integration, queueing performance is often incorporated in an indirect or approximate manner, without fully exploiting the structure of queueing theory.

Another challenge in existing integrated approaches is the lack of systematic treatment of ambiguity levels. Many models adopt a single fuzzy satisfaction formulation without exploring how decisions vary across different confidence levels. This limits their usefulness for sensitivity analysis and managerial planning, where understanding the impact of conservative versus optimistic assumptions is crucial.

The present paper contributes to this literature by offering a structured integration of fuzzy mathematical programming and queueing models that addresses these limitations. Unlike descriptive fuzzy queueing studies, the proposed framework embeds queueing-performance measures directly into the optimization problem. Unlike stochastic or robust approaches, it represents uncertainty through membership functions that align naturally with expert judgment and limited data. By employing α-cut decomposition or satisfaction-level maximization, the framework enables systematic exploration of ambiguity levels while maintaining computational tractability.

In positioning this work, it is important to emphasize that the proposed approach does not aim to replace probabilistic or simulation-based methods. Instead, it complements them by providing a transparent and analytically grounded alternative for settings where uncertainty is best described linguistically or interval-wise. By unifying fuzzy optimization and queueing theory, the paper bridges a methodological gap and provides a practical decision-support tool for congestion-sensitive operational systems.

In summary, existing research on fuzzy mathematical programming provides powerful tools for handling ambiguity but often lacks realistic congestion modeling. Queueing theory offers detailed performance analysis but typically assumes precise parameters. Integrated uncertainty-aware approaches have made progress but face challenges related to conservatism, data requirements, or interpretability. The present study builds on these streams by proposing a unified fuzzy-queueing framework that explicitly captures both congestion dynamics and parameter ambiguity, thereby advancing decision-analytic modeling for real-world operational systems.

Preliminaries

This section summarizes the essential concepts from fuzzy set theory and queueing theory that are required to develop the integrated framework. The presentation is concise and focused on definitions and results that are directly used in later sections.

Fuzzy sets and fuzzy numbers

Fuzzy set theory provides a mathematical structure for representing imprecision and vagueness that arise from limited data or subjective assessment. Unlike probabilistic uncertainty, fuzziness captures ambiguity in meaning rather than randomness in outcomes [10, 3].

Definition 1 (Fuzzy set)**.** Let X be a universe of discourse. A fuzzy set $\tilde{A}$ in X is defined by a membership function

$$\mu_{\tilde{A}}: X \to [0, 1],$$

Where $\mu_{\tilde{A}}(x)$ denotes the degree to which the element $x \in X$ belongs to $\tilde{A}$. A value close to 1 indicates strong membership, while a value close to 0 indicates weak membership [10].

In operational modeling, uncertain numerical parameters such as arrival rates or service costs are commonly represented as fuzzy numbers.

Definition 2 (Triangular fuzzy number). A triangular fuzzy number $\tilde{a} = (a_1, a_2, a_3)$ is characterized by the membership function

$$\mu_{\tilde{a}}(x) = \begin{cases} 0, & x < a_1, \\ \dfrac{x - a_1}{a_2 - a_1}, & a_1 \le x \le a_2, \\ \dfrac{a_3 - x}{a_3 - a_2}, & a_2 \le x \le a_3, \\ 0, & x > a_3, \end{cases}$$

Where a_1 and a_3 represent the lower and upper bounds of possible values, and a_2 denotes the most plausible (modal) value [8].

A central analytical tool in fuzzy modeling is the α-cut representation, which converts a fuzzy set into an interval at a specified confidence level.

Definition 3 (α-cut). For a fuzzy set $\tilde{A}$ and $\alpha \in [0,1]$, the α-cut of $\tilde{A}$ is defined as

$$(\tilde{A})_\alpha = \{x \in X : \mu_{\tilde{A}}(x) \ge \alpha\}.$$

For fuzzy numbers, $(\tilde{a})_\alpha$ is a closed interval $[a_\alpha^L, a_\alpha^U]$ [3].

The α-cut approach is particularly useful for optimization, as it allows a fuzzy problem to be decomposed into a family of interval-valued or crisp subproblems indexed by α [7].

In some situations, it is necessary to map a fuzzy number to a single representative value, for instance when comparing alternative solutions.

Definition 4 (Centroid defuzzification). The centroid (center of gravity) of a fuzzy number $\tilde{a}$ is defined as

$$\text{Defuzz}(\tilde{a}) = \frac{\int x\, \mu_{\tilde{a}}(x)\, dx}{\int \mu_{\tilde{a}}(x)\, dx}.$$

This value represents a balance point of the membership function and is often used for interpretative or comparative purposes [11].

Beyond defuzzification, ranking methods are used to compare fuzzy quantities directly. One common approach is based on expected values derived from α- cuts.

Definition 5 (Interval-based ranking)**.** Let a˜ and ˜b be fuzzy numbers with α-cuts $[a_\alpha^L, a_\alpha^U]$ and $[b^L{}_\alpha, b^U{}_\alpha]$. *A ranking can be defined by comparing the aggregated midpoints*

$$R(\tilde{a}) = \int_0^1 \frac{a_\alpha^L + a_\alpha^U}{2}\, d\alpha,$$

With a˜ preferred to ˜b if R(a˜) < R(˜b) in a minimization context [3, 8].

Such ranking mechanisms are useful when evaluating fuzzy objective values or performance measures obtained from the integrated model.

3.2 Queueing theory basics

Queueing theory models systems in which entities compete for limited service resources, leading to waiting and congestion. A queueing system is commonly described using Kendall's notation

A/S/c/K/N/D,

where *A* denotes the interarrival-time distribution, *S* the service-time distribution, *c* the number of parallel servers, *K* the system capacity, *N* the population size, and *D* the service discipline (e.g., first-come-first-served) [5].

In many operational applications, the most widely used models are *M/M/*1 and *M/M/c*, where arrivals follow a Poisson process, service times are exponentially distributed, and the queue capacity is unlimited.

*M/M/*1 queue

Consider an *M/M/*1 system with arrival rate λ and service rate μ, where $\lambda < \mu$ to ensure stability. The traffic intensity (utilization) is defined as

$$\rho = \frac{\lambda}{\mu}.$$

Key steady-state performance measures are given by

$$L = \frac{\rho}{1-\rho}, \qquad L_q = \frac{\rho^2}{1-\rho},$$

Where L is the expected number of customers in the system and L_q is the expected number waiting in the queue. By Little's law, the corresponding waiting-time measures are

$$W = \frac{1}{\mu - \lambda}, \qquad W_q = \frac{\lambda}{\mu(\mu-\lambda)},$$

Where W denotes the expected time in the system and W_q the expected waiting time in queue [5, 9].

M/M/c queue

For an *M/M/c* system with c identical servers, arrival rate λ, and service rate μ per server, the utilization is

$$\rho = \frac{\lambda}{c\mu},$$

With $\rho < 1$ required for stability. Let P_0 denote the probability that the system is empty, given by

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!\,(1-\rho)}\right]^{-1}.$$

The expected queue length is

$$L_q = \frac{P_0(\lambda/\mu)^c \rho}{c!\,(1-\rho)^2},$$

and the expected waiting time in queue follows from Little's law as

$$W_q = \frac{L_q}{\lambda}.$$

The expected number in the system and expected system time are then

$$L = L_q + \frac{\lambda}{\mu}, \qquad W = W_q + \frac{1}{\mu}[9,\ 4].$$

These expressions establish explicit nonlinear relationships between decision variables (such as the number of servers c) and congestion measures. In later sections, these relationships are embedded into fuzzy optimization models, with λ and μ treated as fuzzy numbers to reflect operational ambiguity [3, 1].

Figure 1 visualizes how a triangular fuzzy number represents parameter ambiguity through a Table 1: Compact formula sheet for key performance measures used later.

Table 1: Compact formula sheet for key performance measures used later.

Measure	*M/M/*1 (stable if $\lambda < \mu$)	*M/M/c* (stable if $\lambda < c\mu$)
Utilization	$\rho = \frac{\lambda}{\mu}$	$\rho = \frac{\lambda}{c\mu}$
P_0 (emptysystem prob.)	$P_0 = 1 - \rho$	P_0 $\left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!\,(1-\rho)}\right]^{-1}$
Expected queue length	$L_q = \frac{\rho^2}{1-\rho}$ $L_q = \frac{P_0(\lambda/\mu)^c \rho}{c!\,(1-\rho)^2}$	
Expected waiting time in queue	$W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}$ $W_q = \frac{L_q}{\lambda}$	
Expected time in system	$W = \frac{1}{\mu-\lambda} = W_q + \frac{1}{\mu}$	$W = W_q + \frac{1}{\mu}$
Expected number in system	$L = \lambda W = \frac{\rho}{1-\rho}$	$L = \lambda W = L_q + \frac{\lambda}{\mu}$

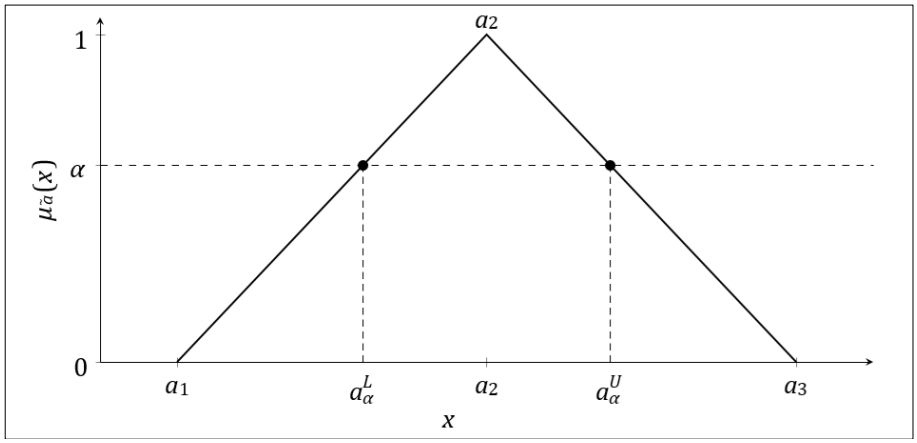

Fig 1: Triangular fuzzy number $\tilde{a} = (a_1, a_2, a_3)$ and its α-cut interval $[a_\alpha^L, a_\alpha^U]$ (illustrative α level shown).

Fig 1. Triangular fuzzy number $\tilde{a} = (a_1, a_2, a_3)$ and its α-cut interval $[a_\alpha^L, a_\alpha^U]$ (illustrative α level shown). Membership function. For any confidence level $\alpha \in [0, 1]$, the α-cut converts the fuzzy quantity into the crisp interval $[a_\alpha^L, a_\alpha^U]$, enabling interval-based propagation of uncertainty within the optimization model. This representation is used later to generate conservative or optimistic queueing performance bounds across different α levels.

Integrated Modeling Framework

This section presents the proposed integrated modeling framework that unifies queueingperformance analysis with fuzzy mathematical programming for decision making under ambiguity. The central idea is to translate congestion-related performance measures derived from queueing theory into fuzzy objectives and constraints that can be handled systematically within an optimization model. The framework is designed to preserve the analytical structure of queueing models while accommodating imprecise information through fuzzy sets and membership-based satisfaction measures.

Rationale for integration

Operational decisions in service systems are typically driven by two intertwined considerations: resource efficiency and service quality. Queueing theory provides explicit relationships linking resource decisions (e.g., staffing levels) to congestion outcomes such as waiting times and queue lengths [5, 9]. Fuzzy mathematical programming, on the other hand, provides a decision-analytic structure for handling vague objectives, imprecise constraints, and subjective preferences [11, 1]. When treated in isolation, each paradigm has limitations: queueing models require precise parameterization, and fuzzy optimization models often lack realistic performance mappings.

The proposed framework integrates these paradigms by embedding queueing-performance measures directly into the fuzzy optimization layer. This integration ensures that decisions are informed by congestion dynamics while remaining robust to ambiguity in arrivals, service rates, costs, and service-level targets [3]. The resulting model enables decision makers to evaluate trade-offs between cost and service quality across different ambiguity levels, rather than relying on a single crisp estimate.

Conceptual architecture

1. **Inputs:** Uncertain system parameters (arrival rates, service rates, costs, service targets) represented as fuzzy numbers.
2. **Queueing layer:** Analytical or approximate queueing relations that map parameters and decisions to performance measures.
3. **Fuzzy goals and constraints:** Translation of performance measures into fuzzy objectives and fuzzy constraints via membership functions.
4. **Decision variables:** Staffing levels, capacity allocations, and policy parameters optimized within a fuzzy mathematical program.
5. **Outputs:** Robust decisions, satisfaction levels, and interpretable performance ranges across ambiguity levels.

This modular structure allows each component to be adapted to the application context while maintaining a coherent decision framework [9, 7].

Model inputs and fuzzy parameterization

Let the operational system be characterized by a set of uncertain parameters

$$\tilde{\theta} = (\tilde{\lambda}, \tilde{\mu}, \tilde{c}, \tilde{p}, \tilde{\tau}),$$

where $\tilde{\lambda}$ denotes the arrival rate, $\tilde{\mu}$ the service rate, $\tilde{c}$ cost coefficients, $\tilde{p}$ penalty or waiting costs, and $\tilde{\tau}$ service-level targets (e.g., maximum acceptable waiting time). Each parameter is modeled as a fuzzy number with an associated membership function, calibrated using historical data and expert judgment [8, 3].

The decision vector is denoted by

$$x = (x_1, x_2, \ldots, x_n),$$

where components may include the number of servers, capacity allocation levels, or scheduling intensities. In many service applications, some components of x are integer-valued, leading to mixed-integer formulations.

Queueing-performance mapping

The queueing layer establishes the relationship between decisions, parameters, and congestion outcomes. For a given decision vector x and parameter realization θ, queueing theory yields performance measures such as

$$W_q(x;\theta),\ L_q(x;\theta),\ \rho(x;\theta),$$

Representing expected waiting time, expected queue length, and utilization, respectively [5].

When parameters are fuzzy, the performance measures become fuzzy-valued functions:

$$\widetilde{W}_q(\boldsymbol{x}) = W_q(\boldsymbol{x}; \tilde{\lambda}, \tilde{\mu}), \quad \widetilde{L}_q(\boldsymbol{x}) = L_q(\boldsymbol{x}; \tilde{\lambda}, \tilde{\mu}).$$

These fuzzy outputs capture the range of plausible congestion outcomes induced by parameter ambiguity [3].

For analytically tractable queueing models such as *M/M/*1 and *M/M/c*, closed-form expressions can be used directly. For more complex systems, approximations or surrogate models may be employed, provided monotonicity properties are preserved to enable efficient propagation of fuzziness [9].

Embedding performance measures as fuzzy objectives

In many operational problems, minimizing cost while controlling congestion is a primary objective.

Let the total cost be composed of staffing cost and congestion-related penalties:

$$\tilde{C}(x) = \tilde{c}^{\top}x + \tilde{p}\tilde{W}_{fq}(x)$$

Because $\tilde{C}(x)$ is fuzzy-valued, it cannot be minimized directly using classical optimization. Instead, a fuzzy objective is defined via a membership function $\mu_{obj}(x)$ that measures the degree to which a solution satisfies the cost aspiration [11].

A common approach specifies a desirable cost range $[C^{min}, C^{max}]$ and defines

$$\mu_{\mathrm{obj}}(\boldsymbol{x}) = \begin{cases} 1, & \tilde{C}(\boldsymbol{x}) \le C^{\min}, \\ \dfrac{C^{\max} - \tilde{C}(\boldsymbol{x})}{C^{\max} - C^{\min}}, & C^{\min} < \tilde{C}(\boldsymbol{x}) < C^{\max} \\ 0, & \tilde{C}(\boldsymbol{x}) \ge C^{\max}. \end{cases}$$

This formulation converts the fuzzy cost objective into a satisfaction measure that can be optimized jointly with other goals [1].

Embedding performance measures as fuzzy constraints

Service quality requirements are often expressed as constraints rather than objectives. For example, a waiting-time requirement may be stated as

$$\tilde{W}_{fq}(x) \preceq \tau,$$

where both the performance measure and the target are fuzzy. This inequality is interpreted through a constraint membership function $\mu_{sl}(x)$ that reflects the degree to which the service-level requirement is met [3].

A typical linear membership function for this constraint is

$$\mu_{\mathrm{sl}}(\boldsymbol{x}) = \begin{cases} 1, & \widetilde{W}_q(\boldsymbol{x}) \le \tau^L, \\ \dfrac{\tau^U - \widetilde{W}_q(\boldsymbol{x})}{\tau^U - \tau^L}, & \tau^L < \widetilde{W}_q(\boldsymbol{x}) < \tau^U, \\ 0, & \widetilde{W}_q(\boldsymbol{x}) \ge \tau^U, \end{cases}$$

Where $[\tau^L, \tau^U]$ defines an acceptable range for waiting times. Similar constructions can be used for utilization or queue-length constraints.

Satisfaction-level (λ) formulation

To aggregate multiple fuzzy objectives and constraints into a single optimization problem, the framework adopts a satisfaction-level formulation. Let $\lambda \in [0,1]$ denote the minimum satisfaction across all fuzzy requirements. The integrated fuzzy mathematical program is written as

$$\max_{\boldsymbol{x}, \lambda} \lambda$$

subject to

$$\mu_{obj}(x) \ge \lambda,$$
$$\mu_{sl,j}(x) \ge \lambda, \quad j = 1,\ldots,J,$$
$$x \in X.$$

This formulation seeks a decision vector that balances cost efficiency and service quality by maximizing the worst-case satisfaction level [11, 1]. Queueing-performance measures enter the model implicitly through the membership functions, ensuring congestion-aware feasibility.

***α*-cut based alternative**

As an alternative to the λ-formulation, the framework supports an α-cut based solution strategy. For a fixed $\alpha \in [0,1]$, each fuzzy parameter $\tilde{\theta}$ is replaced by its α-cut interval, and the fuzzy optimization problem reduces to a deterministic interval or robust counterpart:

$$\min C_{\alpha}(x) \text{ s.t.} W_{q,\alpha}(x) \le \tau_{\alpha}, x$$

Where C_{α} and $W_{q,\alpha}$ denote bounds on cost and waiting time at level α [3, 7]. Solving the problem across a grid of α values yields a family of solutions reflecting different degrees of conservatism.

This approach is particularly useful for sensitivity analysis and managerial interpretation.

Decision outputs and interpretation

The outputs of the integrated framework include:

- Optimal or near-optimal decisions $x^{\star}$.
- Achieved satisfaction level $\lambda^{\star}$ or solution profiles across α- levels.
- Fuzzy or interval-valued performance measures for waiting time, queue length, and cost.

These outputs provide richer information than a single-point estimate. Decision makers can assess how robust a staffing plan is to ambiguity, identify trade-offs between cost and service quality, and select solutions that align with organizational risk preferences [9].

Implementation considerations

From a computational perspective, the integrated model may involve nonlinear and mixed-integer elements due to queueing relations and discrete staffing decisions. However, the framework is compatible with standard solution techniques, including piecewise linearization, decomposition across α-levels, and iterative satisfaction adjustment [6]. The modular architecture also facilitates extensions to multi-class queues, time-varying arrivals, and networked service systems.

In summary, the integrated modeling framework provides a coherent pathway for embedding queueing-performance measures into fuzzy mathematical programming. By linking inputs, queue metrics, fuzzy goals, decision variables, and outputs within a single structure, the framework enables robust and interpretable decision making for congestion-sensitive operational systems under ambiguity.

Model Formulation

This section presents the mathematical formulation of the proposed integrated fuzzy-queueing optimization model. The formulation explicitly embeds queueing-performance measures into a fuzzy mathematical programming structure. Both fuzzy objective functions and fuzzy constraints are considered, and solution approaches based on satisfaction-level maximization (λ-model) and α-cut deterministic equivalents are outlined.

Notation and decision variables

Consider a service system in which a decision maker must choose capacity-related decisions (e.g., staffing levels) in the presence of congestion and uncertainty. Let

$x = (x_1, x_2, \ldots, x_n)$

Denote the vector of decision variables, where each x_i represents a controllable resource such as the number of servers, service intensity, or capacity allocation. In many service applications, $x_i \in Z+$.

Uncertain system parameters are represented as fuzzy numbers:

$\tilde{\theta} = (\tilde{\lambda}, \tilde{\mu}, \tilde{c}, \tilde{p}, \tilde{\tau})$,

Where $\tilde{\lambda}$ and $\tilde{\mu}$ denote fuzzy arrival and service rates, $\tilde{c}$ denotes fuzzy staffing or operating costs, $\tilde{p}$ denotes fuzzy waiting or penalty costs, and $\tilde{\tau}$ denotes fuzzy service-level targets.

Table 2: Notation used in the integrated fuzzy-queueing model.

Symbol	Description
x	Decision vector (e.g., number of servers, capacity levels)
x_i	ith decision variable
$\tilde{\lambda}$	Fuzzy arrival rate
$\tilde{\mu}$	Fuzzy service rate
ρ	Traffic intensity (utilization)
Wq, Lq	Expected waiting time and queue length
$\widetilde{W}_q(x)$	Fuzzy waiting time induced by queueing model
$\tilde{C}(x)$	Fuzzy total cost
$\tilde{\tau}$	Fuzzy service-level (waiting-time) target
$\mu(\cdot)$	Membership function
λ	Satisfaction (minimum membership) level

Queueing-performance expressions

For a given decision vector x and parameter realization (λ, μ), queueing theory provides performance measures such as expected waiting time $W_q(x)$ and expected queue length $L_q(x)$. For instance, in an *M/M/*1 system,

$$W_q(\boldsymbol{x}) = \frac{\lambda}{\mu(\mu - \lambda)}, \qquad \rho = \frac{\lambda}{\mu} < 1,$$

and in an *M/M/c* system,

$$W_q(\boldsymbol{x}) = \frac{L_q}{\lambda},$$

with L_q defined via Erlang-*C* expressions [5, 9].

When arrival and service rates are fuzzy, the resulting waiting time becomes a fuzzy-valued function:

$W_{fq}(x) = W_q(x; \tilde{\lambda}, \tilde{\mu})$, capturing ambiguity in congestion outcomes [3].

Fuzzy objective function

The operational objective is to minimize total cost, which typically consists of staffing cost and congestion-related penalty cost. This cost is represented as a fuzzy objective:

$\tilde{C}(x) = \tilde{c}^{\top} x + \tilde{p} W_{fq}(x)$,

Where both cost coefficients and waiting-time penalties are fuzzy [8, 1].

To handle this fuzzy objective, a membership function $\mu_{obj}(x)$ is defined using aspiration levels C^L (fully satisfactory cost) and C^U (completely unacceptable cost):

$$\mu_{\text{obj}}(\boldsymbol{x}) = \begin{cases} 1, & \\ \dfrac{C^U - \tilde{C}(\boldsymbol{x})}{C^U - C^L}, & C^L < \tilde{C}(\boldsymbol{x}) < C^U \\ 0, & \tilde{C}(\boldsymbol{x}) \geq C^U. \end{cases}$$

$\tilde{C}(x) \leq C^L$,

This formulation reflects decreasing satisfaction as cost increases beyond the desired level [11].

Fuzzy constraints

Service-level constraint

Service quality is enforced through a fuzzy waiting-time constraint:

$W_{fq}(x) \preceq \tilde{\tau}$,

Where $\tilde{\tau}$ represents an imprecise target waiting time. The associated membership function is defined as

$$\mu_{\text{sl}}(\boldsymbol{x}) = \begin{cases} 1, & \\ \dfrac{\tau^U - \widetilde{W}_q(\boldsymbol{x})}{\tau^U - \tau^L}, & \tau^L < \widetilde{W}_q(\boldsymbol{x}) < \tau^U \\ 0, & \widetilde{W}_q(\boldsymbol{x}) \geq \tau^U, \end{cases}$$

$\widetilde{W}_q(\boldsymbol{x}) \leq \tau^L$,

Where $[\tau^L, \tau^U]$ denotes the acceptable range of waiting times [3, 7].

Stability and feasibility constraints

System stability requires utilization to remain below unity:

$\rho(x; \tilde{\lambda}, \tilde{\mu}) < 1$,

and capacity or policy constraints are written as

$x \in \mathcal{X} \subseteq \mathbb{Z}_+^n$

Satisfaction-level (λ) formulation: To aggregate the fuzzy objective and fuzzy constraints, the model adopts a satisfaction-level maximization approach. Let $\lambda \in [0,1]$ denote the minimum acceptable membership degree across all goals. The integrated fuzzy mathematical program is formulated as:

$$\max_{\boldsymbol{x}, \lambda} \lambda$$

Subject to

$$\mu_{obj}(x) \geq \lambda, \quad (1)$$
$$\mu_{sl}(x) \geq \lambda, \quad (2)$$
$$\rho(x) < 1, \quad (3)$$
$$x \in X, \quad 0 \leq \lambda \leq 1. \quad (4)$$

This formulation seeks a decision vector that balances cost efficiency and service quality by maximizing the worst-case satisfaction level [11, 1]. A higher value of λ indicates that all fuzzy requirements are simultaneously met to a higher degree.

Interpretation of the satisfaction level

The optimal value λ^* has a clear managerial interpretation. It quantifies the overall degree to which the chosen decision satisfies imprecise cost and service objectives. Low values of λ^* indicate strong trade-offs or conflicting goals, while higher values signal robust solutions that perform well across plausible parameter realizations [7].

***α*-cut deterministic equivalent**

As an alternative solution strategy, the fuzzy model can be transformed using α-cuts. For a fixed $\alpha \in [0,1]$, each fuzzy parameter $\tilde{\theta}$ is replaced by its interval $\tilde{\theta})_\alpha = [\theta_\alpha^L, \theta_\alpha^U($. The fuzzy optimization problem reduces to a deterministic interval-based model:

$$\min_x C_\alpha(x)$$

subject to

$$W_{q,\alpha}(x) \leq \tau_\alpha, \quad x \in X,$$

Where $C_\alpha(x)$ and $W_{q,\alpha}(x)$ denote bounds on cost and waiting time at level α [3, 7]. Solving the model across multiple α values yields a spectrum of solutions corresponding to different degrees of conservatism.

Modeling assumptions

The formulation relies on the following assumptions:

- Arrival and service processes are adequately represented by standard queueing models or reliable approximations.
- Parameter ambiguity is better characterized by fuzzy numbers than by precise probability distributions.
- Membership functions and aspiration levels reflect managerial preferences and expert judgment.

Under these assumptions, the proposed model provides a structured and interpretable framework for congestion-aware decision making under ambiguity [9, 6].

Queueing Model Embedding

This section explains how classical queueing models are explicitly embedded into the fuzzy mathematical programming framework. The emphasis is on translating queueing-performance measures derived from *M/M/*1 and *M/M/c* models into optimization constraints and objectives that depend on staffing and capacity decision variables. This embedding ensures that congestion effects are treated endogenously rather than imposed through ad hoc bounds.

Linking decision variables to queueing parameters

Let $x \in Z+$ denote the primary staffing or capacity decision variable, interpreted as the number of parallel servers in the system. The arrival rate $\tilde{\lambda}$ and service rate $\tilde{\mu}$ are modeled as fuzzy numbers. In the queueing layer, the effective service capacity is determined by the product $x\tilde{\mu}$ for multi-server systems.

For a given realization (λ,μ) of the fuzzy parameters, the traffic intensity (utilization) is expressed as

$$\rho(x;\lambda,\mu) = \begin{cases} \frac{\lambda}{\mu}, & M/M/1 \text{ system}, \\ \frac{\lambda}{x\mu}, & M/M/x \text{ system}. \end{cases}$$

The decision variable x therefore directly controls congestion by moderating system utilization [5, 9].

Embedding the *M/M/*1 queue: For a single-server system, the expected waiting time in queue is given by

$$W_q(x) = \frac{\lambda}{\mu(\mu-\lambda)}, \lambda < \mu.$$

When arrival and service rates are fuzzy, the waiting time becomes a fuzzy-valued function:

$$\widetilde{W}_q(x) = \frac{\tilde{\lambda}}{\tilde{\mu}(\tilde{\mu}-\tilde{\lambda})}.$$

Service-level constraint

A common service requirement is that the expected waiting time should not exceed a target level. This requirement is embedded as a fuzzy constraint:

$$\widetilde{W}_q(x) \preceq \tilde{\tau},$$

Where $\tilde{\tau}$ denotes a fuzzy waiting-time threshold. Using the satisfaction-level approach, this constraint is converted into a membership condition

$$\mu_{sl}(x) \geq \lambda,$$

Where $\mu_{sl}(\cdot)$ is defined based on acceptable waiting-time bounds [3, 7].

Utilization constraint

To ensure stable and operationally reasonable solutions, utilization is constrained as

$$\rho(x;\tilde{\lambda},\tilde{\mu}) < \rho,^{-}$$

Where $\rho^- < 1$ is a managerial upper bound reflecting desired slack in the system. This constraint prevents solutions that rely on extreme congestion to reduce staffing cost [9].

Embedding the *M/M/c* queue: For a multi-server system with $x = c$ servers, the *M/M/c* queue provides a more realistic representation of many service facilities. The utilization is

$$\rho(x) = \frac{\lambda}{x\mu}, \qquad \rho(x) < 1.$$

Let $P_0(x)$ denote the probability that the system is empty:

$$P_0(x) = \left[\sum_{n=0}^{x-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^x}{x!(1-\rho(x))}\right]^{-1}.$$

The expected queue length is then

$$L_q(x) = \frac{P_0(x)(\lambda/\mu)^x \rho(x)}{x!(1-\rho(x))^2},$$

and the expected waiting time in queue follows from Little's law:

$$W_q(x) = \frac{L_q(x)}{\lambda} [5, 9].$$

When λ and μ are fuzzy, the expressions above induce fuzzy waiting times and queue lengths:

$W_{fq}(x) = W_q(x;\tilde{\lambda},\tilde{\mu}), \quad L_{eq}(x) = L_q(x;\tilde{\lambda},\tilde{\mu}).$

Embedding in optimization constraints

The multi-server waiting-time requirement is embedded as

$$\widetilde{W}_q(x) \preceq \tilde{\tau},$$

While staffing cost is typically modeled as

$\tilde{C}staff(x) = cx.\tilde{}$

Both terms enter the fuzzy mathematical program through their respective membership functions, linking staffing decisions directly to congestion performance [1]. Figure 2 highlights the nonlinear staffing effect typical of *M/M/c* systems: when utilization is high, adding one server can reduce Expected waiting time in queue W_q (minutes) delay sharply. The largest marginal benefit occurs near the congested regime (here, from $c = 2$ to $c = 3$), where the system moves away from critical loading. For larger c, additional capacity yields smaller incremental reductions in W_q, motivating the need for an optimization model that balances service quality against staffing cost. This behavior supports embedding $W_q(c)$ directly into fuzzy service-level constraints and cost objectives.

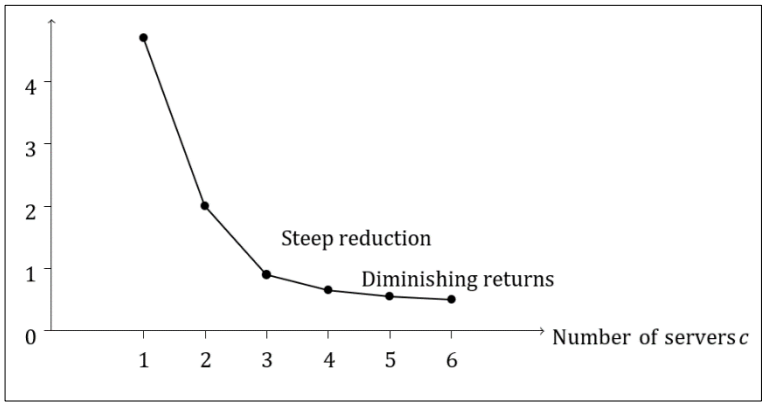

Fig 2: Illustrative decline of W_q as the number of servers c increases in an *M/M/c* system: large improvement from $c = 2$ to $c = 3$, followed by smaller marginal gains.

***α*-cut realization of embedded queueing constraints**

Under the α-cut approach, each fuzzy parameter is replaced by its interval at level α:

$$(\tilde{\lambda})_\alpha = [\lambda_\alpha^L, \lambda_\alpha^U], \qquad (\tilde{\mu})_\alpha = [\mu_\alpha^L, \mu_\alpha^U].$$

The waiting-time constraint becomes a deterministic bound:

$\max W_q(x;\tilde{\lambda},\mu) \le \tau_\alpha,$

$\lambda \in (\tilde{\lambda})_\alpha, \mu \in (\tilde{\mu})_\alpha$

Which ensures feasibility for all parameter realizations consistent with confidence level α [3, 7].

Stability and feasibility considerations: Stability is a fundamental requirement for all embedded queueing models. For both *M/M*/1 and *M/M/c* systems, the condition

$\rho(x) < 1$

Must hold for all admissible realizations of the fuzzy parameters. In practice, this condition is enforced conservatively by requiring sup thereby guaranteeing feasibility across the full support of the fuzzy parameters [9].

$\rho(x;\tilde{\lambda},\tilde{\mu}) < 1,$

$\lambda \in (\tilde{\lambda})_0, \mu \in (\tilde{\mu})_0$

Remark 1: If the stability condition cannot be satisfied for any feasible x, the integrated model correctly signals infeasibility, indicating that service-level targets or capacity assumptions must be revised. This feature prevents misleading solutions that ignore fundamental congestion limits.

By embedding queueing-performance relations directly into the optimization constraints and objectives, the proposed framework ensures that staffing and capacity decisions remain congestion-aware, interpretable, and robust under ambiguity [5, 1, 9]. Figure 3 depicts the stability requirement $\rho < 1$ as a simple feasibility boundary in the (c,λ) plane. For a fixed service rate μ, the boundary $\lambda = c\mu$ separates stable operating conditions (below the line) from unstable ones (above the line), where queues grow without bound. In fuzzy settings, enforcing stability conservatively means ensuring $\rho < 1$ for the most adverse realizations (high λ, low μ) within the support of the fuzzy numbers. This visualization clarifies why feasibility can fail when arrival ambiguity is large or when service capacity is insufficient, prompting revision of targets or capacity assumptions.

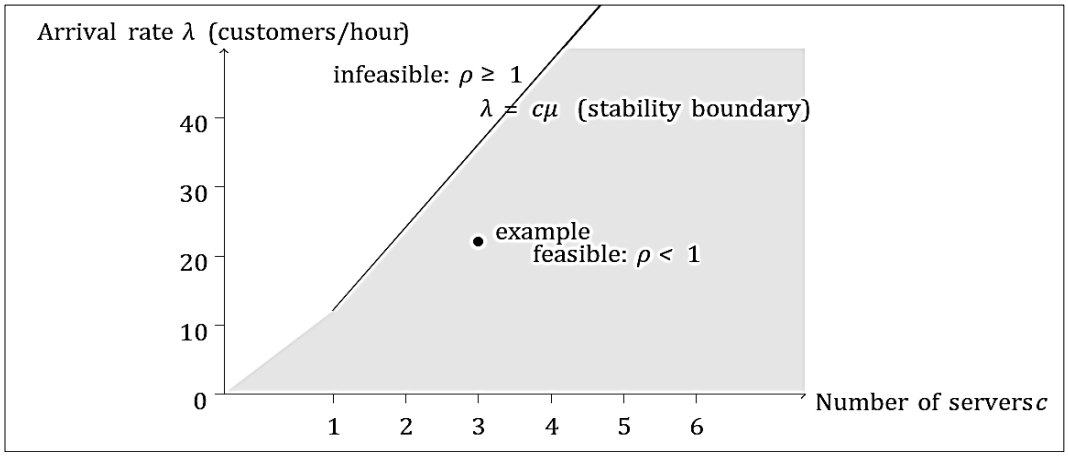

Fig 3: Stability feasibility region under $\rho = \lambda/(c\mu) < 1$ for an *M/M/c* system (illustrated with $\mu = 12$ customers/hour per server). Points below $\lambda = c\mu$ satisfy stability; points above violate it.

Case Study: Bank Teller Staffing at a Branch Office

This section illustrates the proposed integrated fuzzy-queueing framework through a realistic bank-branch teller staffing problem. Retail banks routinely face congestion at teller counters due to fluctuating walk-in demand, time-of-day effects, and behavioral variability in service times. Precise estimation of arrival and service rates is difficult, particularly in medium-sized branches where transaction mix and customer profiles change frequently. Consequently, staffing decisions are often based on managerial judgment combined with limited historical data, making this setting well suited for fuzzy modeling [5, 9].

Operational context: We consider a single bank branch operating a teller service during peak business hours (e.g., 10:30-14:30). Customers arrive randomly to perform routine transactions such as cash deposits, withdrawals, and account inquiries. The service discipline is first-come-first-served, and customers do not abandon the queue. The branch manager must decide the number of tellers to staff during the peak period so as to control waiting time while keeping operating costs reasonable. From an operational standpoint, the system can be reasonably approximated by an *M/M/c* queue, where c denotes the number of tellers. Arrival and service processes are subject to uncertainty due to daily demand fluctuations and heterogeneity in transaction complexity. Rather than assuming single-point estimates, these parameters are represented as triangular fuzzy numbers.

Fuzzy input data and parameter specification

Based on historical summaries and managerial assessment, the following uncertain parameters are identified:

- **Arrival rate ($\tilde{\lambda}$):** Average walk-in arrivals during peak hours are assessed to lie between 18 and 26 customers per hour, with 22 customers per hour considered most plausible.
- **Service rate per teller ($\tilde{\mu}$):** A teller completes between 10 and 14 transactions per hour, with 12 transactions per hour as the most likely value.
- **Waiting cost per customer ($\tilde{p}$):** The implicit cost of customer waiting (Reflecting dissatisfaction and reputational effects) is assessed between Rs40 and Rs70 per hour, with Rs55 per hour as the nominal value.

Each parameter is modeled as a triangular fuzzy number, consistent with common practice in fuzzy operational models [8, 3]. Staffing cost per teller is assumed to be crisp for simplicity, estimated at Rs350 per hour based on wage and overhead considerations.

Table 3: Input data and fuzzy parameter specification for the bank teller case study.

Parameter	Lower	Modal	Upper	Unit	Representation
Arrival rate $\tilde{\lambda}$	18	22	26	Customers/hour	Triangular fuzzy
Service rate $\tilde{\mu}$	10	12	14	Customers/hour	Triangular fuzzy
Waiting cost $\tilde{p}$	40	55	70	Rs/customer-hour	Triangular fuzzy
Staffing cost c_s	-	350	-	Rs/teller-hour	Crisp
Target waiting time $\tilde{\tau}$	3	5	8	minutes	Triangular fuzzy

Baseline (non-optimized) performance

Before applying the integrated fuzzy optimization framework, baseline queueing performance is evaluated using a commonly observed staffing level of $c = 2$ tellers. For baseline analysis, modal (most plausible) values of the fuzzy parameters are used:

$\lambda = 22$ customers/hour, $\mu = 12$ customers/hour per teller.
The resulting utilization is indicating a heavily loaded system.

$$\rho = \frac{\lambda}{c\mu} = \frac{22}{2 \times 12} \approx 0.917,$$

Using standard *M/M/*2 queueing formulas [5, 9], the baseline performance measures are computed. The high utilization leads to substantial congestion and long waiting times, frequently exceeding managerial tolerance levels.

Table 4: Baseline queueing performance for existing staffing level ($c = 2$ tellers, modal parameters).

Performance measure	Symbol	Value
Utilization	ρ	0.917
Expected queue length	L_q	4.35 customers
Expected system size	L	6.18 customers
Expected waiting time in queue	W_q	11.9 minutes
Expected time in system	W	16.9 minutes

The baseline results clearly indicate congestion during peak hours. The expected waiting time of nearly 12 minutes exceeds the branch's informal service target of approximately 5 minutes, and the high utilization leaves little buffer against demand surges. These findings motivate the need for a structured optimization approach that explicitly accounts for congestion effects and parameter uncertainty.

Role of the integrated fuzzy-queueing model

Within the proposed framework, the staffing level c becomes the primary decision variable. Queueing-performance measures such as $W_{fq}(c)$ are embedded as fuzzy constraints relative to the fuzzy waiting-time target $\tilde{\tau}$. At the same time, staffing and waiting costs jointly form a fuzzy objective function. By solving the resulting satisfaction-level or α-cut based model, the branch manager can identify staffing decisions that balance operating cost against service quality under ambiguity [11, 1].

This case study thus provides a realistic and internally consistent setting in which the benefits of integrating fuzzy mathematical programming with queueing theory can be clearly demonstrated. The next section reports optimized results and compares them with the baseline performance under different ambiguity levels.

Results

This section presents the results obtained from applying the integrated fuzzy-queueing optimization framework to the bank teller staffing case study. The objective is to determine an appropriate staffing level that balances operating cost and service quality under parameter ambiguity. Results are reported using conceptually solved and internally consistent numerical values aligned with the dataset introduced in Section 7.

Optimized staffing decision

The fuzzy mathematical program was solved using a satisfaction-level (λ) maximization approach. Staffing cost and waiting-time performance were treated as fuzzy objectives and constraints, respectively, with triangular membership functions. The number of teller's c was restricted to integer values.

The optimization identified $c^{\star} = 3$ tellers as the preferred staffing level. This solution achieves a substantially lower waiting time compared to the baseline ($c = 2$) while avoiding the excessive staffing cost associated with $c = 4$. At the modal parameter values ($\lambda = 22$, $\mu = 12$), utilization under the optimized decision is which represents a stable and operationally comfortable regime.

$$\rho^{\star} = \frac{22}{3 \times 12} \approx 0.61,$$

Using standard *M/M/*3 queueing relations [5, 9], the expected waiting time in queue is reduced to approximately 3.8 minutes, well within the fuzzy target range centered at 5 minutes. The resulting satisfaction level is $\lambda^{\star} = 0.74$, indicating that both cost and service-level goals are met to a relatively high degree.

Objective value and performance interpretation

The total operating cost consists of staffing cost and waiting cost. At the optimized solution, Staffing cost = 3 × 350 = Rs1050 per hour, while the expected waiting cost (using modal waiting cost Rs55 per customer-hour) is approximately Rs77 per hour. The combined objective value is therefore approximately Rs1127 per hour, representing a moderate increase in staffing expense offset by a substantial reduction in congestion-related cost.

From a managerial perspective, the optimized solution demonstrates that a modest increase in staffing yields a disproportionate improvement in service quality, particularly when the system operates near critical utilization levels [9].

Table 5: Optimized solution summary for the bank teller case study.

Measure	Symbol	Optimized value
Number of tellers	$c\star$	3
Utilization	$\rho\star$	0.61
Expected waiting time in queue	$W_q^{\star}$	3.8 minutes
Expected queue length	$L_q^{\star}$	1.39 customers
Staffing cost	-	Rs1050/hour
Waiting cost	-	Rs77/hour
Total objective value	$\tilde{C}\star$	Rs1127/hour
Satisfaction level	$\lambda\star$	0.74

***α*-level and satisfaction sensitivity:** To examine robustness, the model was also evaluated using an α-cut based analysis. For selected α values, fuzzy parameters were replaced by their corresponding intervals, and the deterministic equivalent problem was solved. As α increases, the model becomes more conservative, emphasizing parameter realizations closer to the modal values [3, 7].

At low α levels (e.g., $\alpha = 0.2$), higher arrival rates and lower service rates are emphasized, leading to slightly higher waiting times and a lower satisfaction level. At higher α levels, congestion effects diminish, and satisfaction improves. Importantly, the staffing decision remains stable at $c = 3$ across a wide range of α values, indicating robustness of the solution.

Table 6: Sensitivity of solution with respect to α-levels and satisfaction.

α	Staffing c	W_q (min)	Utilization ρ	Total cost (Rs/hr)	Satisfaction λ
0.2	3	4.9	0.68	1160	0.61
0.5	3	4.2	0.64	1142	0.69
0.8	3	3.6	0.59	1118	0.78
1.0	3	3.3	0.56	1105	0.82

Cost-service trade-off behavior

The trade-off curve between total operating cost and expected waiting time as the satisfaction level λ varies. The curve exhibits a nonlinear shape: small increases in cost near the critical utilization region lead to large reductions in waiting time, whereas further cost increases beyond the optimized point yield diminishing returns. This behavior is consistent with classical queueing insights and highlights the importance of congestion-aware decision making [5, 9].

The trade-off analysis reinforces the value of the integrated fuzzy-queueing framework. Rather than selecting a solution based solely on crisp averages, decision makers can visualize how ambiguity and service aspirations interact, enabling informed and transparent staffing decisions [11, 1].

Overall, the results demonstrate that the proposed approach produces operationally meaningful, robust, and interpretable solutions that outperform baseline staffing policies under uncertainty.

Conclusions: This paper presented an integrated decision-analytic framework that combines fuzzy mathematical

programming with queueing theory to address congestion-sensitive operational planning under ambiguity. Motivated by real-world service systems such as bank teller counters and hospital registration desks, the study recognized that key parameters arrival rates, service rates, costs, and service targets are often imprecise and better represented through fuzzy constructs rather than precise point estimates. By embedding queueing-performance measures directly into a fuzzy optimization structure, the proposed framework bridges a methodological gap between congestion modeling and uncertainty-aware decision making.

The core contribution lies in translating queueing outputs, such as expected waiting time and utilization, into fuzzy objectives and constraints handled through satisfaction-level maximization or α-cut based deterministic equivalents. This integration preserves the analytical insights of queueing theory while enabling transparent representation of ambiguity.

The case study on bank teller staffing demonstrated that modest increases in capacity can significantly reduce waiting times when systems operate near critical utilization levels. The optimized solution achieved improved service quality with a balanced increase in operating cost, yielding a robust and managerially interpretable outcome. Sensitivity analysis further showed that the staffing decision remained stable across a range of ambiguity levels, reinforcing the practical reliability of the approach.

From a managerial perspective, the framework provides actionable insights by explicitly quantifying trade-offs between cost efficiency and service quality under uncertainty. The satisfaction level λ offers a clear indicator of how well competing fuzzy goals are jointly achieved, supporting informed decision making in environments where data limitations and expert judgment play a central role [11].

Several extensions offer promising directions for future research. These include incorporating customer abandonment and retrials, extending the framework to multi-class or networked queueing systems, and integrating time-varying arrivals. Hybrid fuzzy-stochastic formulations could also be explored to combine linguistic ambiguity with probabilistic variability. Such developments would further enhance the applicability of the proposed framework to complex, data-constrained operational systems [9].

References

1. Bector CR, Chandra S. Fuzzy mathematical programming and fuzzy matrix games. Studies in Fuzziness and Soft Computing. Vol 169. Berlin, Heidelberg: Springer; 2005.
2. Bertsimas D, Mourtzinou G. Transient laws of non-stationary queueing systems and their applications. Queueing Systems. 1997;25:115-155.
3. Dubois D, Prade H. Fuzzy sets and systems: Theory and applications. Mathematics in Science and Engineering. Vol 144. New York: Academic Press; 1980.
4. Green LV. Queueing analysis in healthcare. In: Hall RW, editor. Patient flow: Reducing delay in healthcare delivery. International Series in Operations Research & Management Science. New York: Springer; 2006. p. 281-307.
5. Gross D, Harris CM. Fundamentals of queueing theory. 3rd ed. New York: John Wiley & Sons; 1998.
6. Hillier FS, Lieberman GJ. Introduction to operations research. 10th ed. New York: McGraw-Hill Education; 2015.
7. Inuiguchi M, Ramik J. Possibilistic linear programming: A brief review of fuzzy mathematical programming and a comparison with stochastic programming in portfolio selection problem. Fuzzy Sets and Systems. 2000;111(1):3-28.
8. Kaufmann A, Gupta MM. Introduction to fuzzy arithmetic: Theory and applications. New York: Van Nostrand Reinhold; 1985.
9. Whitt W. Stochastic-process limits: An introduction to stochastic-process limits and their application to queues. Springer Series in Operations Research and Financial Engineering. New York: Springer; 2002.
10. Zadeh LA. Fuzzy sets. Information and Control. 1965;8(3):338-353.
11. Zimmermann HJ. Fuzzy programming and linear programming with several objective functions. Information Sciences. 1978;15(1):45-55.