People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

`Number assigned by the library`

Academic Year: 2024/2025

University of Saida Dr Moulay Tahar.
Faculty of MIT, Department of Mathematics

Thesis submitted for the Academic
Master's degree

Sector: Mathematics
Specialty: Stochastic Analysis, Process Statistics and Applications (SASPA)

Presented by

HALIMI **F**atima **H**adjer [1]

Supervised by

Dr. Mokhtar **K**adi

The topic:

Analysis of Customer Impatience in a Simple Multi-Server Queueing System

Board of Examiners

Pr. A. Kandouci	University of Saida Dr. Moulay Tahar	Chair Person
Dr. M. Kadi	University of Saida Dr. Moulay Tahar	Supervisor
Dr. L. Yahiaoui	University of Saida Dr. Moulay Tahar	Examiner

[1] e-mail: hadjerfatima187@gmail.com

Acknowledgements

I would like to express my sincere gratitude to my supervisor, **Dr. M. Kadi**, for suggesting the topic of this thesis and for his continuous support. His patience, valuable advice, and constant encouragement have been truly instrumental throughout the development of this work.

I would also like to extend my heartfelt thanks to the **members of the jury** for the time they have devoted to evaluating this work and for their valuable feedback.

Dedication

I dedicate this humble work to my dear **parents**, whose unwavering support has helped me become who I am today.

To my beloved **brothers and sisters**, wishing them a long life filled with happiness and success.

And to my precious ones, **Meriem** and **Abdelmalek**, with all my love.

Contents

General introduction

The rapid advancements in modern technologies, combined with the constant growth in industrial and service-related needs, have established queueing models as indispensable tools in the modeling and analysis of complex systems. Since the pioneering work of L. Kleinrock, these models have evolved to cover a wide range of domains, including telecommunication networks, logistics, healthcare systems, and customer service environments.

A queueing system relies on fundamental parameters that define its structure and operation: the arrival process of customers, the service time, the queue capacity (finite or infinite), and the service discipline or priority rule. These parameters are formalized through Kendall's notation, a standardized framework that effectively describes and analyzes the various configurations of such systems. Since the first mathematical study conducted by Danish engineer A. K. Erlang in 1917 to optimize the operation of telephone exchanges in Copenhagen, the theory of queueing systems has seen significant advancements [7].

Despite these theoretical developments and the emergence of numerous analytical solutions, challenges remain. These models, while robust, often prove complex to adapt to real-world systems, particularly when human behaviors such as impatience or customer abandonment are considered. This highlights the importance of further research into dynamic systems while integrating stochastic behaviors that accurately reflect real-world service environments.

This thesis aims to study and analyze a specific aspect of queueing systems: customer impatience, a common phenomenon in environments where waiting times are perceived as excessive. Specifically, we focus on multi-server queueing systems, where customer impatience can significantly affect the overall system performance.

Thesis Structure

- **Chapter 1:** Introduction to stochastic processes, which form the essential mathematical foundation for modeling queueing systems. Fundamental concepts include the Poisson process, etc.

- **Chapter 2:** Theory of queueing systems. Presentation of the main notations

(Kendall's notation, Little's law), analysis of key parameters and performance indicators such as average queue length and waiting time.

- **Chapter 3:** Study of customer impatience in a multi-server queueing system (M/M/c). Modeling impatience behavior using exponential patience time distributions, and analyzing performance metrics such as abandonment rates, average waiting time, and the proportion of served customers.

In this study, we focus on the $(M/M/c)$ model, incorporating impatience behavior represented by an exponential distribution of patience times. We derive the equilibrium equations, calculate key performance metrics, and analyze the effects of system parameters on overall behavior. These results provide practical recommendations to optimize the management of queueing systems in various contexts.

Chapter 1

Stochastic Processes

A stochastic process is a collection of random variables indexed by time T and taking values in a set X, allowing us to describe a random phenomenon as it evolves over time.

In this chapter, we will study several key stochastic processes used in modeling queueing systems. We begin with the **counting process**, which tracks the number of events that have occurred up to a given time. Then, we introduce the **Poisson process**, a fundamental example of a counting process widely used to model random arrivals. We will also discuss the **renewal process**, which generalizes the Poisson process by relaxing the assumption of exponential interarrival times. Finally, we conclude with the **birth-and-death process**, a Markov jump process that models the evolution of a population or queue size over time.

Definition 1.0.1. ***(Stochastic Process)***
*A family $X(t)$ of random variables indexed by T and defined on the same probability space is called a **stochastic process**. Generally, $X(t)$ represents the state of the stochastic process at time t.[29]*

- *If T is in $[0, \infty)$, then the stochastic process is called a **continuous-time process**.*
- *If T is countable, i.e., $T \subseteq \mathbb{N}$, then we say that $X(t)_{t \in T}$ is a **discrete-time process**.*

Remark 1.0.1. *The set T is equipped with a total order $\leq$, meaning that for any $(s, t) \in T^2$, either $s \leq t$ or $t \leq s$. We can also consider processes over a finite time horizon:*

- *In the discrete case, we consider $T = \{0, \dots, N\}$, for some final time N.*
- *In the continuous case, we set $T = [0, T]$.*

*The set of values taken by $X(t)$ is called the **state space**, which can be either:*

- ***Discrete** (finite or countably infinite).*
- ***Continuous** (a subset of $\mathbb{R}$ or $\mathbb{R}^n$).*

Therefore, we write $(X_n)_{n\geq 0}$ for a discrete-time process and $(X_t)_{t\geq 0}$ for a continuous-time process.

1.1 Counting Process

Definition 1.1.1. *(***Counting Process***)*
A stochastic process $N(t)_{t\in\mathbb{R}^+}$ is a counting process if $N(t)$ represents the total number of events that have occurred between 0 and t. It must satisfy the following conditions:

- *$N(t) \geq 0$;*
- *$N(t)$ takes only integer values;*
- *for $s < t$, $N(t) - N(s)$ is the number of events that occurred between s and t.*

A counting process is a discrete process in continuous time. A second process can be associated with the process of occurrence times; the interarrival times process $\{T_n, n \in N\}$ where $\forall n \in N$ the random variable T_n represents the waiting time between the $(n-1)^{th}$ and n^{th} occurrences [9], i.e.,

$$T_n = A_n - A_{n-1}$$

where A_n is the arrival time of the n^{th} client.

Proposition 1.1.1. *The following relationships are trivial to verify given that $A_0 = 0$:*

1. *$A_n = T_1 + T_2 + ... + T_n \; \forall n \geq 1$;*
2. *$N(t) = \sup\{n \geq 0 : A_n \leq t\}$;*
3. *$\mathbb{P}[N(t) = n] = \mathbb{P}[A_n \leq t < A_{n+1}]$;*
4. *$\mathbb{P}[N(t) \geq n] = \mathbb{P}[A_n \leq t]$;*
5. *$\mathbb{P}[s < A_n < t] = \mathbb{P}[N_s < n \leq N(t)]$.*

Definition 1.1.2. *(***Renewal Process***)*
A counting process in which the times between two consecutive arrivals are i.i.d. random variables is called a ***renewal process****. The renewal times (or the times of the n-th arrival) are defined as:*

$$A_n = \sum_{i=1}^{n} a_i, \quad n = 0, 1, 2, \ldots$$

We observe that the number of arrivals before time t, i.e., the process

$$(N_t)_{t \in \mathbb{R}_+} = \sup\{k : A_k \leq t\}$$

is a counting process.

Definition 1.1.3. *(***Monotonic Counting Process***)*
A counting process $N(t)_{t \in \mathbb{R}^+}$ is increasing if for all $s \leq t$, $N_s \leq N_t$. The random variable $N_t - N_s$ is then called the increment of the process over $]s, t]$.
Examples:

- *$N(t)$ = number of fish caught in the time interval $[0, t]$;*
- *$N(t)$ = size of a population at time t.*

Definition 1.1.4. *(***Process with Independent Increments***)*
A counting process $N(t)_{t \in \mathbb{R}^+}$ is said to have independent increments if for all $n \in \mathbb{N}^*$ and for all $t_1, ..., t_n$ such that $t_1 < t_2 < ... < t_n$, the increments $N_{t_1} - N_0, N_{t_2} - N_{t_1}, ..., N_{t_n} - N_{t_{n-1}}$ are independent random variables.

Definition 1.1.5. *(***Process with Stationary Increments***)*
A process is said to be stationary (or homogeneous in time) if for all s and t, the increment $N_{t+s} - N_s$ has the same distribution as N_t.

1.2 Poisson Process

We will now introduce a process of a different nature, whose field of applicability is very significant: the Poisson process. In the context that will interest us here, it describes the random and uniform distribution of points on the positive real line.

It can be used to model, for example:

- Telephone calls arriving at a central station.

- Arrival times of customers at a checkout.
- Occurrence times of claims to be compensated by an insurance company, etc.

The arrivals of customers at a queueing system are characterized by the set of arrival times of each customer. The collection of these arrival times can be modeled by the Poisson process.

1.2.1 Definitions

Definition 1.2.1. *(***Poisson Process***)*
Let N_t be the number of occurrences of a random event in the time interval $(0, t]$, $t > 0$ and $N_0 = 0$. If N_t satisfies the following two conditions, we call N a Poisson process with intensity (or rate) $\lambda > 0$,

- *For any two sequences (s_i) and (t_i), with $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2 \leq ... \leq s_n \leq t_n < +\infty$, the random variables $N_{t_1} - N_{s_1}, N_{t_2} - N_{s_2}, ..., N_{t_n} - N_{s_n}$ are independent.*
- *$\forall t > 0$, we have:*

$$\mathbb{P}(N_{t+h} - N_t = k) = \begin{cases} \lambda h + o(h) & \textit{if} \quad k = 1 \\ o(h) & \textit{if} \quad k \geq 2 \\ 1 - \lambda h + o(h) & \textit{if} \quad k = 0 \end{cases}$$

with $\lim\limits_{h \to 0} \dfrac{o(h)}{h} = 0$

Definition 1.2.2. *(***Poisson Process***)*
Let $\lambda > 0$ and $(S_n)_{n \geq 1}$ be a sequence of independent and identically distributed exponential random variables with parameter λ. Define $A_n = S_1 + ... + S_n$. The counting process $N = (N_t)_{t \geq 0}$, taking values in $\mathbb{N} \cup \{+\infty\}$, is given by

$$N_t = \sum_{n \geq 1} \mathbb{1}_{\{A_n \leq t\}}$$

This process is called the Poisson process with intensity λ.

Remark 1.2.1. *The process can also be rewritten as* $N_t = \sup\{n \geq 0 : A_n \leq t\}$*. Conversely, we observe that* $A_n = \inf(t \geq 0 : N_t = n)$*.*
For $t > s$*, we have* $N_t - N_s = \sum\limits_{n \geq 1} \mathbb{1}_{\{s < A_n \leq t\}}$*, meaning* N *is a process with independent and stationary increments.*

Definition 1.2.3. *(***Equivalent Definition of Poisson Process***)*
A Poisson process $N = (N_t)_{t \geq 0}$ *with intensity* λ *is a counting process with right-continuous paths [6] such that:*

- $N(0) = 0$*;*
- N *has independent and stationary increments;*
- *for all* $t \geq 0$*,* N_t *follows a Poisson distribution* $\mathcal{P}(\lambda t)$*.*

1.2.2 Characterization of a Poisson Process by its Arrival Times:

Let A_n be the instant of the n^{th} arrival: $A_n = \inf\{t \geq 0; N_t = n\}$ and T_n be the n^{th} waiting time for $n \in \mathbb{N}^*$: $T_n = A_n - A_{n-1}$ (assuming $A_0 = 0$).
We have $A_n = \sum\limits_{i=1}^{n} T_i$ and $N_t = \max\{n \geq 0; A_n \leq t\}$.

Theorem 1.1. $(N_t)_{t \in \mathbb{R}_+}$ *is a Poisson process with parameter* λ *if and only if the random variables* T_n *are independent and follow an exponential distribution* $\varepsilon(\lambda)$ *with density*

$$f_{T_n}(t) = \lambda e^{-\lambda t} \mathbb{1}_{]0,+\infty[}(t)$$

Proposition 1.2.1. *If* $(N_t)_{t \in \mathbb{R}_+}$ *is a Poisson process with parameter* λ*, the random time* U *separating an instant* θ *from the next event and the random time* V *separating* θ *from the last event both follow an exponential distribution* $\varepsilon(\lambda)$*.*

Proof:

$$P([U > x]) = P([N_{\theta+x} - N_\theta = 0]) = P([N_x = 0]) = e^{-\lambda x}$$

Since [U $> x$] means that during the duration x following θ, there is no arrival. Similarly,

$$P([V > x]) = P([N_\theta - N_{\theta-x} = 0]) = P([N_x = 0]) = e^{-\lambda x}$$

Since [V $> x$] means that during the duration x preceding θ, there was no arrival.

Remark 1.2.2. *We have* $\mathbb{E}(U+V) = \mathbb{E}(U) + \mathbb{E}(V) = \dfrac{2}{\lambda}$ *while* $\mathbb{E}(T_n) = \dfrac{1}{\lambda}$ *for all* $n \in \mathbb{N}^*$. *Thus, as* λ *increases, the average number of arrivals per unit time increases, and the interval between two arrivals decreases, which is intuitively expected. For this reason, the parameter* λ *is also called the intensity of the process.*

1.3 Poisson Distribution and Exponential Distribution

1.3.1 Definitions

Definition 1.3.1. *A discrete random variable* X *takes integer values and follows a Poisson distribution with parameter* $\lambda > 0$ *if:*

$$\forall k \in \mathbb{N}, \qquad \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Definition 1.3.2. *A continuous random variable* Y *takes strictly positive real values and follows an exponential distribution with parameter* $\mu > 0$ *if:*

$$\forall t > 0, \qquad \mathbb{P}(Y = t) = \mu e^{-\mu t}$$

1.3.2 Poisson Distribution:

Let N be a discrete random variable with $N = 0, 1, ...$ following a Poisson distribution. The probability distribution of N is given by:

$$\mathbb{P}(N = n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

The expectation and variance of N are:

$$\mathbb{E}(N) = \lambda \qquad \text{and} \qquad \mathrm{Var}(N) = \lambda, \quad \text{respectively.}$$

The Poisson distribution can also be defined in terms of time t. In this case, the discrete variable n represents the number of occurrences in time t:

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

1.3.3 Exponential Distribution:

Let τ be a continuous random variable with $t \geq 0$ following an exponential distribution. The probability density function of τ is:

$$f(t) = \mu e^{-\mu t}$$

And the corresponding cumulative distribution function is:

$$F(t) = 1 - e^{-\mu t}.$$

The expectation and variance of t are:

$$\mathbb{E}(\tau) = \frac{1}{\mu} \qquad \text{and} \qquad \mathrm{Var}(\tau) = \frac{1}{\mu^2}, \quad \text{respectively.}$$

1.3.4 Relation between Exponential and Poisson Distributions:

The probability density function of an exponential distribution is $f(t) = \mu e^{-\mu t}$. Suppose τ follows an exponential distribution with expectation $\frac{1}{\mu}$, and n follows a Poisson distribution with mean λ. We have:

$$\begin{aligned} \mathbb{P}(\tau > t) &= 1 - F(t) \\ &= e^{-\mu t} \\ &= P(n = 0 \quad \text{in} \quad t) \\ &= P(0,t) \end{aligned}$$

Denoting $P(n,t)$ as the probability of having n units in time t:

$$P(0,t) = e^{-\mu t}$$

$$P(1,t) = \int_{\tau=0}^{t} P(0,\tau) f(t-\tau) d\tau = \mu t e^{-\mu t}$$

$$P(2,t) = \int_{\tau=0}^{t} P(1,\tau)f(t-\tau)d\tau = (\mu t)^2 e^{-\mu t}/2!$$

$$\cdots$$

$$P(n,t) = \int_{\tau=0}^{t} P(n-1,\tau)f(t-\tau)d\tau = (\mu t)^n e^{-\mu t}/n!$$

1.3.5 Memoryless Property of the Exponential Distribution:

When a random variable t follows an exponential distribution, the probability density function is:

$$f(t) = \mu e^{-\mu t},$$

and the corresponding cumulative distribution function is:

$$F(t) = 1 - e^{-\mu t}.$$

For a time increment h, the probability that t exceeds h is:

$$\mathbb{P}(t > h) = e^{-\mu h}$$

Moreover, for $t = (t^{'} + h)$, the probability that t is greater than $(t^{'} + h)$ is:

$$\mathbb{P}(t > (t^{'} + h)) = e^{-\mu(t^{'}+h)}$$

The conditional probability that $t > (t^{'} + h)$ given $t > t^{'}$ is:

$$\mathbb{P}(t > t^{'} + h | t > t^{'}) = \frac{e^{-\mu(t^{'}+h)}}{e^{-\mu t^{'}}} = e^{-\mu h}$$

Since these probabilities are the same, the exponential distribution is referred to as a memoryless probability distribution.

1.3.6 Erlang Process

The Erlang process is a generalization of the Poisson process. An Erlang process is defined as a stochastic process with two main characteristics:

- **Number of events:** The Erlang process is used to model events that occur at a constant rate, similar to the Poisson process.

- **Duration between events:** The duration between each event (called "waiting time") follows an Erlang distribution.

The probability density function of a random variable X following an Erlang distribution of order k and rate λ is given by [14]:

$$f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}, \quad x \geq 0$$

1.4 Birth and Death Process

A process is a collection of random variables $\{Z_t, t \geq 0\}$ indexed by time. Here, it will be used to describe the random evolution over time of the number of individuals in a population or a queueing system.

The random variables Z_t take their values in the set of integers $\mathbb{N}$. The process evolves as a Markov jump process: the number of individuals remains constant for a certain exponential duration, then jumps to another value. Since we are dealing with a population or a queueing system, we will only consider jumps to the two neighboring values: the population size can either increase by 1 (birth or arrival) or decrease by 1 (death or departure). The intensity of these jumps is governed by two sequences of positive real numbers: $(\lambda_n)_{n\in\mathbb{N}}$ (birth rates) and $(\mu_n)_{n\in\mathbb{N}^*}$ (death rates). To avoid special cases, we will assume that these rates are all strictly positive.

1.4.1 Birth Process

Definition 1.4.1. *Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence of strictly positive real numbers. A birth process with birth rates (λ_n) is a Markov jump process $\{Z_t, t \geq 0\}$ taking values in $\mathbb{N}$ such that, for all $n \geq 0$, the transition rate from n to $n+1$ is λ_n.*

- *The birth process is a direct generalization of a Poisson process when the intensity parameter λ depends on the current state of the process. It allows us to introduce the concept of "explosion."*

1.4.2 Death Process

Definition 1.4.2. *Let $(\mu_n)_{n\in\mathbb{N}^*}$ be a sequence of strictly positive real numbers. A death process with death rates (μ_n) is a Markov jump process $\{Z_t, t \geq 0\}$ taking values in $\mathbb{N}$ such that, for all $n \geq 1$, the transition rate from n to $n-1$ is μ_n.*

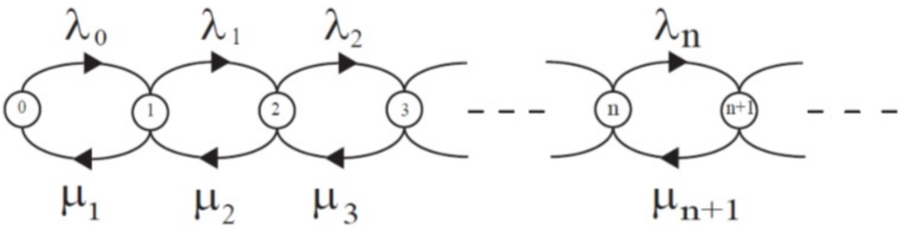


Figure 1.1: Transition graph of a birth and death process

This graph represents the transitions from one state to another. The transition to the right represents a birth, while the transition to the left represents a death.

- If all λ_n are zero, we call it a death process.
- If all μ_n are zero, we call it a birth process.

To understand the dynamics of a birth-and-death process, we can refer back to the construction of a jump process via its embedded chain. Here, it is a Markov chain taking values in $\mathbb{N}$, which jumps from n to $n+1$ with probability $\dfrac{\lambda_n}{\lambda_n+\mu_n}$ and from n to $n-1$ with probability $\dfrac{\mu_n}{\lambda_n+\mu_n}$. The transition from the chain to the process is done by adding independent random sojourn times, which are also independent of the chain, and whose distribution depends on the current state: the sojourn time in state n follows an exponential distribution $\varepsilon(\lambda_n+\mu_n)$.

1.4.2.1 Assumptions:

- The time between two consecutive arrivals is exponentially distributed.
- The service time is also exponentially distributed.

Under these assumptions, a queueing system can be seen as a birth-and-death process:

- Birth ⟷ Arrival of a customer.
- Death ⟷ Departure of a customer from the system after service.

Assumption 1: Birth ⟷ The time between two consecutive births is exponentially distributed.

Assumption 2: Death ⟷ The time between two consecutive deaths is also exponentially distributed.

Assumption 3: Each transition from state n is of the type $n \to (n+1)$ (a single birth) or $n \to (n-1)$ (a single death).

Chapter 2

Queueing system

Queueing theory is a powerful tool for modeling and analyzing systems where waiting phenomena occur. Originating from Erlang's work on telephone networks in Copenhagen in the early 20th century, it has developed further thanks to contributions from mathematicians such as Khintchine, Palm, Kendall, Pollaczek, and Kolmogorov. This field studies arrival flows, priority rules, and execution time modeling, with applications in areas such as air traffic management, service counters, and scheduling of computing tasks.

The primary goal of queueing theory is to optimize resource management by evaluating performance indicators such as the number of customers in the system and the average time a customer spends, which is broken down into waiting time and service time. This helps determine the optimal number of servers or anticipate the impact of operational changes.

This chapter introduces the fundamental concepts of queueing theory, including Kendall's notation and Little's formula, before exploring several classical Markovian models such as M/M/1, M/M/1/N, M/M/c, and M/M/∞, along with performance evaluation methods.

Classification of Queueing Systems: [23]
To describe a queueing system, the following elements must be specified:

1. The nature of the arrival process, defined by the distribution of inter-arrival times.

2. The distribution of the random service time.

3. The number s of service stations.

4. The system capacity N. If $N < \infty$, the queue cannot exceed a length of $N - s$ units. In this case, some arriving customers may not be able to enter the system.

Terminology and Notations: [1]
In relation to the exponential distribution:

- λ: Arrival rate; the average number of arrivals per unit of time.
- $\frac{1}{\lambda}$: The average inter-arrival time.
- μ: Service rate; the average number of customers served per unit of time.
- $\frac{1}{\mu}$: The average service time of a customer in the system.

The analysis of a queueing system depends on the initial state and elapsed time. This represents the transient state, where the study is quite complex. In queueing theory, the analysis is conducted once the system reaches a steady state, where the system states are essentially independent of the initial state and elapsed time. It is assumed that the system has been in operation for a long time.

In a steady-state system, we define: [22]

- P_n: Probability of having n customers in the system.
- L_s: The average number of customers in the system.
- L_q: The average number of customers in the queue.
- W_s: The average time spent in the system (waiting + service).
- W_q: The average waiting time of a customer in the queue.
- c: The number of servers.

2.1 Simple Queue

A queueing system is characterized by a waiting area that contains one or more slots and a service area composed of one or more servers. Customers arrive randomly from outside, wait for an available server, receive service, and then leave the system.

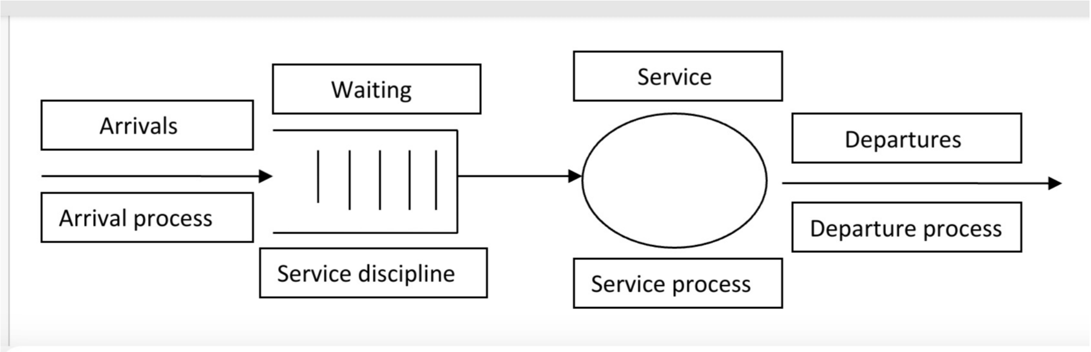


Figure 2.1: Schematic representation of a simple queue.

To specify a queueing system, we consider three main elements:

2.1.1 Arrival Process:

Customer arrivals to the system are described using a stochastic counting process $(N_t)_{t\geq 0}$. If A_n denotes the random variable representing the arrival time of the n^{th} customer, we have: $A_0 = 0$ and $A_n = \inf\{t \geq 0;\ N_t = n\}$.
If T_n denotes the random variable measuring the time between the arrival of the $(n-1)^{\text{th}}$ and the n^{th} customer [19], then:

$$T_n = A_n - A_{n-1}$$

.

2.1.2 Service Process:

First, consider a single-server queue.
Let D_n be the random variable representing the departure time of the n^{th} customer and Y_n the service time of the n^{th} customer (the time between the start and end of service). A departure always corresponds to the end of service but not necessarily to the start of a new service [18]. It is possible for a customer to leave the system, leaving the server idle until the next customer arrives.
Let μ be the service rate:

$$\frac{1}{\mu} \quad \text{is the average service duration.}$$

2.1.3 Queue Structure:

2.1.3.1 Number of Servers:

A service station may have multiple servers in parallel. Let c denote the number of servers. When a customer arrives at the station, either a server is available and the customer is immediately served, or all servers are busy, and the customer joins the queue until a server is freed.
It is generally assumed that servers are identical and operate independently. A special case is the *IS* (infinite servers) station, where the number of servers is infinite, meaning there is no waiting queue.

2.1.3.2 Queue Capacity:

The capacity of a queue to accommodate customers waiting for service can be finite or infinite. Let N be the queue capacity; an unlimited capacity queue satisfies $N = +\infty$.

2.2 Kendall Notation

The following notation, called Kendall's notation [25], is widely used to classify different Queueing systems:

$$A/B/C/K/m/Z$$

where

1. A: indicates the arrival process of customers. The used codes are:
   - M (Markov): Interarrival times of customers are independently and identically distributed according to an exponential distribution. It corresponds to a Poisson point process (memoryless property).
   - D (Deterministic distribution): Interarrival times or service times of customers are constant and always the same.
   - GI (General Independent): Interarrival times follow a general distribution (no assumption on the distribution, but interarrival times are independent and identically distributed).

- ○ G (General): Interarrival times follow a general distribution and may be dependent.
- ○ E_k: This symbol denotes a process where the time intervals between two successive arrivals are independent and identically distributed random variables following an Erlang distribution of order k.

2. B: describes the service time distribution of a customer. The codes are the same as for A.

3. C: number of servers, which is a positive integer.

4. K: queue capacity, the number of places in the system, in other words, the maximum number of customers in the system, including those in service.

5. m: user population.

6. Z: service discipline, which defines how customers are ordered for service. The used codes are:

   - ○ FIFO (First In, First Out) or FCFS (First Come, First Served): The standard queue where customers are served in their order of arrival. Note that FIFO and FCFS are not equivalent in multi-server queues. In FIFO, the first customer to arrive is the first to leave the queue, whereas in FCFS, they are the first to start service, but another customer starting later in a different server might finish earlier.
   - ○ LIFO (Last In, First Out) or LCFS (Last Come, First Served): Corresponds to a stack, where the last customer to arrive (placed on top of the stack) is the first to be served (removed from the stack). Again, LIFO and LCFS are only equivalent in a single-server queue.
   - ○ SIRO (Served In Random Order): Customers are served randomly.
   - ○ PNPN (Priority Service): Customers are served according to their priority. Higher-priority customers are served first, followed by lower-priority ones.
   - ○ PS (Processor Sharing): Customers are served equally. The system's capacity is shared among customers.

Remark 2.2.1. *In its short version, only the first three symbols $A/B/C$ are used. In such cases, it is assumed that the queue follows a FIFO discipline, and both the waiting space and the number of customers in the system are unlimited.*

2.3 Little's Law

Little's law is a very general relation that applies to a wide class of systems. It concerns only the steady-state regime of the system. No assumptions are made about the random variables characterizing the system (interarrival times, service times, etc.). The only condition for applying Little's law is that the system is stable. The system's throughput is then either the input or output rate. Little's law is expressed in the following theorem 2.3.1:

Theorem 2.3.1. *(Little's Formula): The average number of customers L, the average time spent in the system W, and the average throughput of a stable system d in steady-state satisfy the relation:*

$$L = W \times d$$

where d is the system's arrival rate ($d = \lambda$ for an $M/M/1$ queue).
Little's law indicates that there is a relationship between the average number of customers in the queue (waiting or in service) and the total average time a customer spends in the queue (waiting time + service time).

Little's law can also be applied by considering only the waiting time in the queue (excluding service). It then relates the average number of waiting customers L_q to the average waiting time of a customer before service W_q by the relation:

$$L_q = W_q \times d$$

Finally, Little's law can be applied by considering only the server. In this case, it relates the average number of customers in service L_s to the average time a customer spends in the server, which is simply the average service time $\frac{1}{\mu}$, by the relation:

$$L_s = \frac{1}{\mu} \times d$$

Three relations have been derived by applying Little's law successively to the entire system, the queue alone, and the server alone. These three relations are not independent. One can deduce one from the others by noting:

$$W = W_q + \frac{1}{\mu} \qquad \textit{and} \qquad L = L_q + L_s$$

Remark 2.3.1. *Little's law applies to all Queueing models encountered in practice (not just the $M/M/1$ queue).*

2.4 Performance Measurement of a Queue System

The study of a queue or a network of queues aims to evaluate the performance of a system under given operating conditions. In general, a queue is considered stable if the average number of customer arrivals per unit of time, denoted by λ, is less than the average number of customers that can be served per unit of time.

If each server processes on average μ customers per unit of time and the system has c servers, then the queue is stable if and only if:

$$\lambda < c\mu \quad \Leftrightarrow \quad \rho = \frac{\lambda}{c\mu} < 1 \tag{2.1}$$

where ρ is called the traffic intensity.

2.5 Some models of queues

2.5.1 The $M/M/1$ Queue

2.5.1.1 Model description:

The M/M/1 queue is a model characterized by arrivals following a Poisson process with rate λ, service times that are exponentially distributed with parameter μ, and a single server.These service times are also assumed to be independent[20].

Customers arrive at the station according to a Poisson process with rate λ. If the server is idle, the customer is served immediately; otherwise, they join a queue with unlimited capacity, following a FIFO (First In, First Out) discipline.

The queue can be considered as a birth-and-death process with the following transition rates:

$$\lambda_n = \lambda, \quad \forall n \geq 0$$

$$\mu_n = \begin{cases} \mu, & \forall n \geq 1 \\ 0, & \text{if } n = 0 \end{cases}$$

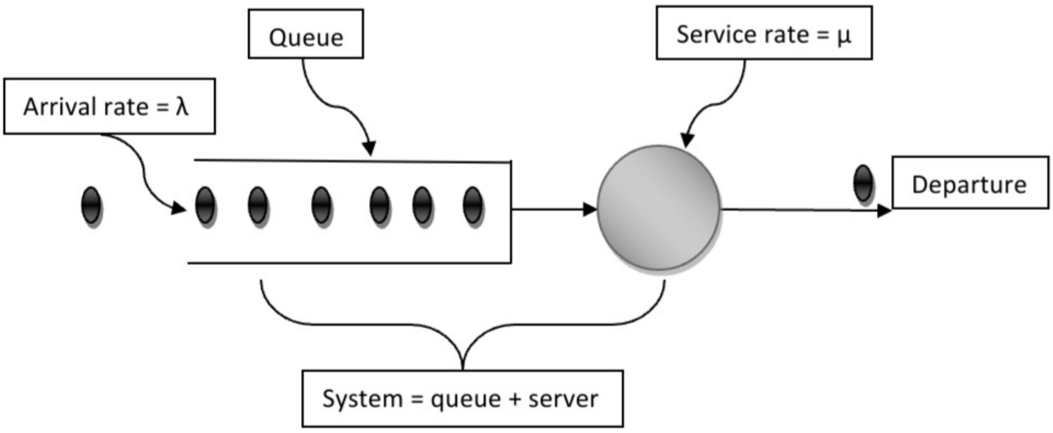


Figure 2.2: The $M/M/1$ Queue

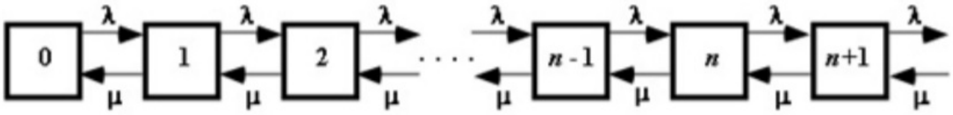


Figure 2.3: Transition diagram of the $M/M/1$ queue.

The state probabilities $p_n(t) = P[N(t) = n]$ can be computed using the following Kolmogorov differential equations, given the initial conditions of the process:

$$p'_n(t) = -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t)$$

$$p'_0(t) = \lambda p_0(t) + \mu p_1(t)$$

Under the assumption that $\lambda < \mu$ (i.e., the arrival rate is lower than the service rate), we define:

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

The steady-state probabilities for the system are given by:

$$p_n = p_0 \rho^n$$

where

$$p_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1 - \rho$$

Thus, the steady-state probability distribution is:

$$p_n = (1-\rho)\rho^n, \quad \forall n \in \mathbb{N}$$

The sequence $p = \{p_n\}_{n \geq 0}$ is called the stationary distribution and follows a geometric law (i.e., the stationary probability of having n customers in the system).

2.5.1.2 System Characteristics:

Expected number of customers in the system:

$$L = \sum_{n \geq 0} n p_n = (1-\rho) \sum_{n \geq 0} n \rho^n$$

which gives:

$$L = \frac{\rho}{1-\rho}$$

Expected number of customers being served:

$$L_S = 1 - p_0 = \rho$$

Expected number of customers in the queue:

$$L_Q = \sum_{n \geq 1} (n-1) p_n = \frac{\rho^2}{1-\rho}$$

Mean residence time W:

The mean residence time can be computed using Little's Law, where the system throughput (denoted d) corresponds to the probability that the system is not empty, multiplied by the service rate μ. Specifically,

$$d = \mathbb{P}(\text{system not empty}) \cdot \mu = (1 - p_0)\mu = \lambda.$$

In the M/M/1 queue, this leads to $d = \lambda$, since the arrival rate $\lambda_n = \lambda$ is constant .

Therefore, Little's Law gives:

$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}$$

Mean service time:

$$W_S = \frac{1}{\mu}$$

Mean waiting time:

$$W_Q = W - W_S = \frac{\lambda}{\mu(\mu - \lambda)}$$

2.5.2 The $M/M/1/K$ Queue

2.5.2.1 Model description:

Let K be the queue capacity, which represents the maximum number of customers that can be present in the system, either waiting or being served. When a new customer arrives while there are already K customers in the system, they are lost. This system is known as the M/M/1/K queue. The state space E is now finite: $E = \{0, 1, 2, \ldots, K\}$ .Since the queue capacity is limited, even if customers arrive on average much faster than the server can process them, any arriving customer is rejected when the queue is full. Therefore, the number of customers in the system can never tend to infinity. Moreover, once a customer is allowed to enter the system, they will eventually leave, as their residence time corresponds to the total service time of all the customers ahead of them, which is bounded by K. Over a long period, the output rate will equal the input rate, ensuring the unconditional stability of the system.

The birth-and-death process modeling this type of queue is defined as:

$$\lambda_n = \begin{cases} \lambda, & \text{if } n < K \\ 0, & \text{if } n \geq K \end{cases}$$

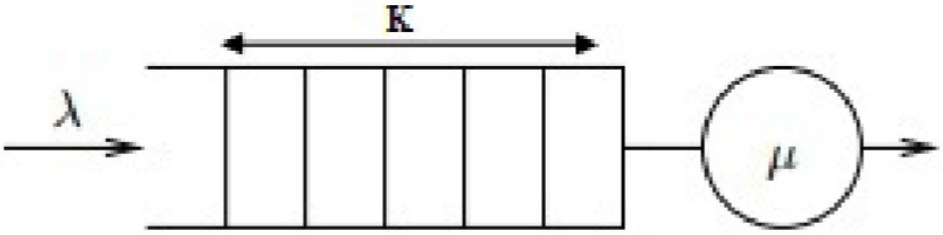


Figure 2.4: The $M/M/1/K$ Queue

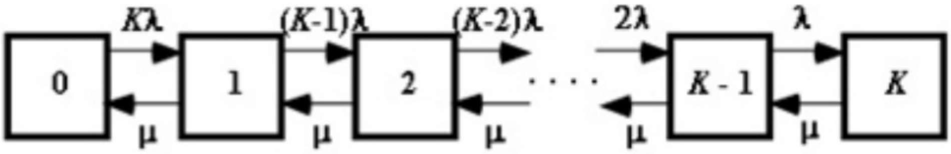


Figure 2.5: Evaluation of the state in the $M/M/1/K$ queue.

The integration of the recurrence equation to compute P_n is given by:

$$p_n = p_0 \rho^n, \quad \text{for } n \leq K$$

$$p_n = 0, \quad \text{for } n > K$$

To determine p_0, we use the normalization condition:

$$\sum_{n=0}^{K} p_n = 1$$

Given that $p_n = p_0 \rho^n$ and $\rho = \frac{\lambda}{\mu}$, the sum forms a geometric series.
-If $\lambda \neq \mu$ ($\rho \neq 1$):

$$p_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

-If $\lambda = \mu$ ($\rho = 1$):

$$p_0 = \frac{1}{K + 1}$$

Thus,

$$p_0 = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}, & \text{if } \lambda \neq \mu; \\ \frac{1}{K+1}, & \text{if } \lambda = \mu. \end{cases}$$

2.5.2.2 System Characteristics:

Average Number of Customers in the System:

$$L = \sum_{n=0}^{K} np_n = \frac{\rho}{1-\rho} \cdot \frac{1-(K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}}$$

When K tends to infinity and $\rho < 1$, this corresponds to the M/M/1 queue:

$$L = \frac{\rho}{1-\rho}$$

Average Number of Customers in the Queue:

$$L_Q = \sum_{n=1}^{K} (n-1)p_n = L - (1 - p_0)$$

Using Little's Law, we can compute the mean time a customer spends in the system, denoted by W, and the mean waiting time in the queue, denoted by W_Q.

Moreover, the system throughput, d, can be calculated in two equivalent ways: either by evaluating the departure rate of customers from the server, d_d, or by computing the effective arrival rate of customers accepted into the system, d_a.

The output rate from the system equals the service rate μ, weighted by the probability that the system is not empty. Thus:

$$d_d = \mathbb{P}(\text{system not empty}) \cdot \mu = \sum_{n=1}^{K} p_n\mu = (1-p_0)\mu = \frac{\rho - \rho^{K+1}}{1-\rho^{K+1}}\mu.$$

The input rate into the queue equals the arrival rate λ, weighted by the probability that the system is not full upon arrival. Therefore:

$$d_a = \mathbb{P}(\text{system not full}) \cdot \lambda = \sum_{n=0}^{K-1} p_n\lambda = (1-p_K)\lambda = \frac{1-\rho^K}{1-\rho^{K+1}}\lambda.$$

Since $\rho = \frac{\lambda}{\mu}$, it follows that $d_d = d_a = d$. .

Mean Time a Customer Spends in the System:

$$W = \frac{L}{d}$$

Mean Waiting Time in the Queue:

$$W_Q = \frac{L_Q}{d}$$

2.5.3 The $M/M/C$ Queue

2.5.3.1 Model description:

We consider a system identical to the $M/M/1$ queue except that it has C identical and independent servers. The assumptions remain the same:

- The arrival process follows a Poisson process with rate λ.
- The service times are exponentially distributed with rate μ.

This system is known as the$M/M/C$ queue [15]. The state space E is infinite, as in the M/M/1 queue : $E = \{0, 1, 2, \dots\}$ The queue has an infinite capacity. If a server is available, an arriving customer is immediately assigned to it. Otherwise, the customer joins a single waiting line shared among all servers. When a server becomes free, the customer at the front of the queue moves to that server. Consequently, the queue discipline is FIFO (First In, First Out).

The birth-death process modeling this queueing system is defined as follows:

$$\lambda_n = \lambda$$

$$\mu_n = \begin{cases} n\mu, & \text{for } n = 1, \dots, C-1, \\ C\mu, & \text{for } n \geq C. \end{cases}$$

The stability condition for this model is:

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

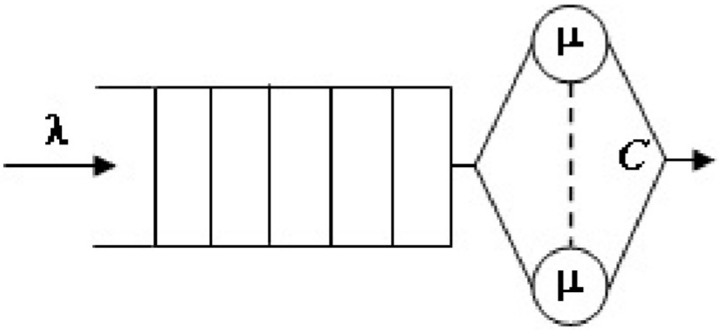


Figure 2.6: The $M/M/c$ Queue

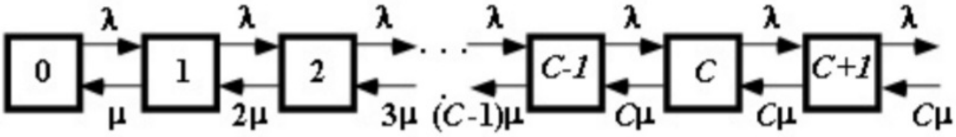


Figure 2.7: Evaluation of the state in the $M/M/c$ queue.

Based on the system diagram and steady-state analysis, using the Chapman-Kolmogorov equations, we obtain the following equations:

$$\begin{cases} \lambda p_0 = \mu p_1 \\ (\lambda + n\mu)p_n = \lambda p_{n-1} + (n+1)\mu p_{n+1}, & \text{pour } 1 \leq n < C \\ (\lambda + C\mu)p_n = \lambda p_{n-1} + C\mu p_{n+1}, & \text{pour } n \geq C \end{cases}$$

with the normalization condition:

$$\sum_{n=0}^{\infty} p_n = 1$$

Solving the above system gives the following stationary distribution:

$$P_n = \begin{cases} P_0 \dfrac{\rho^n}{n!}, & \text{pour } n = 1, 2, \ldots, c-1 \\ P_0 \dfrac{\rho^n c^{c-n}}{c!}, & \text{pour } n \geq c \end{cases}$$

With

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \cdot \frac{1}{1-\frac{\rho}{c}}\right]^{-1}$$

This distribution exists if $\lambda < C\mu$.

2.5.3.2 system characteristics:

From the stationary distribution of the process $\{N(t), t \geq 0\}$, we can compute:
Average Number of Customers in the System:

$$L = \rho + \frac{\rho^{C+1}}{C \cdot C!(1-A)^2} p_0 \quad \text{where } A = \frac{\lambda}{c\mu}$$

Average Number of Customers in the Queue:

$$L_Q = \frac{\rho^{C+1}}{C \cdot C!(1-A)^2} p_0 \quad \text{where } A = \frac{\lambda}{c\mu}$$

Mean Time a Customer Spends in the System:

When the system contains fewer than c customers, they are handled by the c servers at a rate of $n\mu$. When the system contains more than c customers, the service rate becomes constant at $c\mu$. Therefore, the throughput is given by:

$$d = \sum_{n=1}^{c-1} p_n n\mu + \sum_{n=c}^{+\infty} p_n c\mu.$$

By substituting the previously derived expressions for the probabilities p_n and p_0, we find that the queue is indeed stable, and the throughput is:

$$d = \lambda.$$

Applying Little's Law then yields:

$$W = \frac{C\mu\rho^C}{C!(C\mu - \lambda)^2} p_0$$

Mean Waiting Time:

$$W_Q = \frac{1}{\mu} + \frac{\rho^C}{\mu C \cdot C!(1-A)^2} p_0 \quad \text{where } A = \frac{\lambda}{c\mu}$$

2.5.4 The M/M/∞ Queue

2.5.4.1 Model description:

For this queue model, the system consists of an unlimited number of identical and independent servers. As soon as a customer arrives, they are immediately served (there is no waiting time).

In this system, customers arrive at times $0 < t_1 < t_2 < \dots$ following a Poisson process with rate λ, and service times are exponentially distributed with rate μ. This queue is known as the M/M/∞ system.

As for the $M/M/C$ queue, it can be easily demonstrated that the transition rate from state n to state $n-1$ is equal to $n\mu$, which corresponds to the departure rate of one of the n customers being served[14]. Similarly, the transition rate from state n to state $n+1$ is equal to λ, corresponding to the arrival of a customer. Thus, this system follows a birth-and-death process with: $\lambda_k = \lambda$ and $\mu_k = k\mu$ for $k = 0, 1, 2, \dots$.

Let p_n denote the steady-state probability of being in state n. The balance equations give us:

$$p_{n-1}\lambda = p_n n\mu, \quad \text{for } n = 1, 2, \dots$$

So:

$$p_n = \frac{\rho^n}{n!} p_0, \quad \text{for } n = 1, 2, \dots$$

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

Steady-State Probability Distribution:

The normalization condition is:

$$\sum_{n=0}^{\infty} p_n = 1$$

which leads to:

$$p_0 = \left[\sum_{n=0}^{\infty} \frac{\rho^n}{n!}\right]^{-1} = e^{-\rho}$$

Thus, the steady-state probabilities are:

$$p_n = \frac{\rho^n}{n!}e^{-\rho}, \quad \text{for } n = 1, 2, \ldots$$

Since the series $\sum_{n=0}^{\infty} \frac{\rho^n}{n!}$ converges for all values of ρ (and thus for all values of λ and μ), the system is unconditionally stable.

2.5.4.2 System Characteristics:

Average Number of Customers in the System:

The expected number of customers in the system is given by:

$$L = \sum_{n=1}^{\infty} np_n$$

Substituting p_n,

$$L = e^{-\rho} \sum_{n=1}^{\infty} \frac{n\rho^n}{n!}$$

Using the identity:

$$\sum_{n=1}^{\infty} \frac{n\rho^n}{n!} = \rho e^{\rho}$$

we obtain:

$$L = e^{-\rho} \cdot \rho e^{\rho} = \rho$$

Average residence time W:

Using Little's Law where:

$$d = \sum_{n=1}^{+\infty} p_n n\mu = e^{-\rho} \sum_{n=1}^{+\infty} \mu \frac{\rho^n}{(n-1)!} = e^{-\rho} \rho e^{\rho} \mu = \rho\mu = \lambda$$

because the service is performed at a rate of $n\mu$ in each state where the system contains n customers.

Thene:

$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda} = \frac{1}{\mu}$$

Chapter 3

Modeling and Analysis of Customer Impatience in a Multi-Server Queueing System

In a multi-server Queueing system, customer impatience significantly influences the overall performance of the system. When a customer is forced to wait too long to be served, they may choose to leave the queue without receiving service, resulting in losses for the company or service involved. This phenomenon, known as *reneging*, as well as *balking* (refusal to join the queue due to congestion), directly impact the system's efficiency.

This chapter provides an in-depth analysis of *balking* and *reneging*, highlighting their theoretical foundations and historical development. We also examine the interaction of these phenomena and the *balking* functions, which model the probability of a customer joining or not joining the system based on its state. The goal of this study is to offer a comprehensive understanding of customer impatience dynamics and provide analytical tools for optimizing the performance of multi-server Queueing systems. Finally, we analyze a Markovian multi-server Queueing system incorporating *balking* and *reneging*.

3.1 Balking

Definition 3.1.1. *Balking is a phenomenon observed in queueing systems, where a customer arrives at a queueing system and decides not to join the queue due to the current queue length or the perceived waiting time being too long. In other words, balking occurs when customers decide not to enter the queue upon arrival because it is too long.[27]*

3.1.1 The history

The concept of balking in queueing theory dates back to the pioneering work of**Agner Krarup Erlang** in the early 20th century when he studied telephone systems in Denmark. Erlang developed mathematical models to analyze congestion in telephone exchanges, laying the foundations of queueing theory, although he did not formally define the concept of balking [7].

Later, David G. Kendall introduced a standardized notation to classify queueing systems in his paper **Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Kinetic Equations** [18]. In this context, balking was formalized as a behavior in which a potential customer decides not to enter the queue after observing its length.

Between the 1960 and 1980, researchers such as **William Feller** further analyzed the mathematical aspects of balking, particularly through the study of Markov processes and differential equations modeling queue dynamics[8]. Additionally, models such as M/M/1/K, which account for a maximum capacity K, allowed for the study of how the probability of balking varies with queue size[4].

In the 2000, balking was applied to more complex domains such as online services, hospital systems, and traffic management **(Taha, 2006)**.

More recently, an extension of the classical concept, known as reverse balking, was introduced by **Jain et al. (2014)**. Unlike traditional balking, where customers avoid a long queue, reverse balking states that the longer the queue, the higher the probability of joining the system. This phenomenon is observed in contexts where a long queue is perceived as a sign of quality or high demand [17].

3.1.2 Modeling the Probability of Balking in Queueing Systems

3.1.2.1 Linear Model

The probability of balking follows a linear function in relation to the number of customers present in the system[14]:

$$p(n) = \frac{n}{N-1}$$

Where n is the number of customers already present in the system at the moment a new customer arrives,

and N is the maximum number of customers that the arriving customer is willing to tolerate in the system.

This model was proposed by A.K. Erlang in the early 20th century as part of the first works on queueing systems.
It is used when the probability of balking increases progressively with the system's load, but in a linear fashion.

3.1.2.2 Exponential Model

The probability of balking depends exponentially on the number of available spaces left[9]:

$$p(n) = e^{-\alpha(N-n)}$$

Where the parameter α: controls the growth rate of the exponential increase. This model comes from the work of David Kendall in the 1950-1960 on Markov processes and queueing systems.
It applies when the probability of balking increases slowly at first, then accelerates as the system approaches its maximum capacity.

3.1.2.3 Logistical Model (S-curve)

The probability of balking follows an S-shaped curve:

$$p(n) = \frac{1}{1 + e^{-\beta(n-n_0)}}$$

Where parameter β: determines the steepness of the transition in the logistic model.

Popularized by Pierre-FranÃ§ois Verhulst in the 19th century, this model stems from dynamic system theory.
It is ideal for systems where customers react progressively, but with a sharp change once a critical threshold is reached.

3.1.2.4 Power Model

The probability of balking follows a power function:

$$p(n) = \left(\frac{n}{N}\right)^k$$

Where parameter k:

- If $k > 1 \rightarrow$ slow increase at the beginning, then faster.

- If $k < 1 \rightarrow$ fast increase at the beginning, then slower.

Inspired by power laws observed in statistical physics and queueing theory in the $1970 - 1980$.
It is suited when the rate of balking varies non-linearly, following a more complex empirical rule.

3.1.2.5 Threshold Model

The model defines a threshold beyond which the probability of balking becomes immediate[11]:

$$p(n) = \begin{cases} 0 & \text{if } n < n_0 \\ 1 & \text{if } n \geq n_0 \end{cases}$$

This model has been used since the $1950 - 1960$ in the first studies on managing systems with a fixed capacity.
It is applied in situations where customers are willing to wait until a specific threshold before deciding to reject entry.

3.1.3 Reverse balking

Definition 3.1.2. *is a behavior observed in queueing systems where customers hesitate to join the queue when it is empty or nearly empty, but are more likely to join when there are already several people waiting. This may happen, for example, when customers perceive an empty queue as a sign of poor service quality or uncertainty.*

3.2 Reneging

Definition 3.2.1. *After spending some time in the queue, the customer decides to leave the system without being served.*

3.2.1 The history

The study of *reneging* has evolved over the decades with significant contributions from various researchers.
Barrer (1957)[4] was one of the first to explore reneging models using Markovian arrivals and services, where customers could leave the queue if they became impatient before being served.

This approach was extended by **Haight (1959)**[16], who integrated the reneging phenomenon into an *M/M/1* model, considering the abandonment of customers after a certain waiting time.

Later works by **Ghosal (1963)**[13], **Gavish and Schweitzer (1977)**[12], and **Liu ,al. (1987)**[21] further examined reneging in multi-server systems, where a fixed delay before service initiation could influence the abandonment decision.

Other research, such as that of **Bae et al. (2001)**[3] and **Choi et al. (2001)**[5], modeled queueing systems with impatient customers and developed models to measure the system's performance under such conditions.

In subsequent years, more recent research continued to enrich the understanding of *reneging*. **O. Garnett et al. (2002)** [10]studied the behavior of impatient customers in a *call center*, highlighting the impact of abandonment on system performance.

S. Zeltyn and A. Mandelbaum (2005)[30] proposed an $M/M/n + G$ model, characterizing a multi-server system with Poisson arrivals, exponential services, and general waiting times.

W. Whitt (2006) [28]further advanced this analysis by studying fluid models in multi-server queueing systems with abandonment, providing a better understanding of abandonment dynamics in complex environments.

More recently, **H. Shuangchi and J. G. Dai (2011)**[26] examined multi-server queues with abandonment, and **T. Andrey (2013)**[2] analyzed systems consisting of multiple queues and heterogeneous servers, where service times do not necessarily follow an exponential distribution.

These works have contributed to a deeper understanding of the effects of reneging on queueing systems and have aided in optimizing performance in various sectors, including call centers, online services, and computer networks.

3.3 Modeling Customer Impatience in a Multi-Server Queue

A general model is developed to incorporate different forms of customer impatience in a multi-server Queueing system. The model considers *balking*, where customers refuse to enter if the queue is too long, *reneging*, which occurs when customers abandon the queue after excessive waiting, and *reverse balking*, a concept introduced by Jain and al. (2014), where the probability of joining the queue depends on its current size. The impact of these behaviors on system dynamics and performance is analyzed .

3.3.1 System Description

1. The arrival process is Poisson with parameter λ.
2. There are multiple servers, say c. The service times follow exponential distribution with parameter μ such that:

$$\mu_n = n\mu \quad \text{when} \quad n < c \quad \text{and} \quad \mu_n = c\mu \quad \text{when} \quad n \geq c.$$

3. The system capacity is finite, say N.
4. The queue discipline is First-Come, First-Served (FCFS).
5. (a) When the system is not empty, customers balk with probability ($q' = \theta_0 = \theta_1 = ... = \theta_{c-1} < \theta_c < ... < \theta_n < ... < \theta_N = 1$) and do not balk with probability $\bar{\theta}_n = 1 - \theta_n$ and

   (b) When the system is empty, customers balk with probability q' and may not balk with probability $p' (= 1 - q')$.

   The balking described in (a) and (b) is called reverse balking.
6. Each customer, upon joining the queue, waits for some time for their service to begin. If they do not receive service by then, they leave the queue without receiving service (i.e., reneging). The reneging times follow an exponential distribution with parameter ξ.

3.3.2 Stochastic Model Formulation

Let $P_n(t)$ denote the probability that there are n customers in the system at time t. The Chapman-Kolmogorov equations governing the system are as follows:

For $n = 0$ (empty system):

$$\frac{dP_0(t)}{dt} = -\lambda p' P_0(t) + \mu P_1(t) \tag{1}$$

For $1 \leq n < c$ (more than one customer but fewer than c):

$$\frac{dP_n(t)}{dt} = \lambda p' P_{n-1}(t) - (\lambda p' + n\mu) P_n(t) + (n+1)\mu P_{n+1}(t) \tag{2}$$

For $c \leq n \leq N-1$ (from c customers up to $N-1$ customers):

$$\frac{dP_n(t)}{dt} = \bar{\theta}_{n-1}\lambda P_{n-1}(t) - (\bar{\theta}_n\lambda + c\mu + (n-c)\xi)P_n(t) + [c\mu + (n+1-c)\xi]\,P_{n+1}(t) \quad (3)$$

For $n = N$ (full system):

$$\frac{dP_N(t)}{dt} = \bar{\theta}_{N-1}\lambda P_{N-1}(t) - (c\mu + (N-c)\xi)\,P_N(t) \quad (4)$$

Where:

- λ is the arrival rate,
- μ is the service rate,
- c is the number of servers,
- N is the system capacity,
- ξ is the reneging rate,
- p' is the probability that a customer does not balk when the system is empty,
- q' is the probability that a customer balks when the system is empty,
- θ_n is the probability that a customer balks when the system is not empty.

3.3.3 Steady-State Solution

In steady state, we assume:

$$\lim_{t\to\infty} P_n(t) = P_n, \quad P'_n(t) = 0.$$

Thus, the equations become:

$$0 = -\lambda p' P_0 + \mu P_1, \quad n = 0 \quad (5)$$

$$0 = p'\lambda P_{n-1} - (p'\lambda + n\mu)\,P_n + (n+1)\mu P_{n+1}, \quad 1 \leq n < c \quad (6)$$

$$\begin{aligned} 0 = {} & \bar{\theta}_{n-1}\lambda P_{n-1} - \left(\bar{\theta}_n\lambda + c\mu + (n-c)\xi\right) P_n \\ & + [c\mu + (n+1-c)\xi]\,P_{n+1}, \quad c \leq n \leq N-1 \end{aligned} \quad (7)$$

$$0 = \bar{\theta}_{N-1}\lambda P_{N-1} - [c\mu + (N-c)\xi]\,P_N, \quad n = N \quad (8)$$

Solving (5) – (8), we obtain:

$$P_n = \begin{cases} \frac{1}{n!}\left(\frac{p'\lambda}{\mu}\right)^n P_0, & 1 \leq n < c+1 \\ \frac{\lambda^n \prod_{i=c}^{n-1} \bar{\theta}_i}{\prod_{i=1}^{n-c}(c\mu + i\xi)} \frac{1}{c!}\left(\frac{p'\lambda}{\mu}\right)^c P_0, & c+1 \leq n \leq N \end{cases}$$

Using the normalization condition:

$$\sum_{n=0}^{N} P_n = 1,$$

We get :

$$P_0 + \sum_{n=1}^{c} P_n + \sum_{n=c+1}^{N} P_n = 1. \tag{9}$$

Thus,we obtain:

$$P_0 = \left\{1 + \sum_{n=1}^{c} \frac{1}{n!}\left(\frac{p'\lambda}{\mu}\right)^n + \sum_{n=c+1}^{N} \frac{\lambda^n \prod_{i=c}^{n-1} \bar{\theta}_i}{\prod_{i=1}^{n-c}(c\mu + i\xi)} \frac{1}{c!}\left(\frac{p'\lambda}{\mu}\right)^c \right\}^{-1}.$$

3.3.4 Measures of Performance

3.3.4.1 Expected System Size

The expected system size L_s is given by:

$$L_s = \sum_{n=1}^{N} nP_n \tag{10}$$

which can be rewritten as:

$$L_s = \sum_{n=1}^{c} nP_n + \sum_{n=c+1}^{N} nP_n \tag{11}$$

Substituting the expressions for P_n, we get:

$$L_s = P_0\left(\sum_{n=1}^{c} \frac{n(p'\lambda/\mu)^n}{n!} + \frac{(p'\lambda/\mu)^c}{c!} \sum_{n=c+1}^{N} \frac{n\lambda^n \prod_{i=c}^{n-1} \bar{\theta}_i}{\prod_{i=1}^{n-c}(c\mu + i\xi)}\right) \tag{12}$$

3.3.4.2 Average Rate of Reneging

The average rate of reneging R_r is given by:

$$R_r = \sum_{n=c}^{N} (n-c)\xi P_n \tag{13}$$

Substituting the expressions for P_n, we get:

$$Rr = \xi P_0 \cdot \frac{(p'\lambda/\mu)^c}{c!} \sum_{n=c}^{N} (n-c) \cdot \frac{\lambda^n \prod_{i=c}^{n-1} \bar{\theta}_i}{\prod_{i=1}^{n-c} (c\mu + i\xi)} \tag{14}$$

3.3.4.3 Average Rate of Reverse Balking

$$R_b = \sum_{n=0}^{c-1} q'\lambda P_n. \tag{15}$$

$$R_b = p_0 \left(q'\lambda + q'\lambda \frac{1}{n!} \left(\frac{p'\lambda}{\mu} \right)^n \right) \tag{16}$$

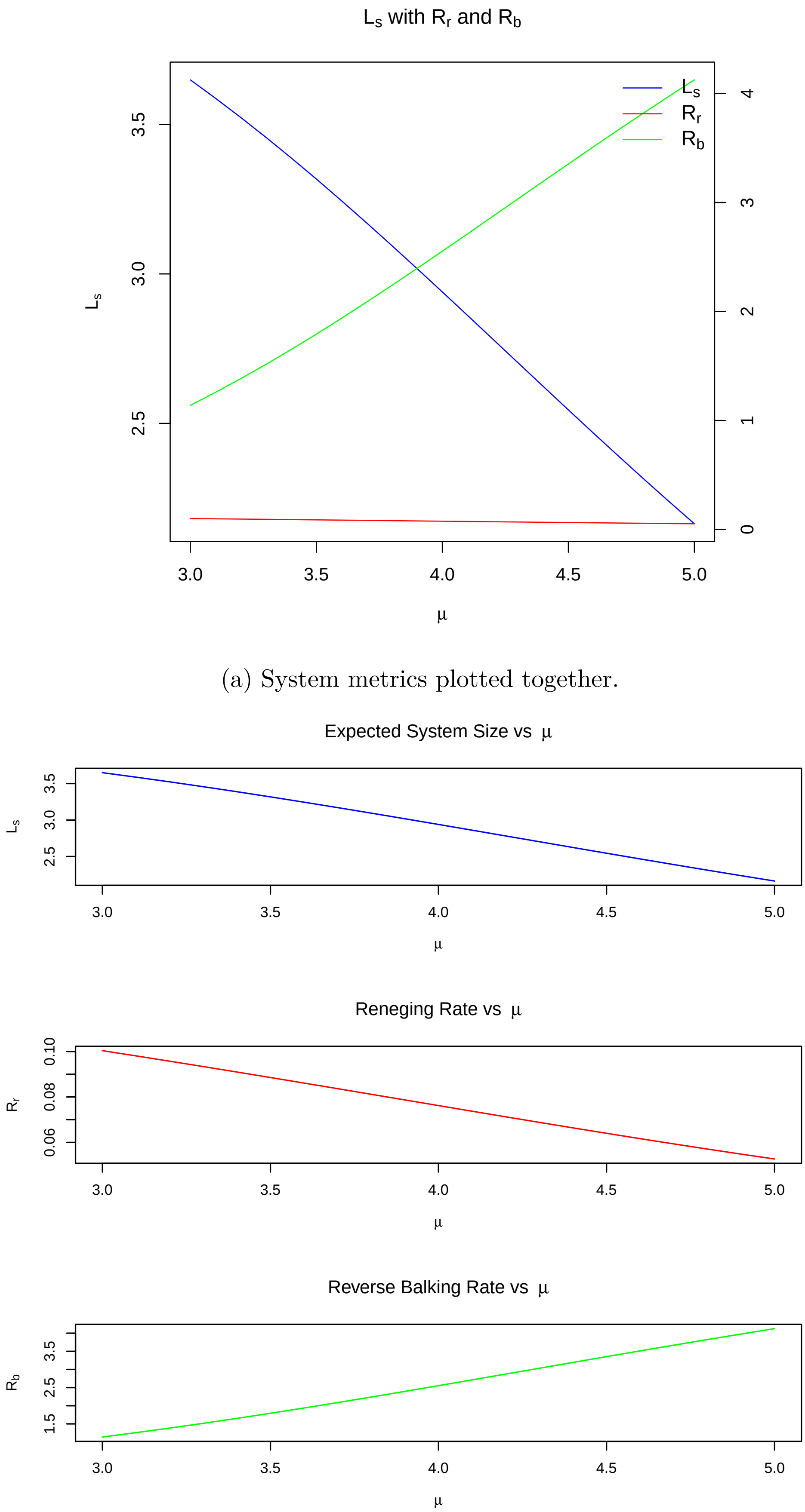
3.3.5 Model sensitivity analysis

3.3.5.1 Impact of Service Rate on System Metrics

This section presents a numerical illustration accompanied by a sensitivity analysis, with results obtained using the R software. The study focuses on the variation of the average system size (L_s), the average reneging rate (R_r), and the average rate of reverse balking (R_b) as a function of the parameter μ. The simulations are conducted considering the following values: $\lambda = 10$, $\xi = 0.1$, $q' = 0.8$, $c = 3$, and $N = 10$.

μ	Expected System Size	Average Rate of Reneging	Average Rate of Reverse Balking
3.0	5.1165	0.2157	0.1284
3.1	5.0570	0.2103	0.1487
3.2	4.9996	0.2053	0.1711
3.3	4.9439	0.2006	0.1958
3.4	4.8899	0.1960	0.2229
3.5	4.8370	0.1917	0.2526
3.6	4.7850	0.1876	0.2849
3.7	4.7338	0.1836	0.3199
3.8	4.6831	0.1798	0.3578
3.9	4.6327	0.1761	0.3986
4.0	4.5824	0.1725	0.4424
4.1	4.5321	0.1690	0.4892
4.2	4.4816	0.1656	0.5393
4.3	4.4308	0.1623	0.5925
4.4	4.3796	0.1591	0.6489
4.5	4.3279	0.1559	0.7087
4.6	4.2755	0.1528	0.7716
4.7	4.2225	0.1497	0.8379
4.8	4.1688	0.1467	0.9074
4.9	4.1143	0.1437	0.9802
5.0	4.0590	0.1408	1.0561

Table 3.1: Variation of System Metrics with Service Rate



(a) System metrics plotted together.

(b) System metrics plotted separately.

Figure 3.1: Variation of system metrics with service rate μ.

Interpretation of the results.

As the service rate μ increases, the system becomes more efficient: customers are served faster, and the waiting queue gradually empties. Consequently, the number of reneging (Rr) decreases, reflecting their reduced waiting time. However, the reverse balking rate (Rb) increases as μ increases. This occurs because, with faster service, the system is more frequently empty or nearly empty. In the context of reverse balking, customers are more hesitant to join when there are few or no waiting customers. Thus, this phenomenon becomes more pronounced, and the reverse balking rate increases due to the growing likelihood of encountering an empty or nearly empty system.

3.3.5.2 Impact of Arrival Rate on System Metrics

This section presents a numerical illustration accompanied by a sensitivity analysis, with results obtained using the R software. The study focuses on the variation of the average system size (L_s), the average reneging rate (R_r), and the average rate of reverse balking (R_b) as a function of the arrival rate (λ). The simulations are conducted considering the following values: $\mu = 3$, $\xi = 0.1$, $q' = 0.8$, $c = 3$, and $N = 10$.

λ	**Expected System Size**	**Average Rate of Reneging**	**Average Rate of reverse Balking**
5	1.278720	0.03258631	3.040312
6	2.544159	0.07834358	2.250466
7	3.630757	0.12235664	1.277883
8	4.261916	0.15227570	0.652815
9	4.607954	0.17245303	0.332896
10	4.825946	0.18806809	0.175840
11	4.990355	0.20174911	0.096927
12	5.132025	0.21461754	0.055640
13	5.263559	0.22712759	0.033114
14	5.390127	0.23945030	0.020339
15	5.513783	0.25163508	0.012841
16	5.635235	0.26367882	0.008304
17	5.754603	0.27555677	0.005486
18	5.871755	0.28723687	0.003693
19	5.986469	0.29868671	0.002528
20	6.098505	0.30987684	0.001757
21	6.207645	0.32078227	0.001238
22	6.313709	0.33138304	0.000884
23	6.416557	0.34166414	0.000638
24	6.516093	0.35161526	0.000465
25	6.612261	0.36123031	0.000343

Table 3.2: Variation of System Metrics with Mean Arrival Rate

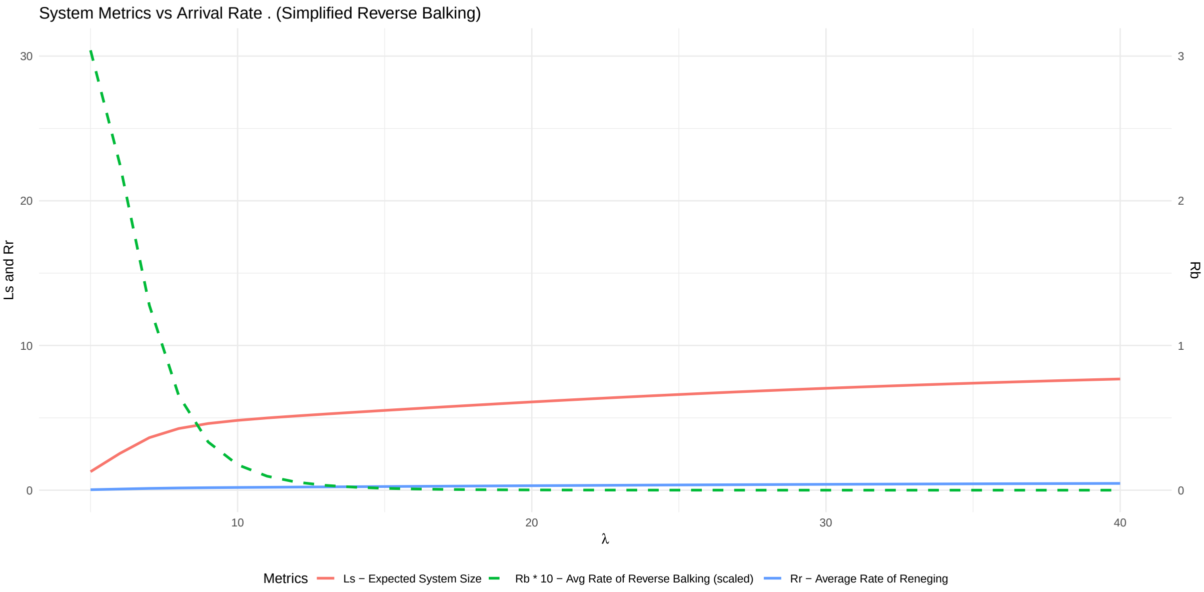


Figure 3.2: Impact of Arrival Rate

Interpretation of the results.

As the arrival rate λ increases, the expected system size L_s grows steadily, confirming that higher customer inflow intensifies system load when the service rate $\mu = 3$ remains constant. For instance, L_s rises from about 1.28 at $\lambda = 5$ to over 6.6 at $\lambda = 25$.

Similarly, the reneging rate R_r increases gradually with λ, indicating more customers abandon the system as waiting times lengthen. R_r moves from roughly 0.03 at $\lambda = 5$ to about 0.36 at $\lambda = 25$.

In contrast, the reverse balking rate R_b decreases monotonically as λ increases, from a high of approximately 3.04 at $\lambda = 5$ to a very low value near 0.00034 at $\lambda = 25$. This suggests that at low arrival rates, reverse balking is significant—possibly due to customers perceiving the system as attractive or available—but as the system saturates, the appeal diminishes sharply, discouraging new entries.

Overall, the results reveal that increasing the arrival rate pushes the system towards saturation, increasing queue lengths and abandonment rates, while simultaneously reducing reverse balking behavior. This highlights the need to either control the inflow of customers or improve service capacity to maintain system efficiency and customer satisfaction.

General Conclusion

This thesis aimed to analyze the impact of customer impatience in a multi-server Queueing system. We studied customer balking,reverse balking and reneging behaviors by modeling impatience through an exponential distribution of patience times. The results allowed us to calculate key performance indicators such as the abandonment rate, average waiting time, and the proportion of served customers. This approach provided an overview of how impatience affects the efficiency of such a system.

Furthermore, this study could be extended to include transient states or systems with infinite server capacities. Extending the model to non-Markovian queues, where customer behaviors may be more complex, could also offer new and interesting perspectives for more accurately representing real-world systems.

A more in-depth analysis, incorporating economic criteria, would help better understand the financial impact of balking and reverse balking and propose optimization strategies accordingly. Thus, this work paves the way for future research to refine queue management in various sectors.

Bibliography

[1] Andrey, T.Queueing Systems with Heterogeneous Servers and Non-Exponential Service Times.Mathematics of Operations Research, 38(4), 506-522, (2013).

[2] Anisimov, V. V., Zakusilo, O. K., and Donchenko, V. S. 1987. Elements of Queueing Theory and Asymptotic System Analysis. Vishcha Shkola, Kiev (in Russian).

[3] Bae .J, Lee .Y, and Choi, H.Queueing with Impatient Customers and Reneging.Computers & Operations Research, 28(10), 943-956, (2001).

[4] Barrer, D. E.Queueing with Reneging and Delayed Arrivals. Journal of the Operations Research Society of America, 5(1), 1-15, (1957).

[5] Choi, H., Kang, M., and Lee, K.Modeling and Performance Analysis of Queueing Systems with Reneging. European Journal of Operational Research, 135(1), 52-64, (2001).

[6] Edward P. C. Kao ,An Introduction to Stochastic Processes ,1997.

[7] Erlang, A. K. The Theory of Probabilities and Telephone Conversations. Nyt Tidsskrift for Matematik,(1909).

[8] Feller, W. An Introduction to Probability Theory and Its Applications (Vol. 1, 3rd ed.). Wiley,(1968).

[9] Frederick Solomon,Probability and Stochastic Processes„ Prentice Hall, 2007.

[10] Garnett, O, Mandelbaum, A,and Zeltyn, S.Call Centers with Impatient Customers Queueing Systems, 40(1), 31-56, (2002). 2(2012), No. 1, 1-5.

[11] Gaver, D. P. Queueing Systems with Balking and Reneging, Operations Research,1964.

[12] Gavish, B, and Schweitzer, P.Queueing with Reneging and Service Delays. Mathematics of Operations Research, 2(3), 282-292,(1977).

[13] Ghosal, S.A Study of the Queueing System with Reneging. Naval Research Logistics Quarterly, 10(2), 163-174, (1963).

[14] Gomez-Corral. A. and M.F. Ramalhoto. On the waiting time distribution and the busy period of a retrial queue with constant retrial rate. Stochastic Modelling and Applications, 3, 37-47, (2000).

[15] Gross, D.and C.M, Harris, Fundamentals of Queueing Theory, Willcy, New york, 1975 1984.

[16] Haight, F. A.Queueing with Reneging. Journal of the Operations Research Society of America, 7(3), 267-275, (1959).

[17] Jain, R., Tiwari, M. K., Gupta,D.Reverse Balking and Reneging in Queueing Systems. International Journal of Computer Applications,(2014).

[18] Kendall, D. G. Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Kinetic Equations. The Royal Society Proceedings A,1953.

[19] Khintchine, A. Y 1969. Mathematical Methods in the Theory of Queueing. Second edition, Hafner Publishing Company, New York (First edition: Griffin, London, 1960; Russian original: 1955). 859-866.

[20] Kumar, R. and Sharma, S.K. An $M/M/1/N$ Queueing Model with Retention of reneged cus- tomers and Balking, American Journal of Operational Research.

[21] Liu, Z, Yao, D, and Li, H.The Effect of Reneging on Multi-Server Queues.Operations Research Letters, 6(1), 19-22, (1987).

[22] Moulay Hachemi. (2015). Files d'attente et applications (p. 12). Cours polycopié, année universitaire 2014/2015.

[23] Newell, G. F. 1982. Applications of Queueing Theory. Second edition, Chapman and Hall, London (First edition: 1971).

[24] Rakesh Kumar and Bhupender Kumar Som,A Multi-Server Queue with Reverse Balking and Impatient Customers, School of Mathematics, Shri Mata Vaishno Devi University, Department of Management, Lloyd Business School. Pak. J. Statist, 2020.

[25] Richard Newman ,Elementary Queueing Theory Notes,University of Floride .(january 1999).

[26] Shuangchi, H. and Dai, J. G.Analysis of Multi-Server Queues with Abandonment. Operations Research Letters, 39(3), 220-229, (2011).

[27] Taha, H. A. Operations Research: An Introduction (9th ed.). Pearson.(2006).

[28] Whitt, W.Fluid Models and Large Queueing Systems. Mathematics of Operations Research, 31(2), 309-325, (2006).

[29] Zakhar Kabluchko, Stochastic Processes (Stochastik II), University of Ulm Institute of Stochastics, (2013-2014).

[30] Zeltyn, S,and Mandelbaum, A.Modeling Call Centers with Impatient Customers. Queueing Systems, 51(1), 47-76, (2005).