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

University of Saida - Dr Moulay Tahar.
Faculty of Sciences.
Department of Mathematics.

No. Assigned by the library

This Thesis was Submitted in Partial Fulfillment of the Requirements the degree of

Academic Master

Discipline : MATHEMATICS
Specialty: Statistical Stochastic Analysis of Processes and Applications

by
Djebli **F**ouzia[1]

Under the supervision of

Dr. Mokhtar **K**adi

Theme:

Working Vacation Policy in Queueing Systems

Defended on 11/06/2026 in front of the committee

Dr. **H**. **B**ouanani	University of Saida Dr. Moulay Tahar	President
Dr. **M**. **K**adi	University of Saida Dr. Moulay Tahar	Supervisor
Dr. **L**. **Y**ahiaoui	University of Saida Dr. Moulay Tahar	Examiner

June, 2025/2026

[1]e-mail: fdjebli29@gmail.com

Dedication

I dedicate this humble work to my beloved parents,
in gratitude for their endless sacrifices, support,
and encouragement throughout my academic journey,
despite all the difficulties and challenges of life.
They have always been my strength and inspiration.

To my dear mother, whose sincere prayers and constant support
gave me the courage to continue and succeed.

To my beloved brothers and sisters,
thank you for your love and continuous encouragement.
I wish you all happiness and success in your lives.

To my dearest friends, ***Shakira*** *and* ***Asma****,*
thank you for all the beautiful memories and unforgettable moments
that will always remain close to my heart.

And finally, to the memory of my dear friend ***Ibtissam****,*
may Allah grant her eternal peace and mercy in Paradise.
Her memory will forever live in our hearts.

Acknowledgements

"All praise is due to God, who illuminated our paths with the beauty of His guidance, and facilitated for us the ways of knowledge and learning."

I extend my sincere thanks and deep gratitude to my supervisor, **Dr. Mokhtar Kadi**, for kindly accepting to supervise this thesis, and for the sound guidance, valuable advice, and great patience he provided throughout the completion period. To him, I express all appreciation and respect.

I also extend my thanks to all **members of the jury**, for accepting to evaluate and review this work, whose scientific remarks I will certainly benefit from in enriching this research.

Contents

List of Figures

List of Tables

General Introduction

Waiting has become a common phenomenon in modern life. It can be observed in many situations such as banks, supermarkets, post offices, hospitals, transportation systems, communication networks, and many other service environments. The study of such phenomena has attracted considerable attention from researchers and has led to the development of queueing theory, which provides mathematical tools for analyzing congestion and service systems.

The origins of queueing theory date back to the pioneering work of the Danish engineer Agner Krarup Erlang in the early twentieth century. His studies on telephone traffic in Copenhagen aimed to determine the number of circuits required to provide an acceptable level of telephone service, laying the foundation for modern queueing theory [43]. Since then, queueing theory has evolved considerably and has become an essential tool for modeling and analyzing a wide variety of stochastic service systems [8].

With the rapid development of computer systems, communication networks, and information technologies, queueing models have found numerous applications in performance evaluation and resource management. Nowadays, applications related to mobile communications, the Internet, multimedia systems, and computer networks represent some of the most active research areas in queueing theory [5].

Despite the significant theoretical advances achieved in this field and the availability of numerous analytical models, several challenges remain. Real-life service systems often involve random customer arrivals, varying service rates, and complex server behaviors that cannot always be accurately represented by classical queueing models. Consequently, researchers have proposed various extensions of traditional queueing systems in order to better reflect practical operating conditions.

One of the most important extensions is the *Working Vacation Policy*, introduced by Servi and Finn [34]. Unlike classical vacation models, where the server completely stops serving customers during vacation periods, the working vacation policy allows the server to continue providing service at a reduced service rate. This mechanism offers a more realistic representation of many practical systems, including manufacturing systems, computer networks, telecommunication systems, and customer service centers.

Since its introduction, the working vacation concept has attracted considerable research interest. Wu and Takagi extended this policy to more general queueing systems, including M/G/1 models [12]. Later, Kim investigated the queue length distribution in M/G/1 working vacation systems [31], while Li and Tian analyzed GI/M/1 and GI/Geo/1 queueing systems under different vacation policies [33]. More recently, Azhagappan studied Markovian queueing systems with working vacations, reneging customers, and server waiting mechanisms, providing a more realistic representation of service environments [3].

The objective of this thesis is to study and analyze queueing systems operating under the *Working Vacation Policy*, with particular emphasis on the M/M/1 queue with working vacation. The main purpose is to investigate the influence of system parameters on various performance measures and to evaluate the effectiveness of this policy in improving system performance.

This thesis is organized into three chapters.

- **Chapter 1** presents the fundamental concepts required for the study of queueing systems. It introduces stochastic processes, including counting processes, renewal processes, Poisson processes, and birth-and-death processes, which constitute the mathematical foundation of queueing theory.
- **Chapter 2** is devoted to the theory of queueing systems. It presents the principal concepts and notations used in queueing analysis, including Kendall's notation and Little's law. Furthermore, the main performance measures of queueing systems, such as the average number of customers in the system, queue length, and waiting times, are discussed.
- **Chapter 3** focuses on the analysis of the M/M/1 queueing model with Working Vacation Policy. The steady-state behavior of the system is investigated, equilibrium equations are derived, and several performance measures are evaluated. Numerical results are provided to illustrate the impact of model parameters on system performance and to highlight the effectiveness of the working vacation mechanism.

Through this study, we aim to provide a better understanding of the behavior of queueing systems operating under working vacation conditions and to contribute to the development of more efficient service systems in various practical applications.

CHAPTER 1

Stochastic Processes

The concept of a random process is considered as a set of random variables, each variable generally indexed by a time parameter. Random processes are widely used in modeling waiting systems (queues) because they help understand the behavior of customers and services over time. The concept of a random process is therefore a generalization of the concept of a random variable. The study of stochastic processes falls within probability theory, of which it constitutes one of the deepest objectives. It raises interesting and often very difficult mathematical problems. Random processes can also be classified into continuous-time random processes and discrete-time random processes.

1.1 Definition and classification

Definition 1.1.1: stochastic process. A stochastic process is a family of random variables $\{X_t, t \in T\}$ where each random variable X_t is indexed by the parameter $t \in T$, if T is a subset of $\mathbb{R}_+$, then t means time.

Generally X_t represents the state of the stochastic process at time t[45].

- If T is countable, i.e., $T \subseteq \mathbb{N}$, then we say that $\{X_t, t \in T\}$ is a discrete-time process.
- If T is an interval of $[0, \infty)$, then the stochastic process is said to be a continuous-time process.

The set of values of X_t is called the state space, which can also be either discrete (finite or countably infinite) or continuous (a subset of $\mathbb{R}$ or $\mathbb{R}^n$), so we write $(X_n)_{n \geq 0}$ for the discrete-time process and $(X_t)_{t \geq 0}$ for the continuous-time process.

Examples

- **Discrete-time process**: The number of customers arriving at a bank, the number of planes landing at an airport.
- **Continuous-time process**: The arrival time of customers in a supermarket, waves at sea.

1.2 Continuous-time stochastic processes

Customer arrivals to a queueing system are phenomena that can be characterized by the set of instants $(A_n)_{n\in\mathbb{N}^*}$ or arrival dates of each customer, which are random variables. They can also be described by the counting process $(N_t)_{t\in\mathbb{R}^+}$, or by the family $(T_n)_{n\in\mathbb{N}^*}$ representing the time intervals between two arrivals. The collection of these arrival dates can be modeled by the Poisson process[27].

1.2.1 Elementary definitions and properties

There are three natural ways to describe such a process:

1. First, one can encode a realization of such a process by a collection

$$0 < T_1(\omega) < T_2(\omega) < \dots$$

   of positive real numbers, corresponding to the position of points on $\mathbb{R}^+$. It is convenient to also set $T_0 = 0$.

2. A second way to encode a realization is to give, for each interval of the form $I = (t, t + s]$, the number of points $N_I(\omega)$ contained in the interval. If we use the simplified notation $N_t = N(0; t]$, we will then have

$$N(t; t + s] = N_{t+s} - N_t.$$

   The relationship between the random variables T_n and N_t is therefore simply

$$N_t(\omega) = \sup\{n \geq 0 : T_n(\omega) \leq t\}, \quad T_n(\omega) = \inf\{t \geq 0 : N_t(\omega) \geq n\}.$$

3. A third natural way to encode this information is to consider the sequence

$$X_1(\omega), X_2(\omega), X_3(\omega), \dots$$

   of positive real numbers corresponding to the successive distances between two points. The relationship between these variables and the T_n is given by

$$X_k = T_k - T_{k-1}, \quad T_k = \sum_{i=1}^{k} X_i.$$

Remark

A mathematical model used to represent systems progressing at specific intervals (for example, through "time slots" or distinct moments) rather than continuously is called a discrete-time stochastic process. In this framework, events (such as customer arrivals or service completions) are only represented at these precise moments, and not at any intermediate instant.

1.3 Counting process

Definition 1.4.0.1. (counting process)

Let $\{N(t), t \geq 0\}$ be a stochastic process. If $N(t)$ represents the total number of events that have occurred before time t, we say that $N(t)$ is a **discrete-state continuous-time counting process**.

Any counting process satisfies the following properties:

1. The number $N(t)$ takes positive integer values, for all $t \geq 0$.
2. The function $t \mapsto N(t)$ is increasing.
3. The difference $N(t) - N(s)$ represents the number of events occurring in the time interval $]s, t]$, for any pair (s, t) with $0 < s < t$.

The inter-arrival time process $\{W_n\}_{n\in\mathbb{N}_0}$ where W_n is the waiting time between the $(n-1)^{\text{th}}$ and the n^{th} occurrence, is a process that can be associated with the occurrence time process, i.e.:

$$W_n = T_n - T_{n-1}$$

with T_n being the arrival time of the n^{th} customer.

Proposition 1.4.0.1:

The following relations are trivial, with $T_0 = 0$:

1. $T_n = W_1 + W_2 + \cdots + W_n, \quad n \geq 1$;
2. $N(t) = \sup\{n \geq 0 : T_n \leq t\}$;
3. $P[N(t) = n] = P[T_n \leq t < T_{n+1}]$;
4. $P[N(t) \geq n] = P[T_n \leq t]$;
5. $P[s < T_n < t] = P[N(s) < n \leq N(t)]$.

Proof: We have $W_n = T_n - T_{n-1}$,The detailed proofs of Properties (1)–(5) can be found in Referencel[9].

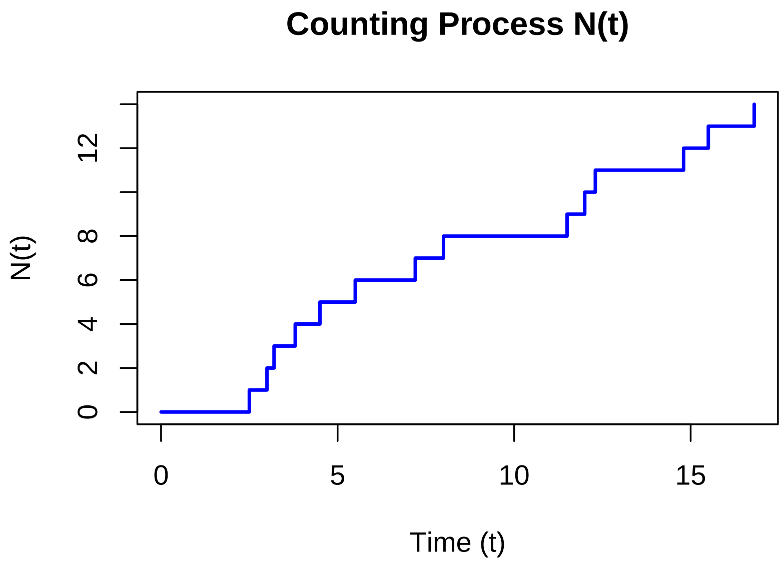

Figure 1.1: Schematic diagram representing the counting process $N(t)$ as a function of t.

Definition 1.4.0.2: (process with independent increments)

A counting process $\{N(t), t \in \mathbb{R}^+\}$ is called a **process with independent increments** if, for all $n \in \mathbb{N}^*$ and for all $t_1, \ldots, t_n$ such that $t_1 < t_2 < \cdots < t_n$, the increments

$$N_{t_1} - N_0, \quad N_{t_2} - N_{t_1}, \quad \ldots, \quad N_{t_n} - N_{t_{n-1}}$$

are independent random variables.

Definition 1.4.0.3: (process with stationary increments)

The process is said to be **stationary** (or **time-homogeneous**) if, for all $s \geq 0$ and for all $t \geq 0$, the increment

$$N_{t+s} - N_s$$

has the same distribution as N_t.

1.4 Renewal process

The renewal process is intended to count the number of occurrences of a specific phenomenon, in the case where the intervals between two successive occurrences are independent and identically distributed random variables.

Definition 1.5.0.1: A counting process for 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:

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

with $a_i, i = 1; 2; \ldots$ being the time between two consecutive arrivals. It is easy to see that the number of arrivals before time t, i.e. the process

$$\{N(t); t \geq 0\} = \sup_k \{k \in \mathbb{N} : A_k \leq t\}$$

is a counting process[15].

Definition 1.5.0.2: A counting process for which the waiting times are i.i.d. is called a renewal process. The Poisson process is the most important example of a renewal process.

Example:

- Number of customers who entered a store up to time t.

1.5 Poisson process

The Poisson process is a mathematical model that describes the distribution of independent random events on the positive real line $(0, \infty)$. In the intended context, it does not describe a uniform distribution as the initial description might suggest, but rather relies on a constant occurrence rate λ, where the intervals between events follow an exponential distribution, and the number of events in a given time period follows a Poisson distribution.

Application examples

It can be used to model random phenomena in various fields, such as:

- Customer arrivals: Number of customers arriving at a store or bank at a given time (for example, 5 customers per hour).
- Particle emission: Number of particles emitted by a radioactive substance in a time period.
- Accidents or failures: Number of road accidents on a highway or failures in a technical system.
- Network requests: Number of requests entering an internet server in one minute.

- Defect distribution: Random distances between defects in an industrial material on a production line.

1.5.1 Definition and properties

Definition 1.6.1.1: (Poisson Process)

The counting process $(N(t))_{t\geq 0}$ is called a **Poisson Process** with intensity $\lambda > 0$ if:

1. $N(0) = 0$.
2. The process has independent increments, that is, for $t > s$, the number of jumps $N(t) - N(s)$ occurring in $(s, t]$ is independent of the number of jumps $N(s)$ occurring before time s.
3. The number of events in an interval of length t is distributed according to a Poisson distribution $P(\lambda t)$:

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

Increments of the same length of the Poisson process all have the same distribution, we say that the process has **stationary increments**[22].

Remark

This parameter λ is called the intensity of the Poisson process $\{N_t,\ t \geq 0\}$. It is equal to the average number of events occurring during a unit length time interval, i.e.

$$\mathbb{E}[N_{t+1} - N_t] = \lambda.$$

1.5.2 Main properties.

- **Number of events at time** t: Follows a Poisson distribution with mean λt.
- **Waiting time between events**: Follows an exponential distribution with rate λ.
- **Independence**: Events do not influence each other.

1.5.3 Homogeneous Poisson process

The Poisson process on the line is a continuous-time process that only takes positive integer values. It is also described as a counting process, denoted by $\{N(t) : t > 0\}$.

We propose to examine the random number $N(t)$ of specific events that occur in a given time period $[0, t]$. Its strong presence in applications stems notably from the fact that many calculations concerning it are explicit.

For a homogeneous Poisson process with rate λ, we have:

$$\mathbb{P}(N(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n = 0, 1, 2, \ldots$$

1.6 Relationship between Counting Process, Renewal Process and Poisson Process

Counting Process (General counting process)

$N(t)$: number of events up to time t

Example: Number of customers entering a store up to time t

↓

Renewal Process (Particular case)

Inter-arrival times $X_1, X_2, \dots$ i.i.d.

$$S_n = X_1 + \cdots + X_n$$
$$N(t) = \max\{n : S_n \leq t\}$$

Example: Number of successive failures of a machine up to time t

↓

Poisson Process (Special case of renewal process)

$$X_i \sim \text{Exponential}(\lambda)$$
$$\Rightarrow N(t) \sim \text{Poisson}(\lambda t)$$

Memoryless property

Example: Number of phone calls received at a call center between 0 and t

↓

Homogeneous Poisson Process

Case where the rate λ is constant over time

Example: Number of cars passing through a toll per minute (on average)

Figure 1.2: Hierarchical structure of counting, renewal and Poisson processes

1.7 Poisson and exponential distributions

1.7.1 Poisson distribution

The Poisson law is attributed to Simeon D. Poisson, a French mathematician (1781–1840). This law was proposed by Poisson in a work he published in 1837 under the title "Research on the probability of judgments in criminal and civil matters."

The Poisson distribution with parameter $\lambda > 0$ is given by:

$$P_k = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad \text{for } k = 0, 1, 2, \cdots \tag{1.1}$$

Let X be a random variable having the Poisson distribution in (1.1). We evaluate the mean, or first moment, by:

$$E[X] = \sum_{k=0}^{+\infty} kP_k = \lambda.$$

To find the variance, we first determine

$$E[X(X-1)] = \sum_{k=2}^{+\infty} k(k-1)P_k = \lambda^2,$$

then

$$E[X^2] = E[X(X-1)] + E[X] = \lambda^2 + \lambda,$$

while

$$\sigma_X^2 = \text{Var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

1.7.2 Exponential distribution

A random variable T follows an exponential distribution with parameter μ if its cumulative distribution function is:

$$F(t) = \begin{cases} 1 - e^{-\mu t}, & t \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The expectation and variance of an exponential distribution are:

$$E(T) = \frac{1}{\mu}, \qquad \text{Var}(T) = \frac{1}{\mu^2}.$$

1.7.3 Relationship between the two distributions

The exponential distribution and the Poisson distribution are intimately linked. In a Poisson process, the inter-arrival times follow an exponential distribution.

The probability density function of an exponential distribution

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

Suppose τ is exponential with mean $\frac{1}{\mu}$, and n is Poisson with mean μ we have:

$$P(\tau > t) = 1 - F(t) = e^{-\mu t} = P(n = 0 \text{ at } t) = P(0, t).$$

Let $P(n, t)$ denote the probability of having n units in time t.

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

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

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

$$= \cdots$$

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

1.7.4 Memoryless property

In probability and statistics, the memoryless property is a property of certain probability distributions, notably the exponential distribution and the geometric distribution. They are said to be distributions without memory .

The memoryless property makes a comparison between the probability distributions of the server's waiting time, and that of the server's waiting time for a customer to arrive after an arbitrary delay following opening. The memoryless property asserts that these distributions are the same.

The only continuous probability distribution with the memoryless property is the exponential distribution, thus the memoryless property characterizes the exponential distribution among all continuous distributions.

Proposition 1.8.4.1. (Memoryless property). A random variable T taking values in $\mathbb{R}^+$ follows an exponential distribution if and only if it satisfies the memoryless property:

$$\forall s, t \geq 0, \quad \mathbb{P}(T > t + s \mid T > s) = \mathbb{P}(T > t).$$

Proof.

Suppose first that T follows an exponential distribution with parameter λ. We have

$$\mathbb{P}(T > s + t \mid T > s) = \frac{\mathbb{P}(T > s + t,\ T > s)}{\mathbb{P}(T > s)} = e^{-\lambda t} = \mathbb{P}(T > t).$$

Conversely, suppose that T satisfies the memoryless property. We set

$$g(t) = \mathbb{P}(T > t), \quad \forall t \geq 0.$$

We have that g is decreasing on $\mathbb{R}^+$ and satisfies

$$\lim_{t \to 0} g(t) = 1 \quad \text{and} \quad \lim_{t \to +\infty} g(t) = 0.$$

Moreover,

$$\mathbb{P}(T > s + t \mid T > s) = \frac{\mathbb{P}(T > s + t,\ T > s)}{\mathbb{P}(T > s)} = \frac{g(s+t)}{g(s)},$$

and

$$\mathbb{P}(T > t) = g(t).$$

Hence

$$g(s)g(t) = g(s + t), \quad \forall s, t > 0.$$

We conclude using the following lemma that

$$g(t) = e^{-\lambda t},$$

and therefore T follows an exponential distribution.

1.7.5 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, similarly to the Poisson process.
- **Duration between events**: The duration between each event (called the "waiting time") follows an *Erlang distribution*.

The probability density function of a random variable X following an Erlang distribution of order k and parameter λ is given by [17]:

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

1.8 Birth and death process

These processes generally allow describing the temporal evolution of the size of a population of a given type. In the case of a queueing system, we consider for example populations comprising all customers who are in the system at time t. Birth and death processes are continuous-time stochastic processes with a discrete state space $n = 0, 1, 2, \ldots$. They are memoryless, and from a given state n, transitions are only possible to one or the other of the neighboring states $(n+1)$ and $(n-1)$ for $n \geq 1$. Markovian (M/M) queueing models are very important special cases of birth and death processes. Their complete study will be carried out in the following chapter.

1.8.1 Birth process

Definition 1.9.1.1: (birth process) The birth process is the direct generalization of a Poisson process when the intensity parameter λ depends on the current state of the process; it will allow us to introduce the concept of "explosion".

If the size of a population undergoes a transition $n \to n + 1$, it corresponds to a birth.

1.8.2 Death process

If the size of a population undergoes a transition $n \to n-1$, it corresponds to a death.

1.8.3 Definition (birth and death process)

This is a particular case of a Markov chain [1], where only transitions from one state to a neighboring state are permitted; we are interested in the continuous case with transition rates. It is the starting point of queueing theory.

We introduce the following data:

- λ_n: birth rate.
- μ_n: death rate.

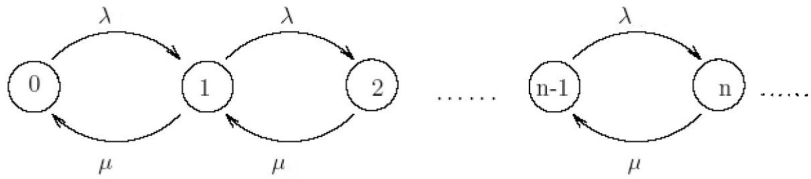

Figure 1.3: Transition graph of a birth and death process.

Example

In a hospital emergency department, it is assumed that the time interval between two successive patient admissions follows an exponential distribution. Furthermore, the patient care time (whether it involves a consultation or treatment) follows an exponential distribution.

Thus, the number of patients in the department can be represented by a birth and death process:

- birth $\Rightarrow$ admission of a patient to the department.
- death $\Rightarrow$ discharge of the patient after treatment.

CHAPTER 2

Queueing Theory

Queueing theory began in 1909 with the pioneering research of the Danish engineer **Agner Krarup Erlang (1878–1929)**, who studied telephone traffic in Copenhagen to determine the number of circuits required to provide an acceptable level of telephone service. Subsequently, queueing models were integrated into the analysis of various fields of activity, leading to rapid developments in the theory.

Queueing theory was soon applied to the performance evaluation of computer systems and communication networks. Researchers in this field developed several new analytical methods that were later successfully extended to other domains, particularly in the manufacturing sector. A renewed interest in practical applications of queueing theory also emerged in more traditional areas of operations research, notably through the contributions of Peter Kolesar and Richard Larson.

Thanks to these developments, queueing theory has become a widely used mathematical tool with numerous practical applications.

This chapter presents the fundamental concepts of queueing theory, including Kendall's notation and Little's formula, before examining several classical Markovian models such as M/M/1, M/M/1/K, M/M/c, and M/M/∞, as well as the main methods used for performance evaluati

Queue classification:

To describe a queue, the following elements must be specified:

- The nature of the arrival process, which is defined by the distribution of the intervals separating two consecutive arrivals.
- The distribution of random service time.
- The number s of service stations.
- The capacity N of the system. If $N < 1$, the queue cannot exceed a length of $N - s$ units. In this case, some customers arriving at the system will not be able to enter it[38].

Terminology and notations:

In connection with the exponential distribution:

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

The examination of a queueing system depends on the initial condition and the time that has elapsed. This corresponds to a transient phase during which the analysis becomes particularly complex. In queueing theory, the analysis is carried out once the system has reached its steady state; a state in which the system conditions are essentially independent of its initial phase and of the time that has already elapsed. It is assumed that the system has been operating for an extremely long period of time [2].

In a steady-state system:

We define[37]:

- P_n : Probability of having n customers in the system.
- $\bar{N}_s$: Average number of customers in the system.
- $\bar{N}_q$: Average number of customers in the queue.
- $\bar{W}_s$: Average time spent in the system (waiting + service).
- $\bar{W}_q$: Average waiting time of a customer in the queue.
- c : Number of servers.

2.1 Description of a Simple Queue

A queueing system's general modeling can be shown as follows. Customers, or service requests, come in; they could be individuals, phone calls, electrical signals, cars, incidents, or other entities at a particular location asking for a particular service. The customer goes to the server, if it is available, and the service is rendered. The customer can either exit the system or join the waiting line if the service is unavailable. A customer is chosen for service at a specific time based on a predetermined discipline. This is shown graphically in Figure (2.1).

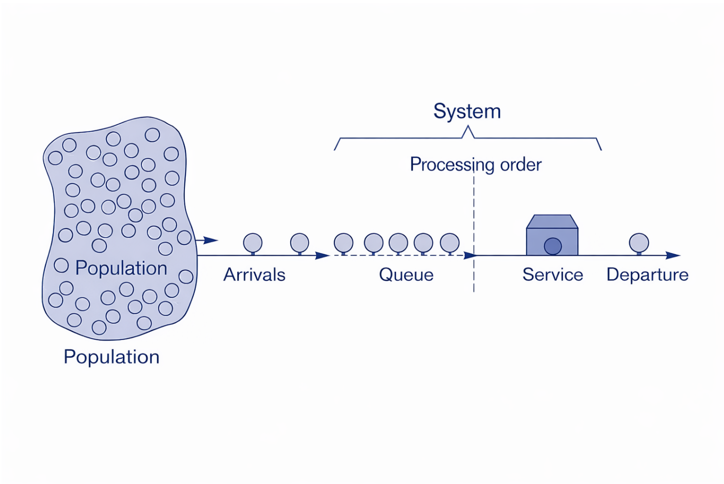

Figure 2.1: Simple queue

2.1.1 Arrival Process

A random process for calculating $(N_t)_{t\geq 0}$ is used to describe customer arrivals at the station.

If A_n denotes the random variable that measures the moment the n-th client enters the system, we then have:

$$A_0 = 0 \quad \text{and} \quad A_n = \inf\{t \geq 0 \, ; \, N_t = n\}.$$

If we define T_n as the random variable that measures the time interval between the arrival of customer number $(n-1)$ and that of customer number n, according to Khintchine [28], we have:

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

2.1.2 Service Process

Let's first examine the case of a single-server queue. We denote by D_n the random variable that indicates when the service of the nth client of the system begins, and by Y_n, the variable that defines the service duration of the nth client (the period between the start and end of the service).

A start time always marks the end of a service, but does not necessarily indicate the beginning of another.

It is entirely possible that a departing client leaves the station unoccupied.

The server therefore remains inactive until the arrival of the next client.

We denote by μ the service level[40]:

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

2.1.3 Queue Structure

Number of Servers 2.1.3.1: A service station may have several servers operating simultaneously. We denote by C the number of servers.

When a customer arrives at the station, two situations may occur:

- A server is available, and the customer is immediately served.
- All servers are busy, and the customer must wait in the queue until a server becomes available.

It is generally assumed that the servers are identical and operate independently of one another.

Specific Case 2.1.3.2: IS Station (Infinite Servers) A notable specific case is the IS station (Infinite Servers), where the number of servers is unlimited ($C = +\infty$).

In this scenario, each customer receives service immediately upon arrival without any delay. Therefore, there is no queue in this type of station.

Queue Capacity 2.1.3.3: The possibility that the queue can accommodate customers waiting for service can be restricted or unrestricted. If K symbolizes the capacity of the queue, then for a queue without capacity limit, we have $K = +\infty$.

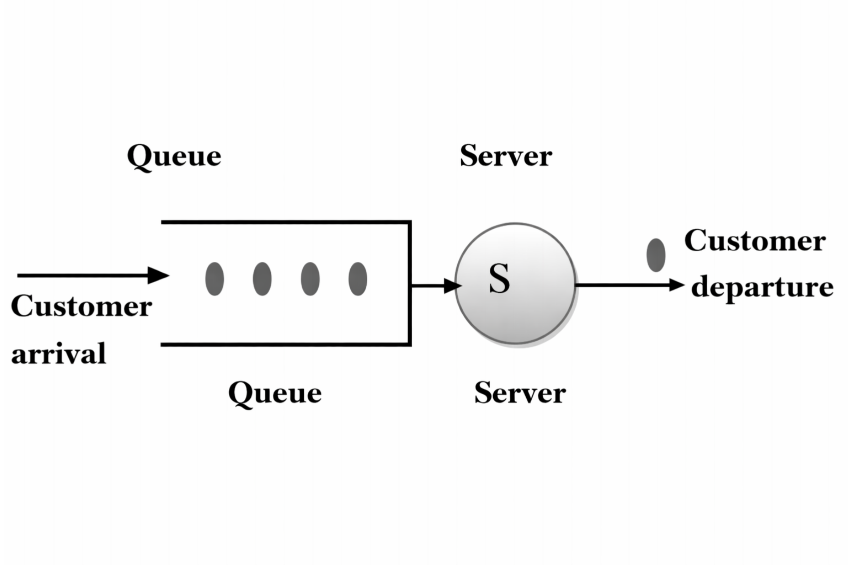

Figure 2.2: Single-server queueing system

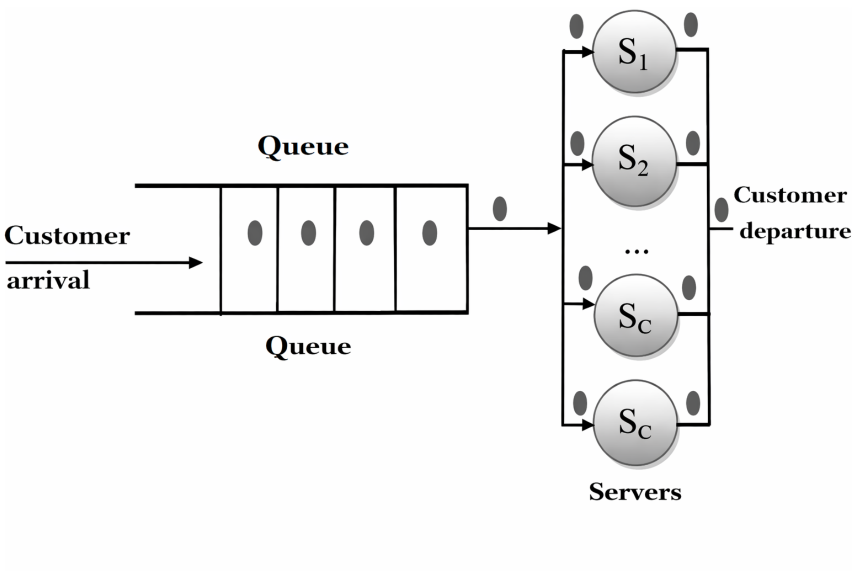

Figure 2.3: Queueing system with C parallel servers

2.2 Kendall Notation

Kendall's notation is a commonly used notation for categorizing various queuing systems[41].

$$Q/R/C/L/K/Z$$

with

1. **Q** : indicates the arrival process of customers. The codes used are:

⧫ **M (Markov)** : The inter-arrival times of customers are independent and identically distributed according to an exponential law. It corresponds to a Poisson point process (memoryless property).

⧫ **D (Deterministic)** : The inter-arrival times of customers or service times are constant and always the same.

⧫ **GI (General Independent)** : The inter-arrival times of customers have a general distribution (no assumption on the distribution) but they are independent and identically distributed.

⧫ **G (General)** : The inter-arrival times of customers have 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. **R** : describes the distribution of service times of a customer. The codes are the same as for Q.

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

4. **L** : queue capacity, i.e., the number of places in the system; in other words, it is the maximum number of customers in the system, including those in service.

5. **K** : customer population.

6. **Z** : service discipline, i.e., the way customers are ordered to be served. The codes used are as follows:

⧫ **FIFO (first in, first out) or FCFS (first come, first served)** : this is the standard queue in which customers are served in their order of arrival. Note that FIFO and FCFS disciplines are not equivalent when the queue has multiple servers. In the first, the first customer to arrive will be the first to leave the queue, whereas in the second, they will be the first to start service. It is then possible that a customer who starts service later on another server may finish before them.

⧫ **LIFO (last in, first out) or LCFS (last come, first served)** : this corresponds to a stack, in which the last customer to arrive (thus placed on the stack) will be the first to be served (removed from the stack). Again, LIFO and LCFS disciplines are only equivalent for a single-server queue.

⧫ **SIRO (Served In Random Order)** : customers are served randomly.

♦ **PNPN (Priority service)** : customers are served according to their priority. All customers of the highest priority are served first, then customers of lower priority, and so on.

♦ **PS (Processor Sharing)** : customers are served equally. The system capacity is shared among the customers.

2.3 Little's Law

Little's Law (1961) is a very general relationship that applies to a large class of systems. It concerns only the steady-state regime of the system. No assumption on the random variables that characterize the system (inter-arrival times, service times, . . .) is necessary. The only condition for applying Little's Law is that the system is stable.

$$d_s = d_e = d$$

Little's Law can be expressed as in the following theorem (Theorem 2.4.1):

Theorem 2.4.1 (Little's Formula)

The average number of customers L, the average time spent in the system W, and the average throughput d of a stable system in steady state are related as follows:

$$L = W \times d$$

Remark 2.4.1

Little's Law applies to all queueing models encountered in practice (not only to the M/M/1 queue).

2.4 Modeling of a Queueing System

2.4.1 Markovian Queueing Systems

In 1906, A. A. Markov began studying a new type of random process, called a Markov chain. These are Markovian queues in which the waiting times between arrivals and service times are exponential. They will be denoted by $M/M/\cdots$ (M as in Markovian...)

2.4.2 Non-Markovian Queueing Systems

A process is referred to as non-Markovian if one of the two stochastic variables—the time between arrivals or the duration of service—is not exponentially distributed. Additional parameters relating to queueing issues make the model non-Markovian. These elements make the mathematical analysis of the model particularly complex, if not impossible. This is why we attempt to convert this model into a Markov process using one of the following examination methods[29]:

1. **Erlang phase technique:** It aims to approximately reproduce the probability distribution that possesses a rational Laplace transform by a Cox distribution (composed of a combination of exponential distributions), the latter being characterized by the memoryless property at each stage.
2. **Induced Markov chain approach:** Its principle consists of selecting a sequence of moments $1, 2, 3, \ldots, n$ such that the induced chain N_n; $n \geq 0$, where $N_n = N(n)$, is Markovian and homogeneous.

Examples

For example:

In a hospital:

The distribution of patients does not follow an exponential pattern.

The duration of treatment varies greatly.

Therefore, instead of $M/M/1$, the appropriate model is $G/G/1$ or $M/G/1$.

In a network of computers:

Intermittent (non-exponential) distribution of packet arrivals may occur.

There is variability in processing times.

The model is no longer Markovian.

2.5 Queueing System Performance

The objective of studying a queue or a network of queues is to evaluate the performance of a system under specific operating conditions. The commonly used indicators are as follows: If the average arrival rate of customers per given period, denoted by λ, is lower than the average rate at which customers can be served during the same period, then the queue remains stable. If each server can process μ customers per unit of time and the total number of servers is λ, a queue is considered stable if and only if $\lambda < \mu$, $\rho = \frac{\lambda}{\mu} < 1$, where ρ is called the traffic intensity.

2.6 Some Queueing Models

2.6.1 M/M/1 Queueing Model

Description of the Model

The M/M/1 queue is a model characterized by arrivals following a Poisson process with rate λ, exponential service times with parameter μ, and a single server. Customers arrive at the station according to a Poisson process with rate λ. If the server is idle, the customer is served immediately; otherwise, the customer joins the queue with unlimited capacity and FIFO discipline. The queue can be considered as a birth-and-death process for which

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

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

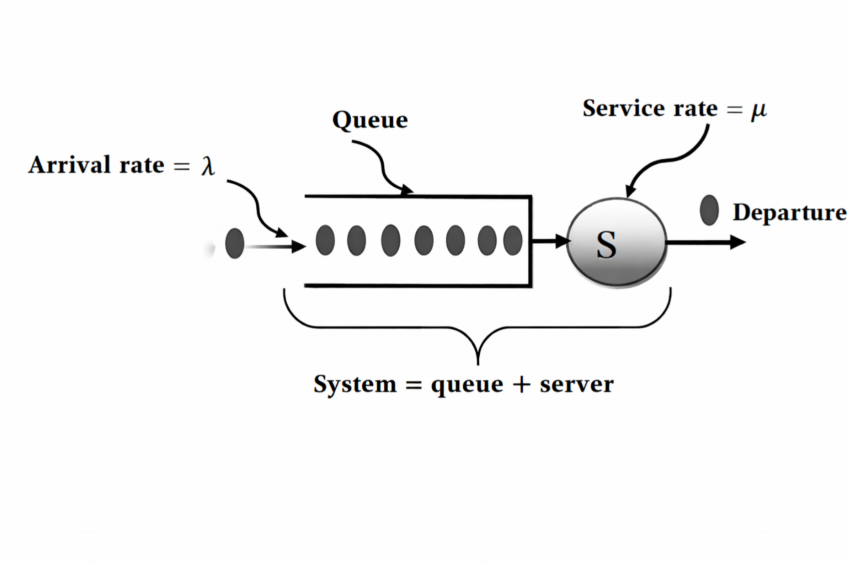

Figure 2.4: The M/M/1 queue

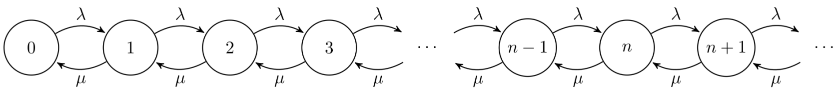

Figure 2.5: Transition diagram of the M/M/1 queue.

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

$$p_n'(t) = -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), \quad n \geq 1 \tag{2.1}$$

and

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \tag{2.2}$$

Under the assumption that $\lambda < \mu$ (the arrival rate is smaller than the service rate), we have:

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

The state probabilities for the stationary regime of the process are given by:

$$\pi_n = \pi_0 \, \rho^n$$

with

$$\pi_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1 - \rho$$

Thus,

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

$\pi = \{\pi_n\}_{n\geq 0}$ is called the stationary distribution; it follows a geometric distribution (the stationary probability of having n customers in the system).

System Characteristics

– The average number of customers in the system is:

$$\begin{aligned}\overline{N} &= E(N) \\ &= \sum_{n=1}^{\infty} n\,\pi_n \\ &= (1-\rho)\sum_{n=1}^{\infty} n\,\rho^n\end{aligned}$$

Hence:

$$\overline{N} = \frac{\rho}{1-\rho}$$

– Average number of customers being served:

$$\overline{N}_S = 1 - \pi_0 = \rho \tag{2.1}$$

– Average number of customers in the queue:

$$\overline{N}_Q = \sum_{n=1}^{\infty} (n-1)\,\pi_n = \frac{\rho^2}{1-\rho}$$

Average sojourn time: The average sojourn time $\overline{W}$ is computed using Little's law:

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

– Average service time:

$$\overline{W}_S = \frac{1}{\mu} \tag{2.2}$$

– Average waiting time:

$$\overline{W}_Q = \overline{W} - \overline{W}_S = \frac{\lambda}{\mu(\mu-\lambda)}$$

2.6.2 M/M/C Queueing Model

Description of the Model

We consider a system that is identical to the M/M/1 configuration, except that it has c servers, all of which are identical and operate independently. We maintain the following assumptions[19]:

- The arrival process of customers follows a Poisson process with rate λ.
- The service time of each customer follows an exponential distribution with rate μ.

We refer to this system as M/M/C. The state space E is similar to that of. The

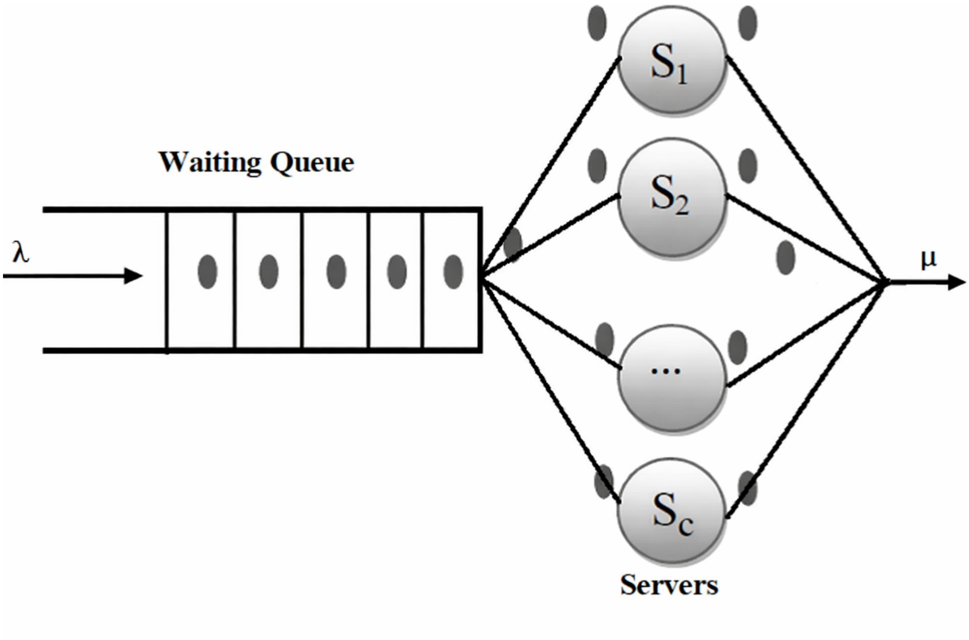

Figure 2.6: The M/M/C Queue

birth–death model representing this type of queue is then formulated as follows:

$$\lambda_n = \begin{cases} \lambda, & \forall n \geq 0 \end{cases}$$

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

The following conclusions can be drawn from the diagram. Studying the system in a steady state using the Chapman-Kolmogorov equations method leads to the following equations:

$$\lambda \pi_0 = \mu \pi_1$$

$$(\lambda + n\mu)\pi_n = \lambda \pi_{n-1} + (n+1)\mu \pi_{n+1}, \quad 1 \leq n < c$$

$$(\lambda + c\mu)\pi_n = \lambda\pi_{n-1} + c\mu\pi_{n+1}, \quad n \geq c$$

with

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

The resolution of the above system yields the following stationary distribution:

$$\begin{aligned} \bar{N}_Q &= \sum_{n=1}^{\infty}(n-1)\pi_n \\ &= \bar{N} - (1 - \pi_0) \end{aligned}$$

$$\pi_n = \frac{\rho^c}{c!}(A)^{n-c}\pi_0, \quad n \geq c \tag{2.5}$$

or

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!}\sum_{n=c}^{\infty} \rho^{n-c}\right]^{-1}$$

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

and

$$A = \frac{\lambda}{c\mu}$$

The latter exists if: $\lambda < C\mu$

System Characteristics

From the stationary distribution of the process $\{N(t), t \geq 0\}$, we can calculate the system characteristics. Indeed,

- **The average number of customers in the system is:**

$$\bar{N} = \rho + \frac{\rho^{c+1}}{c \cdot c!(1-A)^2}\rho_0 \tag{2.6}$$

- **The average number of customers in the system is:**

$$\bar{N}_Q = \frac{\rho^{c+1}}{c \cdot c!(1-A)^2}\rho_0 \tag{2.7}$$

$$\bar{W} = \frac{c\mu\rho^c}{c!(c\mu - \lambda)^2}\rho_0 \tag{2.8}$$

- **Average waiting time:**

$$\bar{W}_Q = \frac{1}{\mu} + \frac{\rho^c}{\mu c \cdot c!(1-A)^2}\rho_0 \tag{2.9}$$

2.6.3 M/M/1/K Queueing Model

Description of the Model

The **M/M/1/K queueing model** is a fundamental framework in random process research for analyzing single-server systems with limited capacity K, which represents the maximum number of customers that can be in the system, either waiting or being served.

This system is characterized as follows:

- Customer arrivals to the queue follow a Poisson process with rate λ.
- Customer service times are exponentially distributed with rate μ.
- Let K denote the queue capacity, which is the maximum number of customers that can be in the system, either waiting or being served. When a customer arrives and there are already K customers in the system, the customer is lost (for example, in a waiting room of a small clinic).

This system is known as an **M/M/1/K queueing system**.

The state space is finite and given by:

$$E = \{0, 1, 2, \ldots, K\}.$$

Since the capacity of the queue is limited, even if customers arrive on average much faster than they can be served, once the queue is full, new customers are turned away before the server can process them. As a result, the number of customers in the system can never become infinite.

Additionally, once a customer is admitted into the system, they will eventually leave, and their waiting time will come to an end because the total number of customers in the system is limited to K, which corresponds to the service time of all customers ahead of them.

The departure rate will eventually equal the arrival rate, which reflects the overall stability of the system. The birth and death processes that describe this type of queue are therefore well defined.

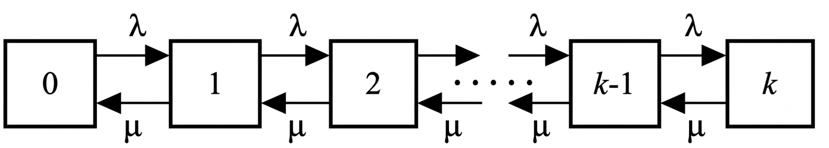

Figure 2.8: State transition diagram of the M/M/1/K queueing system

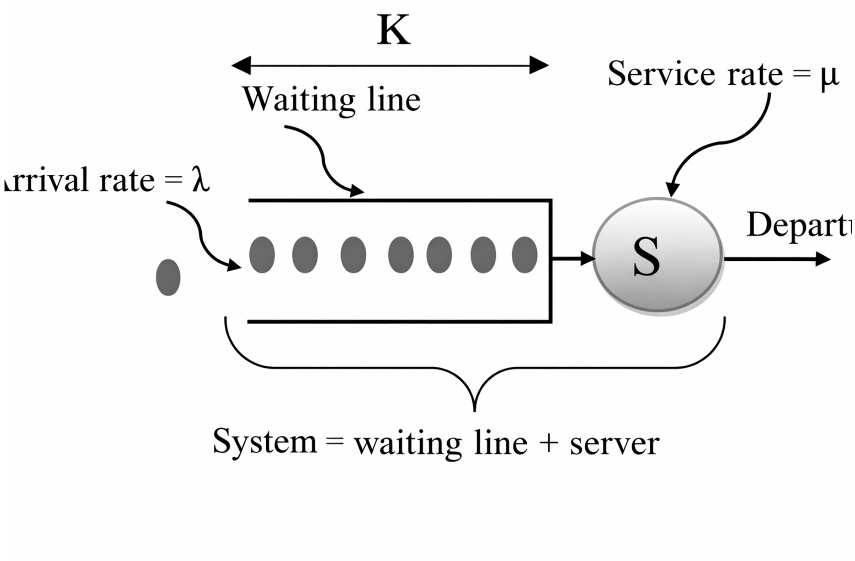

Figure 2.7: Figure 2.6 – The M/M/1/K Queue

The integration of the recurrence equation allowing the calculation of r_n is then carried out as follows[42]:

$$\pi_n = \pi_0 \rho^n, \quad 0 \leq n \leq K$$

$$\pi_n = 0, \quad n > K$$

and

$$\pi_0 = \begin{cases} \dfrac{1}{\sum_{n=0}^{K} \rho^n} = \dfrac{1-\rho}{1-\rho^{K+1}}, & \text{if } \lambda \neq \mu, \\ \dfrac{1}{K+1}, & \text{if } \lambda = \mu. \end{cases}$$

System Characteristics

Mean number of customers in the system The mean number of customers in the system is given by:

$$\bar{N} = \sum_{n=0}^{K} n\pi_n$$

$$\bar{N} = \frac{\rho}{1-\rho} \cdot \frac{1-(K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}}$$

When $K \to \infty$ and $\rho < 1$, we recover the M/M/1 queue case:

$$\bar{N} = \frac{\rho}{1-\rho}$$

Mean number of customers in the queue

$$\bar{N}_Q = \sum_{n=1}^{\infty}(n-1)\pi_n$$

$$\bar{N}_Q = N - (1 - \pi_0)$$

Using Little's law, we obtain the mean time a customer spends in the system $\bar{W}$ and the mean waiting time in the queue $\bar{W}_Q$.

Mean time a customer spends in the system

$$\bar{W} = \frac{N}{\lambda} \tag{2.3}$$

Mean waiting time in the queue

$$\bar{W}_Q = \frac{N_Q}{\lambda} \tag{2.4}$$

2.6.4 $M/M/\infty$ Queueing Model

Description of the Model

In this queueing model, the system comprises an unlimited number of identical, autonomous servers. As soon as a customer arrives, they are served instantly (i.e. there is no waiting time)[4]. In this queue, customers arrive at times 0 < t1 < t2 < : : :, constituting a Poisson process with a rate Service times follow an exponential distribution with a rate . The frequency of transition from one state to another The exit rate of a customer among the n customers in service, i.e. the probability of transition from state n to state n-1, is given by the following formula: n $\mapsto$ n-1. Thus, the transition rate from state n to state n+1 is equivalent to:

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

$$\pi_{n-1}\lambda = \pi_n n\mu, \quad n = 1, 2, \ldots$$

So:

$$\pi_n = \frac{\rho^n}{n!}\pi_0, \quad n = 1, 2, \ldots$$

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

Steady-State Probability Distribution: The normalization condition is:

$$\sum_{n=0}^{\infty} \pi_n = 1,$$

which leads to:

$$\pi_0 = \left(\sum_{n=0}^{\infty} \frac{\rho^n}{n!}\right)^{-1} = e^{-\rho}.$$

Thus, the steady-state probabilities are:

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

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.

System Characteristics

Average Number of Customers in the System: The expected number of customers in the system is given by:

$$\bar{N} = \sum_{n=1}^{\infty} n\pi_n$$

Substituting r_n:

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

Using the identity:

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

we obtain:

$$\bar{N} = e^{-\rho} \cdot \rho e^{\rho} = \rho.$$

Average Residence Time $\bar{W}$:

Using Little's Law, where:

$$\lambda = \sum_{n=1}^{\infty} \pi_n n\mu = e^{-\rho} \sum_{n=1}^{\infty} \frac{\rho^n}{(n-1)!}\mu = e^{-\rho} \cdot \rho e^{\rho}\mu = \rho\mu = \lambda,$$

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

Then:

$$\bar{W} = \frac{\bar{N}}{\lambda} = \frac{\rho}{\lambda} = \frac{1}{\mu}.$$

CHAPTER 3

Working Vacation Policy in Queueing Systems

The working vacation policy in queueing systems is an innovative concept that allows servers to provide services at a reduced rate during vacation periods instead of stopping service completely. This model was first introduced by **Servi, L. D. and Finn, S. G. (2002)**[34].

In this context, Markovian queueing systems with working vacations have been widely studied and have attracted considerable attention because of their numerous applications in information systems, communication networks, manufacturing, inventory management, and healthcare. Models such as $M/M/1$ and $M/G/1$ illustrate how this policy can be incorporated into practical systems.

Research has also shown the existence of stochastic decomposition structures for queue length and waiting time in these systems. Analytical methods such as quasi-birth-and-death processes and the matrix-geometric method have been used to obtain distributions and performance measures under different system parameters.

This approach is particularly useful for modeling systems in which a complete service interruption may lead to significant costs or substantial customer losses.

Therefore, the working vacation policy represents a realistic extension of traditional vacation models in queueing systems. It improves system flexibility and efficiency by allowing the server to continue serving customers at a reduced rate, thereby increasing customer satisfaction.

3.1 Working Vacation

3.1.0.1 Definition

Within the framework of queueing systems analysis, a "working vacation" is a policy whereby the server does not completely stop working during the vacation period. Instead, it continues to serve customers, but at a lower rate. This allows the server to maintain some activity and reduce customer dissatisfaction that might arise if the service were to stop completely. This approach is adopted to improve service efficiency and queue management.

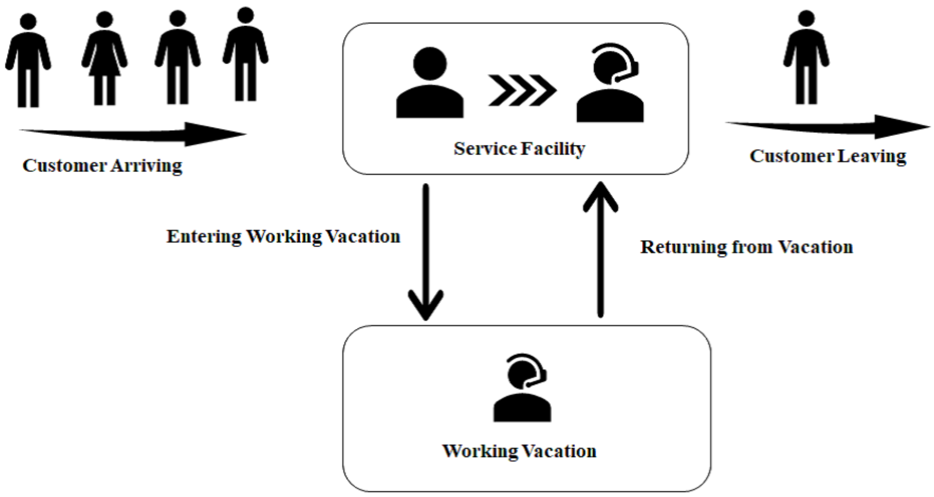

Figure 3.1: Queueing system with working vacation

3.1.0.2 The history

Over time, research on queueing models has changed significantly. In order to improve system performance, the scientists **Servi and Finn** [34] introduced the idea of a "working vacation" in **(2002)**. During this period, the server continues to provide service at a reduced rate.

After that, **Wu and Takagi** [12] expanded this research to include more complex models, such as $M/G/1$ models, in **(2006)**.

In order to better understand the number of customers in the system, **Kim** [31] and other researchers studied the distribution of queue length in an $M/G/1/WV$ system with working vacations in **(2003)**.

Li and Tian [33] examined the performance of $GI/M/1$ and $GI/Geo/1$ systems under different vacation and rest conditions **(2009)**.

However, **Zhang** used supplementary variable analysis and matrix techniques to analyze $GI/M/1/N$ models and randomized vacations, in addition to $M/M/1$ models with multiple vacationsin .

Azhagappan A. [3] studied a Markovian queue with working vacation, reneging, and a waiting server in **(2019)**, which improved the understanding of system behavior under more realistic conditions.

In general, these studies helped expand the theory of queueing models and enhance their applications in fields such as industry, transportation, and communication networks.

3.1.0.3 Classical Vacation vs Working Vacation

▲ Traditional Vacation ▲

- • During the vacation period, the server is totally inactive, meaning it ceases to provide any services.
- • During the vacation, the system is temporarily shut down and no clients are accepted.
- • Usually used for maintenance or service outage analysis.

⧫ Working Vacation ⧫

- • During the vacation, the server either keeps running at a slower pace or offers a service with less activity than it would during regular business hours.
- • During the vacation, it can still take clients, but its performance will be less than usual.
- • It is used to analyze systems that enable the server to run and complete a portion of a task during maintenance or vacation periods, increasing efficiency and time utilization.

These differences affect probability distributions, performance analysis, and queueing model system management strategies[34].

3.1.0.4 Applications of Working Vacation Models

Working vacation models have been applied in various fields due to their ability to represent situations in which the server or system continues to operate at a slower pace during periods of inactivity or rest, including[25]:

1. **Information and network systems:** Used to model file transfers, web server management, and messaging services, even though the server may continue to handle some requests with little activity.

2. **Production and fabrication management:** To maximize the use of idle or partially maintained machinery or production lines, allowing for a reduced but continuous activity during these periods.

3. **Health sector:** Use in the management of patient files in hospitals or medical centers, where staff may continue to provide a service at a reduced pace during periods of rest or low income.

4. **Reservation systems and queues:** In situations involving reservation or rework systems, the fact that the operator continues processing at a reduced pace helps decrease the total waiting time and optimize service availability.

5. **Telecommunication networks:** To simulate data transfer or request management, particularly in systems that experience interruptions or failures, where the

system can operate in a degraded mode rather than completely stopping its operations.

6. **Control and maintenance systems:** When a system can remain operational at a restricted level during repairs or interventions, thus reducing total interruptions.
7. **Applications in web services and file transfer:** The partial maintenance of a service during slowdown or maintenance phases helps to optimize the overall performance of the system.

3.2 Types of Vacations

3.2.1 Simple Vacation

The server takes only one break; once the system is idle, it remains so until a client arrives. For example, in hospitals, when a doctor has finished treating all patients, they may take a short break. If there are no patients when they return, they remain in a waiting state until a new patient arrives. This represents simple idle behaviour.

3.2.2 Multiple Vacation

Every server can take breaks on their own thanks to the Multi-Vacation (MV) model. If the system is idle when it returns, it can take another break; if not, it continues to operate. This model is appropriate for complex systems and provides a high degree of flexibility. For instance, in communication systems, a server goes into an idle state (waiting period) if it finds no data to process. This procedure is repeated multiple times until new data is received if the lack of data continues. Several idle states are represented by this.

3.2.3 Synchronized Vacation

When every customer in the queue has the opportunity to leave at the same time, rather than leaving one by one, this is referred to as "simultaneous departure" in queueing theory. It is as though everyone decides whether to stay or leave based on a shared probability at the end of the waiting period. For example, in factories, all machines may stop at the same time for routine maintenance, and as soon as the maintenance is complete, they may all start up again.

3.3 M/M/1 Queue With Working Vacation

Model Description

We consider an M/M/1 queue with working vacation. Customers arrive according to a Poisson process with rate λ. The server serves customers at an exponential rate μ during the normal busy period, and at a lower rate μ_n during the working vacation period, where $\mu_n < \mu$. The server switches to a working vacation whenever the system becomes empty. FCFS (First Come, First Served) is adopted, meaning that customers are served in the order of their arrival. It is also assumed that service times during the working period and vacation times are exponentially distributed with rates η and θ, respectively. If at least N customers are waiting in the system at the moment a service is completed during the working period, let $Q(t)$ be the number of customers in the system at time t, and let $J(t)$ denote the state of the server[24].

$$J(t) = \begin{cases} 0, & \text{when the server stays in a WV period at time } t, \\ 1, & \text{when the server stays in non-vacation period at time } t. \end{cases}$$

3.3.0.1 Assumptions

- λ: arrival rate
- μ: service rate during normal period
- μ_v: Service rate during working vacation ($\mu_v < \mu$)
- η: Vacation ending rate (rate of vacation completion)
- $Q(t)$: the number of customers in the system at time t
- The service discipline is First-In, First-Out (FIFO).
- The system has an infinite capacity.

3.3.0.2 State Space

The process $\{(Q(t), J(t)),\, t \geq 0\}$ represents a two-dimensional continuous-time Markov chain with the state space

$$S = \{(0,0)\} \cup \{(k,j) : k \geq 1,\ j = 0,1\}.$$

Let $P_{n,j} = \lim_{t\to\infty} P\Big\{Q(t) = n,\ J(t) = j\Big\}$, for $n \geq 0$ and $j = 0,1$, denote the steady-state probabilities of the system states,The steady-state balance equations are given as follows.

3.3.0.3 Transition Rate Diagram

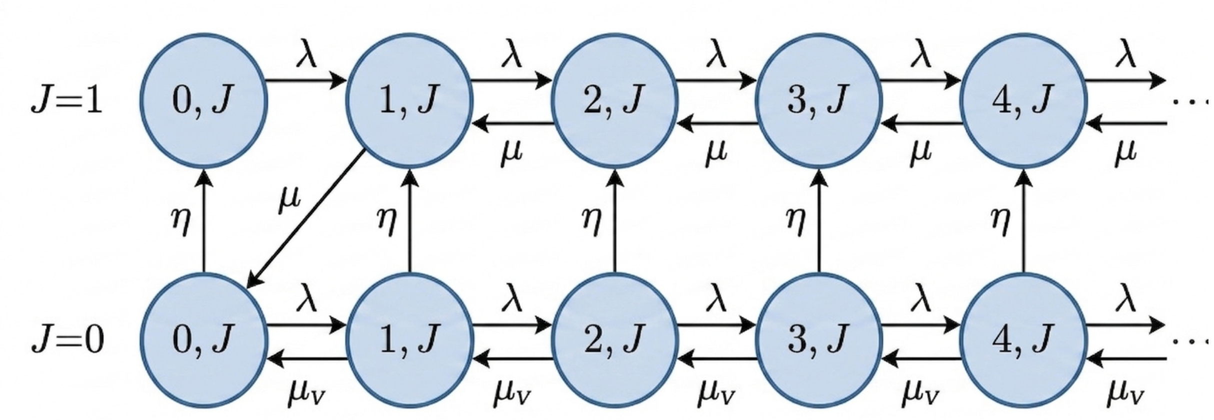

Figure 3.2: Generic M/M/1/WV queueing system

3.3.1 Steady-State Solution

Normal Service Period ($j = 1$)**.** The balance equations are given by

$$\begin{cases} \lambda P_{1,0} = \mu P_{1,1} + \eta P_{0,0}, & n = 0, \\ (\lambda + \mu) P_{1,n} = \lambda P_{1,n-1} + \mu P_{1,n+1} + \eta P_{0,n}, & n \geq 1. \end{cases} \tag{3.1}$$

Working Vacation Period ($j = 0$)**.** The balance equations are given by

$$\begin{cases} (\lambda + \eta) P_{0,0} = \mu_v P_{0,1}, & n = 0, \\ (\lambda + \mu_v + \eta) P_{0,n} = \lambda P_{0,n-1} + \mu_v P_{0,n+1}, & n \geq 1. \end{cases} \tag{3.2}$$

Consider the partial probability generating functions (PGFs) defined as follows:

$$G_j(z) = \sum_{n=0}^{\infty} P_{j,n} z^n, \quad 0 \leq j \leq 1. \tag{3.3}$$

and let

$$G_j'(z) = \frac{dG_j(z)}{dz} = \sum_{n=1}^{\infty} n P_{j,n} z^{n-1}, \qquad j = 0, 1$$

The partial generating functions $G_0(z)$ and $G_1(z)$, for $0 \leq z \leq 1$, are defined as[32]:

$$\begin{cases} G_0(z) = \sum_{n=0}^{\infty} P_{0,n} z^n, & j = 0 \\ G_1(z) = \sum_{n=0}^{\infty} P_{1,n} z^n, & j = 1 \end{cases} \qquad 0 \leq z \leq 1 \tag{3.4}$$

They satisfy the normalization condition:

$$\sum_{n=0}^{\infty} P_{0,n} + \sum_{n=0}^{\infty} P_{1,n} = 1 \tag{3.5}$$

This equation can be rewritten as:

$$G_0(z) = g_0(z) P_{0,0}, \quad G_1(z) = g_1(z) P_{0,0}. \tag{3.6}$$

$$g_0(z) = \frac{1-r}{1-rz}, \tag{3.7}$$

$$g_1(z) = \frac{\alpha(1-r)}{1-rz}, \tag{3.8}$$

where

$$r = \frac{\lambda + \mu_v + \eta - \sqrt{(\lambda + \mu_v + \eta)^2 - 4\lambda\mu_v}}{2\mu_v},$$

and

$$\alpha = \frac{\eta}{\lambda - \mu_v r}$$

is a parameter depending on the arrival rate λ and the working vacation service rate μ_v. Now, using the normalization condition, we obtain:

$$P_{0,0} = \frac{1}{1+\alpha} \tag{3.9}$$

This completes the evaluation of the steady-state probabilities.

Proof. The partial generating functions $G_0(z)$ and $G_1(z)$, for $0 \leq z \leq 1$, are defined as:

$$G_0(z) = \sum_{n=0}^{\infty} P_{0,n} z^n, \quad G_1(z) = \sum_{n=0}^{\infty} P_{1,n} z^n$$

From the structure of the model, the stationary probabilities follow a geometric distribution scaled by a normalization factor.

To satisfy both the functional equation and the normalization condition, we assume:

$$P_{0,n} = P_{0,0}(1-r)r^n, \quad 0 < r < 1$$

where r is the characteristic root of the quadratic equation:

$$\mu_v r^2 - (\lambda + \mu_v + \eta)r + \lambda = 0$$

which yields:

$$r = \frac{(\lambda + \mu_v + \eta) - \sqrt{(\lambda + \mu_v + \eta)^2 - 4\lambda\mu_v}}{2\mu_v}$$

Therefore, for the vacation state ($j = 0$):

$$G_0(z) = \sum_{n=0}^{\infty} P_{0,0}(1-r)r^n z^n = P_{0,0}\frac{1-r}{1-rz}$$

Similarly, for the regular state ($j = 1$), we have

$$P_{1,n} = \alpha P_{0,n}$$

where

$$\alpha = \frac{\eta}{\lambda - \mu_v r}$$

Thus:

$$G_1(z) = \alpha P_{0,0}\frac{1-r}{1-rz}$$

Finally:

$$G_0(z) = g_0(z)P_{0,0}, \quad G_1(z) = g_1(z)P_{0,0}$$

where

$$g_0(z) = \frac{1-r}{1-rz}, \quad g_1(z) = \frac{\alpha(1-r)}{1-rz}$$

□

3.3.2 Measures of Performance

Mean Number of Customers in the System

During Working Vacation: The mean number of customers in the system during a working vacation is given by:

$$E[\bar{N}_0] = G_0'(1)$$

Where:

$$E[\bar{N}_0] = \frac{P_{0,0}\, r}{(1-r)^2}$$

During Regular Service: The mean number of customers in the system during regular service is given by:

$$E[\bar{N}_1] = G_1'(1)$$

Where:

$$E[\bar{N}_1] = \frac{\alpha P_{0,0}\, r}{(1-r)^2}$$

Total Mean Number of Customers: The total mean number of customers in the system is:

$$\bar{N} = E[\bar{N}_0] + E[\bar{N}_1]$$

Where:

$$\bar{N} = \frac{(1+\alpha)P_{0,0}\, r}{(1-r)^2}$$

Mean Queue Length The mean number of customers waiting in the queue is:

$$\bar{N}_q = \bar{N} - P_{busy}$$

Mean Waiting Time in Queue: By Little's law, the mean waiting time in the queue is:

$$\bar{W}_q = \frac{\bar{N}_q}{\lambda}$$

Probability of Vacation State: The probability that the server is in vacation is:

$$P_{vac} = G_0(1)$$

Thus:

$$P_{vac} = P_{0,0}$$

Probability of Server Being Busy: The probability that the server is busy (normal service) is:

$$P_{busy} = 1 - P_{vac}$$

Mean Time in the System: The mean time a customer spends in the system is given by:

$$\bar{W} = \frac{\bar{N}}{\lambda}$$

3.3.3 Model Sensitivity Analysis

3.3.3.1 Impact of Arrival Rate on System Metrics

This section presents a numerical illustration accompanied by sensitivity analysis, with results obtained using the RStudio software. This study focuses on analyzing the impact of the arrival rate (λ) on the system performance measures, namely the probability of an empty system (P_{00}), the probability of working vacation (P_{vac}), the probability of server busy state (P_{busy}), the mean number of customers in the system ($\bar{N}$), and the mean waiting time ($\bar{W}$), illustrates the system performance metrics while keeping the parameter values fixed at $\mu = 4.0$, $\eta = 0.5$, and varying $\mu_v \in \{1.2, 1.6, 2.0\}$.

	$\mu_v = 1.2$				$\mu_v = 1.6$				$\mu_v = 2.0$			
λ	P_{vac}	P_{busy}	$\bar{N}$	$\bar{W}$	P_{vac}	P_{busy}	$\bar{N}$	$\bar{W}$	P_{vac}	P_{busy}	$\bar{N}$	$\bar{W}$
0.05	0.0292	0.9708	0.0309	0.6186	0.0210	0.9790	0.0223	0.4460	0.0165	0.9835	0.0175	0.3503
0.10	0.0578	0.9422	0.0651	0.6509	0.0409	0.9591	0.0463	0.4632	0.0318	0.9682	0.0359	0.3592
0.15	0.0859	0.9141	0.1028	0.6850	0.0604	0.9396	0.0723	0.4819	0.0466	0.9534	0.0556	0.3704
0.20	0.1134	0.8866	0.1442	0.7212	0.0792	0.9208	0.1005	0.5027	0.0610	0.9390	0.0772	0.3861
0.25	0.1403	0.8597	0.1899	0.7595	0.0976	0.9024	0.1309	0.5235	0.0748	0.9252	0.0999	0.3996
0.30	0.1667	0.8333	0.2400	0.8000	0.1154	0.8846	0.1636	0.5454	0.0882	0.9118	0.1241	0.4137
0.35	0.1924	0.8076	0.2950	0.8428	0.1327	0.8673	0.1989	0.5684	0.1011	0.8989	0.1500	0.4284
0.40	0.2175	0.7825	0.3552	0.8881	0.1495	0.8505	0.2372	0.5930	0.1136	0.8864	0.1775	0.4438
0.45	0.2420	0.7580	0.4211	0.9359	0.1658	0.8342	0.2785	0.6190	0.1256	0.8744	0.2070	0.4601
0.50	0.2658	0.7342	0.4931	0.9863	0.1816	0.8184	0.3235	0.6470	0.1374	0.8626	0.2384	0.4768
0.55	0.2890	0.7110	0.5716	1.0393	0.1970	0.8030	0.3725	0.6773	0.1488	0.8512	0.2721	0.4947
0.60	0.3115	0.6885	0.6571	1.0952	0.2119	0.7881	0.4259	0.7098	0.1598	0.8402	0.3082	0.5137
0.65	0.3333	0.6667	0.7500	1.1538	0.2263	0.7737	0.4842	0.7449	0.1705	0.8295	0.3470	0.5338
0.70	0.3545	0.6455	0.8508	1.2154	0.2403	0.7597	0.5478	0.7826	0.1808	0.8192	0.3888	0.5554
0.75	0.3750	0.6250	0.9600	1.2800	0.2538	0.7462	0.6174	0.8232	0.1908	0.8092	0.4338	0.5784
0.80	0.3948	0.6052	1.0781	1.3476	0.2670	0.7330	0.6937	0.8671	0.2005	0.7995	0.4826	0.6033
0.85	0.4140	0.5860	1.2055	1.4182	0.2798	0.7202	0.7774	0.9146	0.2099	0.7901	0.5356	0.6301
0.90	0.4325	0.5675	1.3428	1.4920	0.2922	0.7078	0.8694	0.9660	0.2191	0.7809	0.5932	0.6591
0.95	0.4503	0.5497	1.4904	1.5688	0.3043	0.6957	0.9708	1.0219	0.2280	0.7720	0.6560	0.6905
1.00	0.4675	0.5325	1.6488	1.6488	0.3161	0.6839	1.0829	1.0829	0.2366	0.7634	0.7247	0.7247
1.05	0.4841	0.5159	1.8185	1.7319	0.3275	0.6725	1.2072	1.1497	0.2450	0.7550	0.8000	0.7619

Table 3.1: System Performance Metrics and Probabilities vs λ for $\mu_v \in \{1.2, 1.6, 2.0\}$ with $\eta = 0.5$

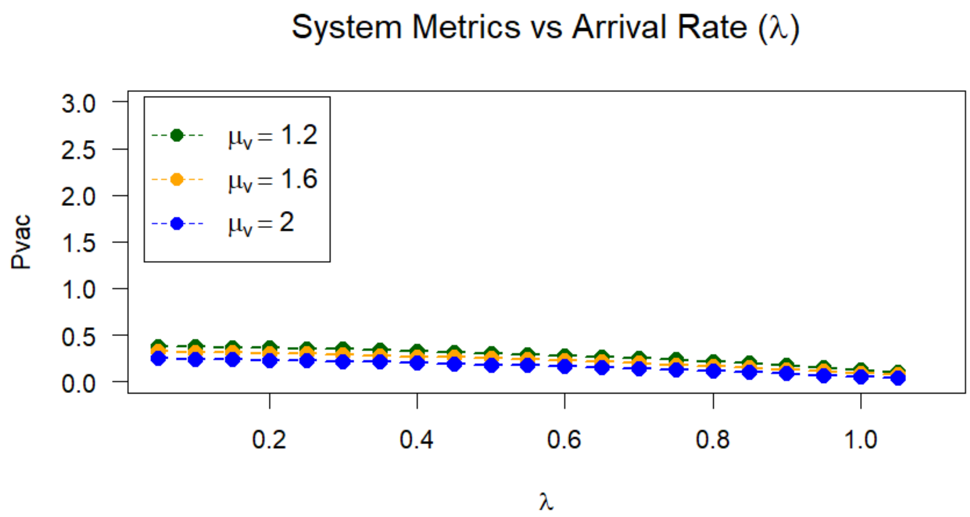

Figure 3.3: Probability of the server being on vacation P_{vac} vs λ for different values of μ_v.

Figure (3.3) The probability of the server being on vacation P_{vac} decreases gradually with the increase in the arrival rate λ. This decrease is due to the fact that as customer arrivals increase, the normal service period lengthens to reduce the workload, which in turn reduces the vacation time. For a fixed arrival rate λ, we observe that an increase in the service rate during vacation μ_v leads to lower values of P_{vac}. The blue curve ($\mu_v = 2$) appears at the bottom, while the green curve ($\mu_v = 1.2$) rises to the top. The interpretation is that a higher vacation service rate pulls the server quickly back to the normal mode, thus reducing the probability of it remaining on vacation.

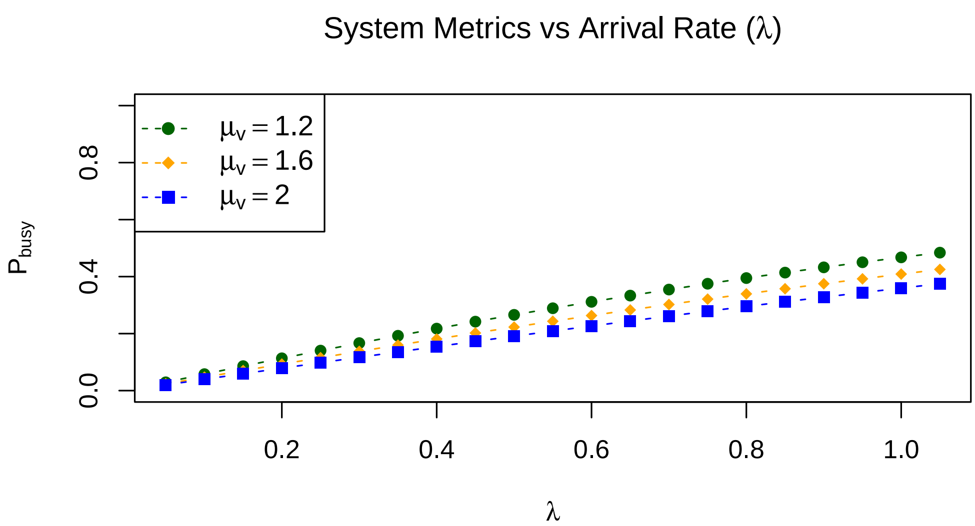

Figure 3.4: Probability of the server being busy P_{busy} vs λ for different values of μ_v.

Figure (3.4) The probability of the server being busy P_{busy} increases gradually and significantly with the increase in the arrival rate λ. This increase is natural; as the customer arrival rate rises, the workload pressure in the system increases and requests accumulate, which raises the probability of the server being busy. For a fixed arrival rate λ, we observe that an increase in the service rate during vacation μ_v leads to lower values of P_{busy}. The blue curve ($\mu_v = 2$) appears at the bottom, while the green curve ($\mu_v = 1.2$) rises to the top. The interpretation is that higher performance speed during vacation contributes to quickly clearing the queue, thus reducing the probability of the server remaining busy.

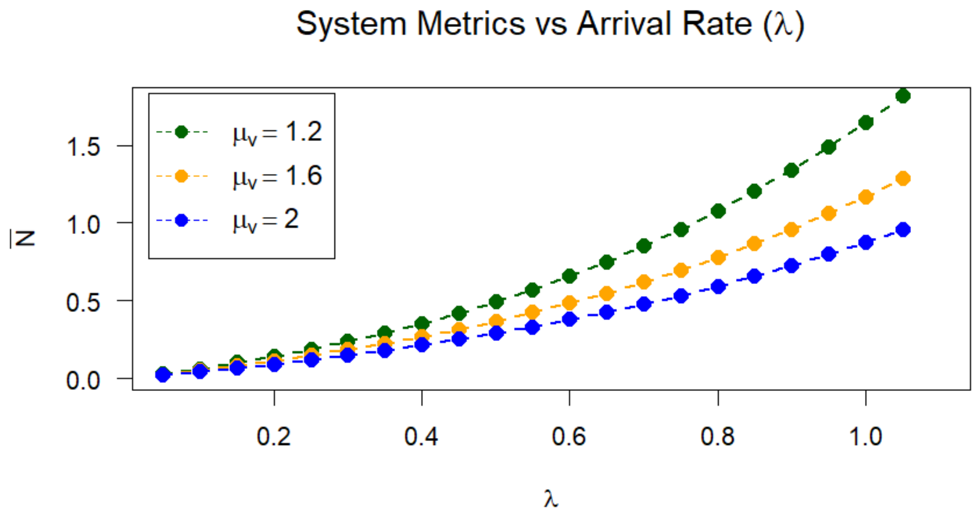

Figure 3.5: $\bar{N}$ as a function of λ for different values of μ_v.

Figure (3.5) shows that as the arrival rate λ increases, the average system length $\bar{N}$ increases for a fixed value of the service rate μ_v in the normal service state. This is due to the decrease in the average time between customer arrivals. For a fixed value of λ, as μ_v increases, the average service time decreases, which leads to a reduction in the expected system length.

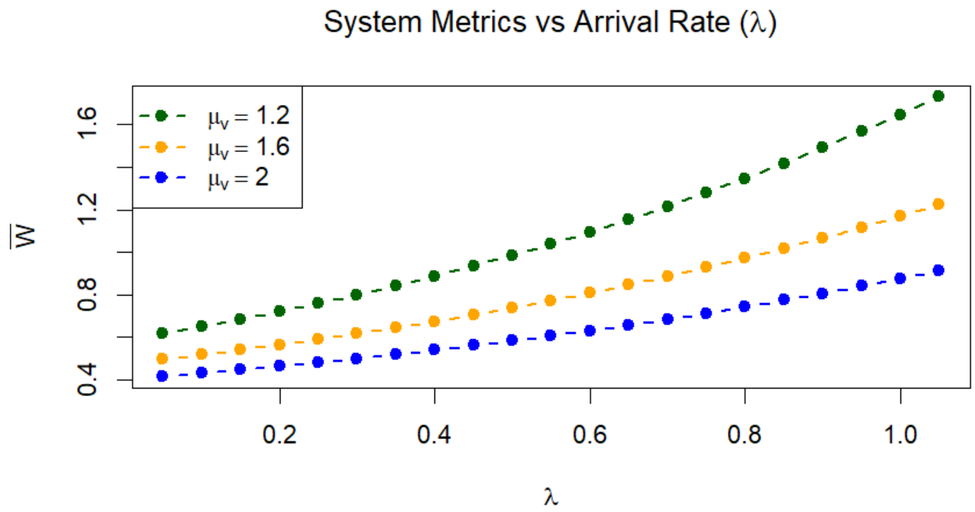

Figure 3.6: $\bar{W}$ as a function of λ for different values of μ_v.

Figure (3.6) shows that the average waiting time $\bar{W}$ increases as the arrival rate λ increases for a fixed value of the vacation service rate μ_v. This is due to the decrease in the time between customer arrivals, which increases congestion. For a fixed value of λ, as μ_v increases, the average service time decreases, leading to a reduction in the expected waiting time.

General Conclusion

In this research, we have conducted a detailed study of the vacation policy in queueing systems involving temporary periods of inactivity, highlighting its theoretical foundations and practical significance. By extending the traditional concept of server vacation periods, the key findings demonstrate that this policy allows for a substantial reduction in the server's energy consumption while maintaining an acceptable level of service.

In these results, we have further enabled the evaluation of essential performance metrics, such as the mean waiting time, the average number of customers in the system, and the probabilities of the server being in a vacation state or a busy state. This approach provides a comprehensive overview of how a working vacation policy influences the overall efficiency of such systems.

Furthermore, the scope of this study can be extended to systems with infinite server capacities. This work significantly contributes to the development of Markovian models for systems incorporating vacation periods. It may also provide new and insightful perspectives for a more accurate representation of real-world systems.

Bibliography

[1] Arnaud GUYADER, *Jump Markov Processes*, University of Rennes, 2006/2007.

[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] Azhagappan, A. Transient behavior of a Markovian queue with working vacation, variant reneging and a waiting server. TOP, 27, 351–370, 2019.

[4] A. Gomez-Corral 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, vol. 3, pp. 37–47, 2000.

[5] Adan, I. J. B. F., and Resing, J. A. C., *Queueing Theory*, Department of Mathematics and Computing Science, Eindhoven University of Technology, 2001.

[6] Berkane, K., "Analysis of a Queueing System with General Service Time Distribution, Server Vacations and Impatient Customers", Master's Thesis, University of Saida – Dr Moulay Tahar, p.13, (2021).

[7] B. K. Kim and D. H. Lee, "The M/G/1 queue with disasters and working breakdowns," *Applied Mathematical Modelling*, Vol. 38, no. 5–6, pp. 1788–1798, 2014.

[8] Bhat, U. N., *An Introduction to Queueing Theory: Modeling and Analysis in Applications*, 2nd ed., Birkhäuser, Boston, 2015.

[9] Caumel, Yves, *Probability and Stochastic Processes*, Springer-Verlag France, 2011.

[10] Divya, K. and Indhira, K., "A Literature Survey on Queueing Model with Working Vacation," *RT&A*, Vol. 19, No. 1(77), pp. 40–47, March 2024.

[11] D. Fiems, *Queues with Working Vacations: A Survey*, Mathematics, Vol. 13, Article 1894, 2025.

[12] D.-A. Wu and H. Takagi, "M/G/1 Queue with Multiple Working Vacations," *Performance Evaluation*, vol. 63, no. 7, pp. 654–681, 2006.

[13] Fezza, Kheira, *Queueing System with Balking: M/M/s Model with Balking*, University Dr Tahar Moulay - Saïda, Master's Thesis, pp.11,17,18,19, (2016/2017).

[14] Fezza, K., *Queueing System with Balking: M/M/s Model with Balking*, Master's Thesis in Mathematics, University Dr Tahar Moulay – Saïda, p.11, (2016/2017).

[15] Florin Avran, *Markov and Lévy Processes, Queueing, Actuarial Science and Reliability*, 2009–2010.

[16] Guérin, Hélénec, *Course Handout (Mat-3017 - Stochastic Processes)*, (2012).

[17] Gomez-Corral, A. and Ramalhoto, M. F., *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.

[18] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA, 1999.

[19] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, Wiley, New York, 1975, 1984.

[20] Hammache, Saouaguia, *Queueing Systems with Discouragement*, University of Saïda - Dr Moulay Tahar, p.8, 2017/2018.

[21] J. H. Li, N. S. Tian, Z. G. Zhang and H. P. Luh, "Analysis of the M/G/1 queue with exponentially working vacations—A matrix analytic approach," *Queueing Systems*, Vol. 61, no. 2, pp. 139–166, 2009.

[22] Hafsi, K., *Study of a Queueing System with Heterogeneous Servers, Balking and Customer Abandonment*, Master's Thesis, University of Saïda – Dr Moulay Tahar, Examples used on pp.10-11-12-15, (2018).

[23] M. O. Abou El-Ata and A. M. A. Hariri, *The M/M/c/N queue with balking and reneging*, Computers and Operations Research, 19 (1992), No. 13, 713–716.

[24] Majid, S., and Manoharan, P. (2019). *Analysis of an M/M/1 Queue with Working Vacation and Vacation Interruption*. Applications and Applied Mathematics: An International Journal (AAM), Vol. 14, No. 1, pp. 19–33, June 2019.

[25] M. Kobielnik and W. M. Kempa, *Output stream analysis in a queueing model with working vacation mechanism as a power reduction strategy*, PLoS One, vol. 21, no. 1, e0341422, 2026.

[26] Keddari, Rabie Abdel-Mohcen, *Markovian Queueing System with Single Vacations*, University Dr Tahar Moulay - Saïda, Discipline: Mathematics, Specialization: ASSPA, Master's Thesis, p.8, (2019/2020).

[27] Kadi, Mokhtar, *Handout on Introduction to Queueing Systems: Course and Solved Exercises*, University Dr Tahar Moulay – Saïda, Faculty of Sciences, Department of Mathematics, pp.12-13.

[28] Khintchine, A. Y. (1969). *Mathematical Methods in the Theory of Queueing*. 2nd edition, Hafner Publishing Company, New York. (First edition published by Griffin, London, 1960; Russian original edition, 1955), pp. 859–866.

[29] Kadi, Mokhtar, *Introduction to Queueing Systems: Course and Solved Exercises*, Course handout, University Dr Tahar Moulay Saida, p28-29.

[30] Krishnamoorthy, A. and Sreenivasan, C., "An M/M/2 Queueing System with Heterogeneous Servers Including One with Working Vacation," *International Journal of Stochastic Analysis*, 2012.

[31] Kim, J.D., Choi, D.W., Chae, K.C. Analysis of queue-length distribution of the M/G/1 queue with working vacations (M/G/1/WV). In: Proceedings of Hawaii International Conference on Statistics and Related Fields, June 5–8, 2003.

[32] Kadi, M., Bouchentouf, A. A., and Yahiaoui, L. (2025). *Modèle M/M/c avec feedback, unique et multiple vacances synchronisées, impatience et retention des clients abandonés*. Presentation, Laboratory of Stochastic Models, Statistics and Applications (LMSSA), University of Saida - Dr. Moulay Tahar.

[33] Li, J., Liu, W., Tian, N. The discrete-time GI/Geo/1 queue with multiple working vacations. Queueing Systems, submitted for publication.

[34] L. D. Servi and S. G. Finn, "M/M/1 queues with working vacations (M/M/1/WV)," *Performance Evaluation*, vol. 50, no. 1, pp. 41–52, 2002.

[35] Melle Aya BOUDALI, *Different Types of Discrete-Time Queues*, University of Saïda - Dr Moulay Tahar, Faculty of Sciences, Master's Thesis, Academic year, pp.13-14 (2022/2023).

[36] Mentefa, Abdelmadjid, *Markovian Queueing System with Multiple Vacations and Group Service*, University of Saïda - Dr Moulay Tahar, Faculty of Sciences, Department of Mathematics, p.4, (2022).

[37] Moulay Hachemi, *Queueing Systems and Applications*, Course handout, page 12, 2014/2015.

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

[39] N. Tian and Z. G. Zhang, *Vacation Queueing Models: Theory and Applications*, Vol. 93, Springer Science and Business Media, 2006.

[40] Petito, M. (2010). *Introduction to Network Modeling*, p. 21, October 26, 2010.

[41] Richard Newman, *Elementary Queuing Theory Notes*, University of Florida, January 1999.

[42] Rabie Abdel-Mohcen Keddari, *Markovian Queueing System with Single Vacation*, Master's Thesis, Université Dr Tahar Moulay - Saïda, September 2020, pp. 27–28.

[43] Stordahl, K., "The History Behind the Probability Theory and the Queuing Theory," *Telektronikk*, vol. 103, no. 2, pp. 123–140, 2007.

[44] V. M. Chandrasekaran, K. Indhira, M. C. Saravanarajan, and P. Rajadurai, "A Survey on Working Vacation Queueing Models," *International Journal of Pure and Applied Mathematics*, Vol. 106, No. 6, pp. 33–41, 2016.

[45] Zakhar Kabluchko, *Stochastic Processes (Stochastik II)*, University of Ulm, Institute of Stochastics, (2013–2014).