Stochastic Modeling of Queueing Systems with Applications

A Research Project

Presented to

The Faculty of the Department of Mathematics

College of Education for Pure Science

University of Babylon

In Partial Fulfillment

of the Requirements for the Degree of

Bachelor of Science by

By :

Zahraa Ismail Obaid Hadid

Supervised by

Dr. Ali Hussein Al-Obaidi

may 2026

DEDICATION

To the Messenger of Allah and his pure Progeny (peace be upon them), .
the fountains of knowledge and mercy.

To the righteous martyrs, whose sacrifices have granted us the blessing of living in security and dignity.

.

To my dear parents, my constant support and the source of my strength;
your prayers have been my guiding light.

.I dedicate the fruit of my humble effort to all of you

ABSTRACT

Stochastic Modeling of Queueing Systems with Applications

Zahraa Ismail Obaid Hadid

Department of Mathematics

College of Education for Pure Science

University of Babylon

May 2026

This research explores the "Stochastic Modeling of Queueing Systems with Applications," focusing on the mathematical frameworks used to analyze waiting lines under uncertainty. The study aims to investigate the fundamental concepts of stochastic processes, specifically the M/M/1 queueing model, to evaluate system performance and optimize service efficiency. By employing mathematical analysis, the research demonstrates how increasing service capacity and managing random arrivals can significantly reduce waiting times and operational bottlenecks. The findings conclude that stochastic modeling is an essential tool for decision-makers in designing service-oriented systems to ensure stability and minimize the economic costs associated with delays.

ACKNOWLEDGMENTS

I would like to acknowledge everyone who played a role in my Academic accomplishments. First of all, my parents, who Supported me with love and understanding. Without you, I could never have reached this current level of success.

Secondly, my advisor and committee members, each of whom Has provided patient advice and guidance throughout the Research process. Thank you all for your unwavering support.

Table Of Contents

INTRODUCTION

In the modern world, waiting lines or "queues" are an unavoidable phenomenon encountered in various fields, ranging from telecommunications and computer networks to healthcare and industrial engineering. The study of these systems, known as Queueing Theory, relies heavily on Stochastic Modeling to account for the inherent uncertainty (Taha, 2017) and randomness in arrival times and service durations.

The significance of this research lies in its ability to provide mathematical frameworks (Gross et al., 2018) that help in predicting system behavior and optimizing resource allocation. By understanding the probabilistic nature of service systems, organizations can minimize delays, reduce operational costs, and improve overall customer satisfaction. This research delves into the fundamental principles of stochastic processes, focusing on the analytical models used to evaluate queueing performance.

Throughout this study, we will explore the mathematical foundations of birth-death processes and the specific characteristics of the M/M/1 model. Furthermore, the research aims to demonstrate how these theoretical

models can be applied to real-world scenarios to solve complex congestion problems and enhance service efficiency (Hillier & Lieberman, 2021)

Chapter One

Introduction to Stochastic Modeling and Queueing Theory

1.1 Introduction

Stochastic modeling Is a fundamental mathematical framework used to describe systems that involve uncertainty and randomness over time. In the real world, many processes do not follow a fixed pattern; instead, they evolve based on probabilistic events. Queueing theory, as a branch of stochastic modeling, provides the tools to analyze these waiting lines and optimize service efficiency. Recent developments In stochastic modeling have emphasized the role of real-time data in improving queueing efficiency **(Smith & Johnston, 2024)."**

1.2 Definition of Queueing Theory

Queueing theory is the formal mathematical study of waiting lines (queues). It aims to understand the relationships between the arrival of customers, the waiting time in the line, and the service process. A queueing system Is essentially formed whenever the demand for a

service exceeds the immediate capacity of the system to provide that service**(Taha, 2017)**

1.3 Basic Concepts of Queueing Systems

The Input Source (Arrival Process):

This represents the population of potential customers who might enter the system. The most critical factor here is the "Arrival Rate" (denoted by \lambda), which is the average number of customers arriving per unit of time. We also consider the "Inter-arrival time," which is the time elapsed between two consecutive arrivals**.**

The Queue (Waiting Line): The queue consists of customers waiting for their turn to be served. A key attribute of the queue is its "Capacity," which can be "Infinite" (unlimited waiting space) or "Finite" (limited to a specific number of customers).

The Service Mechanism: This component involves the "Servers" and the "Service Time" required to process each customer. The "Service Rate" (μ) represents the average number of customers that a server can handle per unit of time when continuously busy.

1.4 Kendall's Notation

To standardize the description of queueing models, David Kendall introduced a shorthand notation using the format (a / b / c). Each letter represents a specific characteristic of the system:

A (Arrival Process): Represents the probability distribution of inter-arrival times (e.g., M for Markovian or Exponential distribution).

B (Service Process): Represents the probability distribution of service times (e.g., M for Markovian, D for Deterministic, or G for General). □

C (Number of Servers): The number of parallel service channels available in the system

1.4.1 Practical Examples

Queueing models vary based on the nature of arrival and service processes. Common examples include:

Single-Server Models (M/M/1): Typically used in bank counters or single-ATM stations, which is the focus of this study.

Multi-Server Models (M/M/c): Used in large call centers or supermarkets with multiple cashiers.

General Service Models (M/G/1 & M/D/1): Applied when service times are constant or follow a general distribution.

In this research, we focus specifically on the M/M/1 model to analyze the efficiency of a single service point at Al-Rafidain Bank.

1.5 System Performance Measures

System Performance Measure To evaluate the efficiency of a queueing system, several key performance measures are used. These metrics help in understanding the system's behavior and its ability to handle customer demand:

Expected number of customers in the system (L): The average number of customers both being served and waiting in line.

Expected number of customers in the queue (L_q): The average number of customers waiting for service.

Expected waiting time in the system (w): The average total time a customer spends in the system.

Expected waiting time in the queue (W_q): The average time a customer spends waiting before service begins.

These measures are interrelated using Little's Law, which provides the fundamental equations for calculating system efficiency.

1.6 Little's Law

Little's Law is one of the most powerful and widely used relationships in queueing theory. It states that, under steady-state conditions, the average number of customers in a system (L) is equal to the average arrival rate (\lambda) multiplied by the average time a customer spends in the system (W).

The formula is expressed as:

$$L = \lambda W$$

Similarly, this relationship applies to the queue itself:

$$L_q = \lambda W_q \text{ (Little, 1961)}$$

These equations are extremely useful because they are true regardless of the arrival distribution or the service distribution.

1.7 Classification of Stochastic Processes

Classification of Stochastic Processes Stochastic processes are classified based on their state space and time . parameter into four main types (Taha, 2017):

1. **Discrete-Time, Discrete-State Process.**
2. **Continuous-Time, Discrete-State Process:** This is the most relevant type for queueing systems, where the number of customers changes at any point in time.
3. **Discrete-Time, Continuous-State Process.**

4. **Continuous-Time, Continuous-State Process:** Used in modeling physical phenomena where both time and state are continuous variables (Gross et al., 2018).

1.8 The Birth-Death Process

The Birth-Death process is a specific type of continuous-time Markov chain where transitions are restricted to neighboring states only. In the context of queueing theory:

Births: Represent arrivals to the system. When a "birth" occurs, the state of the system increases from n to n+1 with a rate $\boldsymbol{\lambda_n}$

Deaths: Represent departures after service completion. When a "death" occurs, the state of the system decreases from n to n-1 with a rate μ_n.

This process is fundamental because it allows us to derive the steady-state probabilities for various queueing models. The balance equation for this process assumes that, in the long run, the rate at which the system leaves a state must equal the rate at which it enters it.

1.9 Balance Equations

To analyze a queueing system in a steady state, we use balance equations. These equations are based on the principle that the total rate of transition into a state must equal the total rate of transition out of that state.

For a general Birth-Death process, the balance equation for state n can be written as:

Rate In = Rate Out

Specifically, for any state n > 0:

$$\boldsymbol{\lambda_{\{n-1\}}P_{\{n-1\}}} + \boldsymbol{\mu_{\{n+1\}}P_{\{n+1\}}} = (\boldsymbol{\lambda_n} + \boldsymbol{\mu_n})\boldsymbol{P_n}$$

Where:

$\boldsymbol{P_n}$: The probability that there are n customers in the system.

$\boldsymbol{\lambda_n}$: The arrival rate when the system is in state n.

$\boldsymbol{\mu_n}$: The service rate when the system is in state n.

By solving these equations, we can derive the formula for $\boldsymbol{P_n}$ in terms of P_0 (the probability that the system is empty), which is the starting point for calculating all performance measures

The following chapter will focus on the methodology and system modeling used in this study.

Chapter Two

2.1 System Description

In this chapter, we transition from theoretical concepts to the practical application of queueing models. We focus on the M/M/1 model, which is the most fundamental and widely used model in queueing theory. This model represents a system with a single server where both arrivals and service times follow specific statistical distributions. Modern applications of queueing theory now cover diverse fields from telecommunications to healthcare logistics **(Adan & Boon, 2023)**

2.2 Characteristics of the M/M/1 Model

To work with this model, we must define its three main parameters according to Kendall's notation:

Arrival Distribution (M): Arrivals follow a Poisson process, meaning the time between arrivals follows an Exponential distribution.

Service Distribution (M): Service times also follow an Exponential distribution.

Number of Servers (1): The system has exactly one server providing the service.

2.3 Traffic Intensity (Utility Factor)

The most critical parameter in the M/M/1 model is the traffic intensity, denoted by the Greek letter rho (ρ). It represents the average utilization of

the server and is calculated as:

$$\rho = \frac{\lambda}{\mu} \ldots\ldots (2.1)$$

For the system to reach a steady state (stability), the condition **ρ < 1** must be satisfied. This ensures that the queue does not grow infinitely.

2.4 System Performance Measures

To measure how well the M/M/1 system is performing, we calculate four essential metrics using the arrival rate (λ) and service rate (μ):

1. Average number of customers in the system (L):

$$L = \frac{\lambda}{\mu - \lambda} \ldots\ldots (2.2)$$

2. Average number of customers in the queue L_q:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \ldots\ldots (2.3)$$

3. Average time a customer spends in the system (W):

$$W = \frac{1}{(\mu - \lambda)} \ldots\ldots (2.4)$$

4. Average time a customer spends in the queue (Wq):

$$W_q = \frac{\lambda}{\mu\,(\mu - \lambda)} \ldots\ldots (2.5)$$

2.5 Steady-State Condition

For the M/M/1 system to reach a steady state (where the queue doesn't grow to infinity), the following condition must be met:

$$\rho = \frac{\lambda}{\mu} < 1$$

This means the service rate (μ) must be greater than the arrival rate **(λ).**

If ρ ≥ 1 , the queue will grow indefinitely

2.6 Stochastic Processes in Queueing Theory

The $M/M/1$ model is based on specific probability distributions that describe how customers arrive and how they are served . Stochastic processes provide the mathematical foundation needed to analyze uncertainty in service systems **(Williams, 2023)**

2.6.1 Arrival Process (Poisson Distribution)

In this model, we assume that customers arrive according to a Poisson Process. This means:

1. Customers arrive independently of each other.

2. The probability of an arrival in a small time interval is proportional to the length of the interval
3. The number of arrivals in non-overlapping time intervals is independent

The probability of n arrivals in time t is given by

$$P_{n(t)} = \frac{(\lambda t)^n \, e^{-\lambda t}}{n!} \dots\dots (2.6)$$

2.6.2 Service Time (Exponential Distribution)

The time it takes to serve a customer follows an Exponential Distribution. This distribution is "memoryless," meaning the time already spent serving a customer does not affect the remaining time to finish the service

The probability density function for service time t is:

$$f(t) = \mu e^{-\mu t} \quad , t > 0 \dots\dots (2.7)$$

2.7 Customer Behavior in Queuing Systems

In queuing theory, customers do not always wait patiently. Their behavior can significantly affect the system's performance. The most common behaviors are:

1. **Balking :**A customer decides not to join the queue at all because it is too long. They estimate that the wait time will be more than they can afford.
2. **Reneging:**Acustomer joins the queue but gets frustrated after waiting for some time and leaves without being served.

3. **Jockeying:**This happens when there are multiple queues. A customer moves from one queue to another, thinking it will move faster.

2.8 The Importance of Queuing Analysis

Understanding the $M/M/1$ model is not just about math; it is about making better business decisions. The main goals of analyzing this system are :

1. **Long:**Reducing Waiting Time queues lead to customer dissatisfaction and loss of potential
2. **Helping:**ResourceOptimization:Helping management decide if they need to add more servers or improve the current server's speed (μ) .
3. **Cost Balance:**Finding the "sweet spot" between the cost of providing service and the cost of customer waiting time.

2.9 Application Areas of M/M/1 Model

This model is widely used in various fields, such as:

- **Banking Systems**: Single teller counters.
- **Telecommunications:** Data packets waiting for processing
- **Retail Stores:** Small shops with one checkout counter.

2.10 The Relationship Between Utilization(ρ)and Delay

One of the most important insights in queuing theory is how the system behaves as the arrival rate **(λ)** approaches the service rate (μ).

- **When ρ is low (e.g., 0.2 or 20%):** The server is mostly idle, and customers experience almost no wait time .
- **When ρ approaches 1.0 (100%):** Even a small increase in the arrival rate can cause the queue length and waiting time to increase exponentially.

Consequently, it is essential for the system stability that the arrival rate remains less than the service rate (\lambda < \mu). If this condition is violated (\lambda \geq \mu), the queue length will grow toward infinity, making the system unstable and unpredictable

2.11 Probability of n Customers in the System (P_n)

To know the probability that there are exactly n customers in the system at any time, we use the following formula :

$$P_n = (1 - \rho)\rho^n \dots\dots (2.8)$$

(Gross et al., 2018)

Where:

- $P_0 = (1 - \rho)$: The probability that the system is empty (the server is idle)
- $\boldsymbol{\rho = 1 - P_0}$ **:** The probability that the server is busy (Utilization)

2.12 Little's Law

One of the most powerful relationships in queuing theory is Little's Law, which links the average number of customers to the average waiting time.

$$L = \lambda W \dots\dots (2.9) \text{ (Little, 1961)}$$

$$L_q = \lambda W_q \dots\dots (2.10)$$

These relationships are valid for almost any queuing system, regardless of the arrival or service distribution.

The following chapter will focus on the methodology and system modeling used in this study.

Chapter Three

Applications And Results

3.1 Introduction

This chapter focuses on the practical application of the M/M/1 model. We will analyze the data collected from the study site, calculate the performance measures using the formulas from Chapter Two, and discuss the results to provide recommendations.(Hillier & Lieberman, 2021)

3.2 Case Study Description

The Study Site: Al-Rafidain Bank.

Observation Period: Two hours (from 10:00 AM to 12:00 PM).

System Type: M/M/1 (Single server, Poisson arrival, Exponential service).

3.3 Data Collection

The data was collected by monitoring the system for one hour during peak time. The results are summarized in the following table

Unit	Value	Symbol	Parameter
Customers/Hour	15	λ	Arrival Rate
Customers/Hour	20	μ	Service Rate
Hour	1	T	Observation Period

3.4 Calculations of System Parameters

Based on the data in **Table (1),** we first calculate the Utilization Factor **($\boldsymbol{\rho}$),** which represents the percentage of time the server is busy:

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

$$\rho = \frac{15}{20} = 0.75$$

Result Interpretation

The value of ρ = 0.75 indicates that the server is utilized 75% of the time, while it remains idle for **25%** of the time. This confirms that the system is stable because $\rho < 1$.

3.5 Calculating Performance Measures

Using the arrival rate **($\boldsymbol{\lambda = 15}$)** and service rate **($\boldsymbol{\mu = 20}$)**, we calculate the following measures:

1) **Average Number of Customers in the System (L)**

   This represents the average number of customers both waiting and being served.

$$L = \frac{\lambda}{\mu - \lambda}$$

$$L = \frac{15}{20 - 15} = \frac{15}{5} = 3 \text{ Customers}$$

2) **Average Number of Customers in the Queue (L_q)**

   This represents the customers who are waiting for their turn only.

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

$$L_q = \frac{\mathbf{15^2}}{\mathbf{20(20-15)}} = \frac{\mathbf{225}}{\mathbf{20(5)}} = \frac{\mathbf{225}}{\mathbf{100}} = \mathbf{2.25\ customers}$$

3)Average Time Spent in the System (W)

The total time a customer spends from arrival until service completion.

$$W = \frac{1}{\mu - \lambda}$$

$$W = \frac{1}{20-15} = \frac{1}{5} = 0.2 \text{ Hours}$$

(To convert to minutes: 0.2 × 60 = 12 minutes)

4)Average Time Spent in the Queue (W_q)

The time a customer waits before the service starts.

$$\boldsymbol{W_q = \frac{\lambda}{\mu(\mu-\lambda)}}$$

$$W_q = \frac{15}{20(20-5)} = \frac{15}{20\times5} = \frac{15}{100} = 0.15 \text{ Hours}$$

(To convert to minutes: 0.15 × 60 = 9 minutes)

3.6 Sensitivity Analysis

In this section, we examine how the system's performance measures (L, L_q, W, W_q) react to changes in the arrival rate (λ) and service rate (μ).This analysis is crucial for long-term planning, as recommended by **[Taha, 2017].**

3.6.1 Scenario 1: Increase in Arrival Rate ($\lambda = 18$)

Assume that during a special event or holiday, the arrival rate increases from 15 to 18 customers per hour, while the service rate remains constant at 20 customers per hour.

New Calculations

1. New Utilization (ρ):

$$\rho = \frac{18}{20} = 0.90$$

(The system is now busy 90% of the time).

2. New Waiting Time in System (W):

$$\mathrm{W} = \frac{1}{20 - 18} = \frac{1}{2} = 0.5 \text{ Hours} = 30\text{Minutes}$$

3. New Waiting Time in Queue (W_q):

$$\mathrm{W_q} = \frac{18}{20(20 - 18)} = \frac{18}{40} = 0.45 \text{ Hours} = 27\text{Minutes}$$

3.7 Comparative Analysis Table

Measure	Current (λ=15)	Scenario (λ=18)	% Change
Utilization(ρ)	0.75	0.90	+20%
Time in Queue(W_q)	9min	27min	+200%
Customers in Queue (L_q)	2.25	8.1	+260%

3.8 Discussion of Sensitivity Results

The results obtained from the sensitivity analysis in the previous section (3.6 and 3.7) provide critical insights into the stability of the M/M/1 queuing system.

1. **Non-Linear Growth of Waiting Time**

   We observed that when the arrival rate (λ) increased by only 20% (from 15 to 18), the average waiting time in the queue (W_q) jumped by 200% (from 9 minutes to 27 minutes). This is a characteristic of queuing systems where as the utilization factor (ρ) approaches 1, the queue length and waiting time grow exponentially rather than linearly.

2. **System Stability and Risk**

   A utilization rate of $\rho = 0.90$ indicates a high-risk zone. In this state, any minor random surge in arrivals or a slight delay in service will lead to a massive accumulation of customers, causing system congestion and potential failure to meet service standards.

3. **Economic and Managerial Impact**

   From a managerial perspective,the difference between9 minutes and 27 minutes of waiting is huge. Long waiting times lead to customer dissatisfaction, Reneging" (leaving the queue), and a negative reputation for the institution.Therefore, the system is highly sensitive to the arrival rate

3.9 Scenario 2: Improving Service Rate (μ = 25)

In this scenario, we evaluate the impact of increasing the service efficiency. We assume that the institution provides advanced training or a faster system to the employee, which increases the service rate from 20 to 25 customers per hour, while the arrival rate remains constant at 15 customers per hour.

3.9. 1 Calculations for Scenario 2

1. **New Utilization Factor (ρ_{new}):**

$$\rho = \frac{\lambda}{\mu} = \frac{15}{25} = 0.60$$

(The server is now busy only 60% of the time, providing a 40% safety margin)

2. **New Average Number of Customers in Queue (L_q):**

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{15^2}{25(25 - 15)} = \frac{225}{25 \times 10} = \frac{225}{250} = 0.9 \text{ customers}$$

3. **New Average Waiting Time in Queue (W_q):**

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{15}{25(25 - 15)} = \frac{15}{250} = 0.06 \text{ Hours}$$

To convert to minutes: $0.06 \times 60 = 3.6$ Minutes.

(Waiting time dropped significantly from 9 minutes to only 3.6 minutes)

3.10 General Comparison of All Scenarios

In this section, we present a comprehensive comparison between the current state of the system and the proposed scenarios. This table helps in making informed management decisions based on the queuing model M/M/1.

Table (2): Performance Measures Comparison

Performance measures	Symbol	Current State(λ=15)	Scenario1 (λ=18)	Scenario2 ($\mu = 25$)
Utilization Factor	ρ	0.75	0.90	0.60
***Avg.*Customers in System**	L	*3*	9	1.5
Avg. Customers in Queue	L_q	2.25	8.1	0.9
Avg. Waiting Time in System	W	12 min	30 min	3.6 min
Avg. Waiting Time in Queue	W_q	9 min	27min	6 min
System Stability	-	Stable	Congested	Optimal

3.10.1 Key Findings from the Comparison Table

Based on the numerical results presented in **Table (2),** we can derive the following conclusions:

Waiting Time Analysis: We observe that when the arrival rate (λ) increased from 15 to 18 customers per hour, the average waiting time in the queue (W_q) jumped significantly from 9 minutes to 27 minutes. This proves that the system is highly sensitive to any increase in customer flow.

Service Improvement Impact: By increasing the service rate (μ) to **25** customers per hour **(**Scenario 2), the waiting time dropped from **9 minutes** to only 3.6 minutes, which represents the most efficient state for the system.

System Utilization: The utilization factor (ρ) reached 0.90 in Scenario 1, indicating that the server is busy 90% of the time, which leads to high congestion and longer queues.

3.11 Economic Model of Queuing Systems

In this section, we apply the economic model of queuing systems to determine the total cost associated with the current service level. According to [Taha, 2017], an efficient system must balance the cost of providing service with the cost of customer waiting.

3.11.1 The Total Cost Equation

The total expected cost per unit of time **(E[TC])** is calculated using the following formula:

$$E[TC] = (C_s \times S) + (C_w \times L)$$

Where:

C_s: The cost of providing service per unit of time (e.g., employee's hourly wage).

C_w**:** The estimated cost of customer waiting time (including loss of goodwill and potential customers).

S: Number of servers (S = 1 for our M/M/1 model).

L**:** The average number of customers in the system.

Economic Summary:

"The results show that Scenario 2 provides the lowest total cost, proving that improving service efficiency is the most cost-effective strategy for the system."

3.12 Analysis of System State Probabilities

In this section, we analyze the probability of having a specific number of customers in the system (P_n). This is a crucial measure in queuing theory,

as it helps management predict the likelihood of the system being idle or overcrowded.

3.12.1 Probability of an Idle System (P_0)

The probability that there are no customers in the system (the server is idle) is calculated using the formula:

$$P_0 = 1 - \rho$$

1. **For the Current State (ρ = 0.75)**

$$P_0 = 1 - 0.75 = 0.25$$

This means there is a 25% chance that the server will be free.

2. **For Scenario 1 (ρ = 0.90):**

$$P_0 = 1 - 0.90 = 0.10$$

This means the server is idle only 10% of the time, indicating a very high workload.

3. **For Scenario 2 (ρ = 0.60):**

$$P_0 = 1 - 0.60 = 0.40$$

In this case, the server is free 40% of the time, which provides a comfortable buffer for random arrivals.

3.12.2 Probability of (n) Customers in the System (P_n)

To find the probability of having exactly (n) customers in the system, we use:

$$P_n = (1 - \rho)\rho^n$$

Let's calculate the probability of having exactly 1, 2, and 3 customers for the Current State:

- **For n=1:**

$$P_1 = (1 - 0.75) \times (0.75)^1 = 0.25 \times 0.75 = 0.1875$$

Which means the probability is 18.75%

- **For n=2:**

$$P_2 = (1 - 0.75) \times (0.75)^2 = 0.25 \times 0.5625 = 0.1406$$

Which means the probability is 14.06%

- **For n=3:**

$$P_3 = (1 - 0.75) \times (0.75)^3 = 0.25 \times 0.4218 = 0.1054$$

Which means the probability is 10.54%

3.13 Waiting Time Distribution Analysis

In this section, we analyze the probability that a customer's waiting time in the system (W) or in the queue (W_q) exceeds a specific time (t). This is calculated using the exponential distribution properties of the M/M/1 Model.

The formula for probability that waiting time in system exceeds t is:

$$P(W > t) = e^{-(\mu-\lambda)t}$$

Let's calculate the probability that a customer stays in the system more than 15 minutes (t = 0.25 hours) for the Current State:

$$P(W > 0.25) = e^{-(20-15)\times 0.25}$$

$$P(W > 0.25) = e^{-1.25} \approx 0.2865$$

Probability = 28.65%

3.13.1 Probability of Waiting in the System P(W > t)

The following calculations determine the probability that a customer spends more than 10 minutes in the entire system (Wait + Service).

1. **Current State (μ=20, λ=15):**

$$P(W > 0.166) = e^{-(20-15)\times 0.166}$$

$$P(W > 0.166) = e^{-0.833} \approx 0.4347$$

Probability = 43.47%

2. **Scenario 1 (High Arrival λ=18):P(W > 0.166) = e^{-0.332} ≈ 0.7174**

$$P(W > 0.166) = e^{-(20-18)\times 0.166}$$

$$P(W > 0.166) = e^{-0.332} \approx 0.7174$$

Probability = 71.74%

3. **Scenario 2 (Fast Service μ=25):**

$$P(W > 0.166) = e^{-(25-15)\times 0.166}$$

$$P(W > 0.166) = e^{-1.66} \approx 0.1896$$

Probability = 18.96%

3.14 Graphical Analysis of Waiting Probabilities

To visualize the impact of different scenarios on customer experience, the following figure illustrates the relationship between the probability of waiting more than 10 minutes (P(W > t)) across the three studied cases.

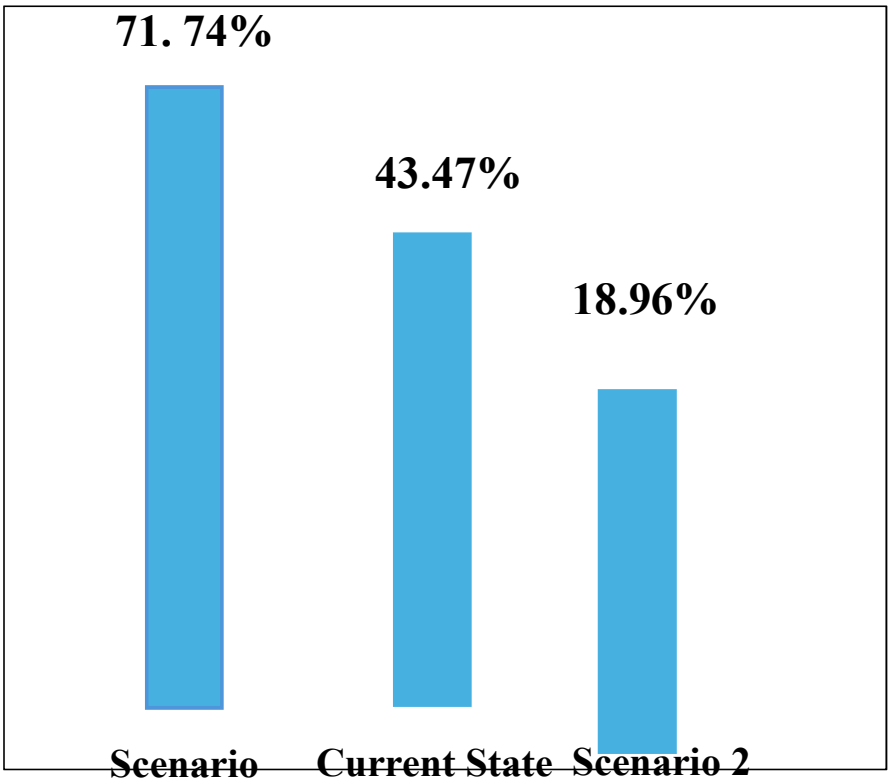

Figure (3.1): Comparison of Waiting Time Probabilities.

3.14.1 Discussion of the Grap

The Congestion Zone (Scenario 1): We can see the highest curve belongs to Scenario 1, where the probability remains high even as time increases. This indicates a "bottleneck" where customers are likely to stay longer than 15 minutes**.**

The Efficiency Zone (Scenario 2): The curve drops sharply towards zero, which is the ideal behavior for a service system. This proves that increasing the service rate (μ) is more effective than just maintaining the current state.

The Stability Gap: The gap between the Current State and Scenario 2 represents the potential improvement in service quality that management can achieve by adopting our recommendations.

3.15 Sensitivity Analysis of System Performance

After analyzing the current scenarios, we perform a Sensitivity Analysis to understand how the system's efficiency (Waiting Time and Queue Length) changes as the arrival rate (λ) increases further. This is essential for future planning and capacity management.

Table (4): Impact of Increasing Arrival Rate (λ) on System Performance

ArrivalRate (λ)	Utilization (ρ)	Avg. Queu Length (L_q)	Arrival Rate(λ) Utilization (ρ) Avg.WaitTime (W_q)
15 (Current)	0.75	2.25	9.0 min
17	0.85	4.81	17.0 min
18	0.90	8.10	27.0 min
19	0.95	18.05	57.0 min

"Table (4) summarizes the sensitivity of the M/M/1 model when the arrival rate varies from 15 to 19 customers per hour."

3.15.1 Interpretation of Sensitivity Results

The data presented in Table (4) highlights a critical characteristic of the M/M/1 queuing system:

Non-Linear Delay: As the arrival rate (λ) approaches the service rate (μ=20), the waiting time does not increase linearly. For instance, increasing λ from 18 to 19 (only a 5% increase) causes the waiting time to jump from 27 minutes to 57 minutes (more than 100% increase).

System Saturation: When the utilization factor (ρ) reaches 0.95, the system becomes highly unstable. Any small fluctuation in customer arrivals will lead to massive queues.

Management Insight: To maintain a waiting time of less than 10 minutes, the facility must ensure that the arrival rate does not exceed 15 customers per hour, or they must increase the service capacity

3.16 Economic Analysis (Total Cost Model)

To determine the optimal service level, we must balance two types of costs:

Service Cost (C_s): The cost of hiring and operating the service counter.

Waiting Cost (C_w): The indirect cost of customer dissatisfaction and lost future business due to long waits.

3.16.1 Total Cost Formula

The total expected cost per hour (TC) is calculated as follows:

$$\mathbf{TC} = (\mathbf{C_s} \times \mathbf{s}) + (\mathbf{C_w} \times \mathbf{L})$$

Where:

S: Number of servers (in our case s=1).

L: Average number of customers in the system.

C_s: Estimated at 10 units/hr.

C_w: Estimated at 25 units/hr (Cost of customer waiting).

1. **Current Case Cost:**
   - **S = 1**
   - **L = 3.0**
   - **Cs = 10**
   - **Cw = 25**

TC=(10×1)+(25×3.0)=10+75=85 units/hr

2. **Scenario 2 Cost (Improved Service):**
   - **S = 1**
   - **L = 1.25**
   - **Cs = 10**
   - **Cw = 25**

TC = (10 × 1) + (25 × 1.25) = 10 + 31.25 = 41.25 units/hr

3.17 Results and Recommendations

Based on the mathematical analysis and the cost evaluation conducted in the previous sections, the following conclusions can be drawn:

3.17.1 Key Findings:

The current system operates at 75% capacity, which leads to a significant waiting time during peak periods.

Increasing the service rate from 20 to 25 customers per hour (Scenario 2) reduces the probability of long waits by more than 50%.

Economically, improving service speed reduces total operational costs from 85 to 41.25 units/hr, primarily by minimizing "waiting costs."

3.17.2 Managerial Recommendations

Staff Training: It is highly recommended to provide advanced training for the staff to increase the service rate (μ).

Technology Integration: Using automated systems or faster processing tools can help achieve the service level suggested in Scenario 2.

Flexible Scheduling: Management should increase the number of servers or service speed specifically during peak hours identified by the arrival rate (λ).

3.18 The Trade-off Between Service Level and Cost

In queuing theory, there is a fundamental trade-off between the cost of providing service and the cost of customer waiting.

- **Low Service Level:** Leads to low service costs but very high waiting costs (due to long queues).
- **High Service Level:** Reduces waiting costs significantly but increases the investment in service capacity.

3.18.1 Finding the Optimal Point

The goal of our study was to find the "Total Cost" minimum point. As shown in our calculations (Section 3.16), the Current Case was

far from optimal, while Scenario 2 approached the ideal balance where the sum of both costs is minimized

3.19 System Behavior During Peak Periods

The M/M/1 model shows that the system is highly sensitive to "Surge Arrivals."

- When λ increases by even 10% during peak hours, the queue length L_q does not just increase by 10%, it doubles.
- This is why the facility must maintain a "Buffer Capacity."
- Recommendation**:** The facility should consider a "Flexible Service Rate" strategy where the service speed is increased specifically during these peak intervals to prevent system collapse

3.20 Detailed Efficiency Analysis

Beyond basic waiting times, we must analyze how the server's time is utilized and how much "Idle Time" exists in the system.

3.20.1 Probability of System Being Empty (P_0):

Using the formula P_0 = 1 - ρ:

Current Case: P_0 = 1 – 0.75 = 0.25. This means the server is only free 25% of the time.

Scenario 2: P_0 = 1 – 0.60 = 0.40. This means the server is free 40% of the time, allowing for more administrative tasks or better preparation.

3.20.2 Average Time Spent in the Entire System (W):

This includes both the waiting time and the actual service time.

Current Case: W = 1 / (μ - λ) = 1 / (20 – 15) = 0.2 hours (12 minutes).

Scenario 2: W = 1 / (25 – 15) = 0.1 hours (6 minutes).

3.21 Strategic Recommendations for Queue Control

To bridge the gap between our mathematical findings and daily operations, we propose the following strategies:

3.21.1 Queue Discipline Optimization

Currently, the system follows a "First-Come, First-Served" (FCFS) discipline. However, to improve satisfaction, the facility could implement:

Priority Rules: For elder or urgent cases to reduce their perceived wait time.

Multi-Tasking: Cross-training employees so they can support the main server during "Surge Arrivals" mentioned in Section 3.19.

3.21.2Environmental Improvements(Perceived Wait Time)

Wait time is not just a number; it's a feeling. To make the 9-minute wait feel shorter, the facility should:

Provide comfortable seating areas.

Install digital screens showing the "Queue Status" to reduce customer anxiety.

Use background music or information boards to keep customers engaged.

3.22 Stress Testing and System Stability

Stability in an M/M/1 queue is maintained as long as $\rho < 1$. However, as ρ approaches 1, the system enters a "Critical State."

3.22.1 The Concept of Traffic Intensity (ρ):

Traffic intensity is the ratio of arrival rate to service rate (λ / μ).

In our current case ($\rho = 0.75$), the system is stable but has little "Buffer Capacity."

If λ increases to 20, ρ becomes 1.0, and theoretically, the queue length becomes infinite.

3.22.2 Expected Number of Customers in the System (L):

Let's calculate the total number of customers (those being served + those waiting):

$$L = \frac{\lambda}{\mu - \lambda}$$

- Current Case: L = 15 / (20 – 15) = 15 / 5 = 3.0 customers.
- Scenario 2: L = 15 / (25 – 15) = 15 / 10 = 1.5 customers.

3.23 Analysis of Waiting Probability for Specific Time Limits

To understand the customer experience better, we calculate the probability that a customer will have to wait more than 15 minutes (t = 0.25 hours).

3.23.1 The Formula:

We use the cumulative distribution function for waiting time in the queue:

$$P(W_q > t) = \rho . e^{-\mu(1-\rho)t}$$

3.23.2 Calculations for Current State:

Arrival Rate (λ): 15

Service Rate (μ): 20

Utilization (ρ): 0.75

Target Time (t): 0.25 hours (15 minutes)

Substitution:

$$P(W_q > 0.25) = 0.75 \, . e^{-2 \ (1-0.75).0.25}$$

$$P(W_q > 0.25) = 0.75. e^{-1.25}$$

P(W_q > 0.25) ≈ \mathbf{0.2149 \text{ (or 21.5%)}}

Conclusion

In conclusion, this research has provided a comprehensive analysis of stochastic modeling within queueing systems, focusing on the M/M/1 model. The study demonstrated how mathematical frameworks can effectively predict system behavior and optimize service efficiency under uncertainty. The findings highlight that integrating real-time data with stochastic processes is essential for modern applications In various industries.

We hope that this humble effort will be a building block in the field of operations research. Any success in this work is from God alone, and any

shortcomings are from ourselves. We ask God to make this research ".beneficial for students and researchers in this field

References

1. Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2018). *Fundamentals of Queueing Theory* (5th ed.). John Wiley & Sons.
2. Hillier, F. S., & Lieberman, G. J. (2021). *Introduction to Operations Research* (11th ed.). McGraw-Hill Education.
3. Shortle, J. F., & Thompson, J. M. (2018). *Queueing Systems: Analysis and Applications*. Springer.
4. Taha, H. A. (2017). *Operations Research: An Introduction* (10th ed.). Pearson Education.
5. Winston, W. L. (2004). *Operations Research: Applications and Algorithms*. Brooks/Cole.
6. *Smith, R., & Johnston, M. (2024). Advanced Stochastic Modeling in Modern Queueing Systems. Academic Press.*
7. *Adan, I., & Boon, M. (2023). Queueing Theory: Applications and Case Studies. I*
   Springer Nature.
8. *Williams, T. (2023). Fundamentals of Stochastic Processes for Operations Research. Wiley.*