People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Saïda - Dr. Moulay Tahar

Faculty of Mathematics, Computer Science and Telecommunications.

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 : Stochastic Analysis, Statistics of Processes and Applications

by

Ms Hadjira BOUDIA [1]

Under the supervision of

Dr. L.YAHIAOUI

Theme:

Advanced Discrete-Time Queueing Models.

Defended on 11/06/2026 before the jury composed of

Dr. H. ZOUAOUI	University of Saïda Dr. Moulay Tahar	President
Dr. L. YAHIAOUI	University of Saïda Dr. Moulay Tahar	Supervisor
Dr. M. KADI	University of Saïda Dr. Moulay Tahar	Examiner

Academic year: 2025/2026

[1] e-mail: boudiahadjira7@gmail.com

Acknowledgments

First and foremost, thank ***Almighty God****, who granted me the strength, patience, and perseverance to bring this graduation project to completion.*

I would like to express my deepest and most sincere gratitude to my supervisor, ***Professor YAHIAOUI Lahcene****, for invaluable guidance, unwavering support, insightful advice, and constant encouragement throughout the preparation of this thesis. His dedication has been a beacon throughout this journey.*

I also extend my warmest thanks to all the professors of the ***Mathematics Department*** *for their passion, dedication, and for shaping my academic education during my university years. Their teaching has left an indelible mark on my mind.*

My sincere appreciation goes as well to the members of the ***examination committee*** *for accepting the task of reviewing and evaluating this work. Their thoughtful engagement honors this achievement.*

Finally, words cannot capture the depth of my gratitude to my ***parents and family****. Their unconditional support, boundless encouragement, and constant love have been the quiet foundation upon which all my studies rest. To them, I dedicate this work.*

$\leq$ To all, thank you $\geq$

Dedications

I dedicate this humble work

To my **dearest parents,** no dedication could ever fully express the love, respect, and gratitude I have for you. Thank you for your countless sacrifices, your unwavering support, your constant encouragement, and your prayers that have always given me the strength to move forward.
May God protect you and grant you a long life filled with happiness and health.

To my dear brothers **Mohamed Amine** and **Mohamed Miloud**, and my dear sister **Imene** for their affection, their moral support, and their steadfast encouragement. Thank you for always being by my side through difficult times as well as moments of joy.

To my dear niece, **Djinane Kaouather,** a true source of joy and happiness! I wish her a future full of success, health, and happiness.

To all my family and to all the people who are dear to me.

Hadjira BOUDIA

Abstract

This thesis is devoted to the study of advanced discrete-time queueing models, which are widely used to model and analyze modern service systems operating in time-slotted environments such as communication networks and computer systems.

The work begins with a presentation of the fundamental theoretical concepts necessary for the analysis of queueing systems.

Next, we investigate classical discrete-time queueing models based on Markov chains, including the Geo/Geo/1 and Geo/Geo/s models. For these models, key performance measures such as the stationary distribution, server utilization, and mean queue length are derived and analyzed.

The main contribution of this thesis lies in the study of advanced queueing models with retrial phenomena, namely the Geo/Geo/s with retrial and Geo/G/1 with retrial models. These models are more realistic because they account for customers who, upon finding the server busy, leave and retry after a random delay.

Finally, a numerical approach has been developed to illustrate the theoretical results and highlight their practical relevance for decision-making and system performance evaluation.

This work provides a comprehensive understanding of discrete-time queueing systems, from basic concepts to advanced models, and emphasizes their importance in modeling real-world stochastic systems.

Contents

Introduction

Queues are a phenomenon we experience in reality, or more specifically, in our daily lives. We find them in many places, such as hospitals, banks, airports, transportation hubs, etc. This leads us to study them from a mathematical perspective.

The knowledge of queuing theory is attributed to the Danish engineer Erlang in 1909. He conducted experiments on the problem of excessive phone calls, where callers experienced delays because operators could not handle incoming requests as quickly as they arrived. Erlang addressed the problem by calculating the delay for a single operator at that time *[4]*. In 1917, research on this problem was revisited, but for more than one operator, and thus, queuing theory emerged *[3]*. Its use extended to solving many similar administrative problems, and this theory is now widely used in all production facilities as a mathematical tool to assist management in making the most appropriate decisions about these types of problems.

Queuing can be defined as the mathematical equations and relationships that can be employed to determine the operating characteristics of a specific waiting line.

It can also be defined as the theory concerned with developing the necessary mathematical methods to solve problems related to situations characterized by bottlenecks, or the formation of waiting lines as a result of units requesting service and waiting their turn to receive it, provided that arrival at the service location is random and follows a specific distribution, and the service time for each unit can take a random form. It also measures the service center's ability to achieve its intended purpose by calculating the mean waiting time to receive the service.

Discrete-time queuing models are also a modern extension of the classical queuing theory developed by the engineer Erlang *[5]*. With the development of digital networks and computer systems, discrete-time models have become essential for representing systems in which events occur at discrete time intervals, such as communication networks and data transmission systems *[6]*.

The objective of this work is to present advanced discrete-time queueing models.
This work consists of four chapters. In the first chapter, we introduced the elements we will be using in this work. This chapter gives the theoretical background needed to study discrete-time queueing systems. First, it introduces a review of probability theory, covering random variables, Bernoulli and geometric distributions, and the memoryless property. Next, the main concepts of stochastic processes are presented, such as Markov chains, birth and death processes, and Bernoulli processes.

The chapter then presents the basic elements of queueing systems, such as arrival flows, queue structures, system capacity, and service disciplines. Finally, Kendall notation *[5]*, Little's law, and the main performance measures of queueing systems are presented, since they are the essential tools for the analysis and evaluation of queueing models.

In the second chapter, we introduced the classical Markovian models of discrete-time queues *[11]*. We began by introducing discrete-time Markovian queues, and then discussed discrete-time queueing models, such as the Geo/Geo/1 and Geo/Geo/s models. The Geo/Geo/1 queueing model is a mathematical model used to study queues in a service system *[10]*. It is based on two main assumptions. Customer arrivals follow a Bernoulli process, and service time is geometrically distributed. And for the Geo/Geo/s queueing model *[13]*, this model assumes that customer arrivals follow a Bernoulli distribution and that service times follow a geometric distribution with multiple servers.

Third chapter: We will discuss the advanced model of discrete-time queues. We have studied advanced queueing models, and after all the advanced models, by choosing queueing models with recall, we will consider the following two models: the Geo/Geo/s model with recall *[12]*, and the Geo/G/1 model with recall *[16]*.

Fourth chapter: a simulation was carried out for the models that were studied in the second and third chapters.

Chapter 1

Theoretical Foundations of Discrete-Time Queueing Systems

Introduction

The fundamental mathematical tools in the field of queuing theory are stochastic processes, which refer to random systems whose behavior over time is studied. More specifically, Markov chains are central to this research, as their zero-memory property stipulates that the future state depends solely on the present state.

Discrete-time systems are particularly relevant for this type of modeling because their behavior is observed at well-defined time points, a characteristic shared by many modern systems, such as computer networks and digital systems. Discrete-time queues thus allow us to measure system performance in terms of queue length, waiting time, and utilization rate.

The objective of this chapter is to describe the theoretical foundations of discrete-time queueing systems by introducing the concepts of stochastic processes and Markov chains before presenting the basic concepts of queueing systems. This serves as a cornerstone, laying the groundwork for the treatment of classical and advanced models in the following chapters.

1.1 Probability Theory Review

1.1.1 Random variables

Definition 1.1.1.1.

A random variable is said to have a **discrete distribution** if there exists an at most countable set $T \subset \mathbb{R}$ such that

$$P(X = x) > 0 \quad \text{for all } x \in T$$

and

$$\sum_{x \in T} P(X = x) = 1.$$

Here, T is called the **support** of X. The function

$$p_X : \mathbb{R} \to [0, 1]$$

defined by

$$p_X(x) = \begin{cases} P(X = x) & \text{if } x \in T \\ 0 & \text{otherwise} \end{cases}$$

is called the **probability mass function** (PMF) of X.

And,

$$F_X(x) = \sum_{\substack{y \in T \\ y \leq x}} p_X(y) \quad \text{and} \quad p_X(x) = F_X(x) - F_X(\bar{x}).$$

1.1.2 Bernoulli Distribution

Definition 1.1.2.1. A random variable X such that $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ is said to be a Bernoulli random variable with parameter p. Note $\mathbb{E}X = p$ and $\mathbb{E}X^2 = p$, so $\operatorname{Var} X = p - p^2 = p(1 - p)$.

We denote such a random variable by $X \sim \text{Bern}(p)$.

1.1.3 Geometric Distribution

Definition 1.1.3.1. A random variable X has the geometric distribution with parameter λ, $0 < \lambda < 1$, if

$$\mathbb{P}(X = k) = (1 - \lambda)^{k-1}\lambda \quad \text{for } k = 1, 2, \ldots$$

Using a geometric series sum formula, we see that.

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} (1 - \lambda)^{k-1}\lambda = \frac{1}{1 - (1 - \lambda)}\lambda = 1.$$

In i.i.d. Bernoulli (p) trials, let X be the first time we observe a success; then X is a geometric random variable.
There is another form of geometric distribution, which is $(1 - \lambda)^k\lambda$

For example: if we toss a coin over and over and X is the first time we get a head, then X will have a geometric distribution. To see this, to have the first success occur on the k nth trial, we have to have $k - 1$ failures in the first $k - 1$ trials and then a success. The probability of that is $(1 - p)^{k-1}p$.

1.1.4 memoryless

The memorylessness property is demonstrated by some probability distributions in the areas of probability and statistics, such as the geometric and exponential distributions. For these two examples of probability distributions, the phrase "memoryless" is attached.

Furthermore, both distributions exhibit the following characteristic: when comparing the probabilities associated with the customer's waiting time at the server from the time of arrival until the time the server is open and an arbitrary amount of time has elapsed since the server opened, both distributions exhibit the property of memorylessness.

Of all discrete probability distributions, only the geometric probability distribution exhibits the property of memorylessness. Therefore, when we compare the geometric distribution to other discrete probability distributions, we can say that 'memorylessness' is one property of the geometric probability distribution.

Proposition 1.1.4.1.

Let X be a random variable following a geometric distribution with parameter λ, denoted by $X \sim \text{Geometric}(\lambda)$. Then X satisfies the *memoryless property*. For all integers $m \geq 0$ and $n \geq 0$, we have:

$$P(X > m + n \mid X > n) = P(X > m). \tag{1.1}$$

Interpretation.

If the first n trials result in failures, the probability that the next m trials are also failures is equal to the probability that the first m trials are failures. This means that the geometric distribution has no memory. The past does not influence the future.

Proof.

Step 1: Geometric Series Formula

$$1 + q + q^2 + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q}. \tag{1.2}$$

Step 2: Computation of $P(X > n)$

$$P(X > n) = 1 - P(X \leq n). \tag{1.3}$$

Since

$$P(X \leq n) = \sum_{k=1}^{n} P(X = k),$$

and for a geometric distribution

$$P(X = k) = \lambda(1 - \lambda)^{k-1},$$

we obtain

$$P(X > n) = 1 - \sum_{k=1}^{n} \lambda(1 - \lambda)^{k-1}.$$

Factoring ou:λ:

$$P(X > n) = 1 - \lambda \left[1 + (1 - \lambda) + \cdots + (1 - \lambda)^{n-1}\right].$$

Using the geometric series formula:

$$P(X > n) = 1 - \lambda \frac{1 - (1 - \lambda)^n}{1 - (1 - \lambda)}.$$

Since $1 - (1 - \lambda) = \lambda$, we get:

$$P(X > n) = (1 - \lambda)^n. \tag{1.4}$$

Step 3: Conditional Probability

$$\begin{aligned} P(X > m + n \mid X > n) &= \frac{P(X > m + n)}{P(X > n)} & (1.5) \\ &= \frac{(1 - \lambda)^{m+n}}{(1 - \lambda)^n} & (1.6) \\ &= (1 - \lambda)^m. & (1.7) \end{aligned}$$

Since

$$(1 - p)^m = P(X > m),$$

We conclude that

$$P(X > m + n \mid X > n) = P(X > m). \tag{1.8}$$

1.2 Stochastic Processes

Definition 1.2.1. A random process is a family of random variables or vectors $(X_t)_{t \in T}$, indexed by the set T of countable or continuous times, defined on a probability space $(\Omega, \mathcal{F}, P)$ and with values in a state space E. The random variable or vector X_t describes the state of the process at time t. Each occurrence ω is matched with the trajectory of the realization $(X_t(\omega))_{t \in T}$ of the process defined by the application:

$$t \in T \longmapsto X_t(\omega) \in E.$$

Explanation of Notation

- $\mathbf{T}$: Index set (time) - can be countable (discrete-time) or continuous
- $(\boldsymbol{\Omega}, \mathcal{F}, \mathbf{P})$: Probability space
  - Ω: Sample space
  - $\mathcal{F}$: σ-algebra of events
  - P: Probability measure
- $\mathbf{E}$: State space (e.g., $\mathbb{R}$, $\mathbb{R}^d$, or more general spaces)
- $\mathbf{X_t}(\omega)$: Value of the process at time t for outcome ω
- $(\mathbf{X_t}(\omega))_{\mathbf{t \in T}}$: Entire trajectory (path) for a fixed ω

1.2.1 Counting Process

Definition 1.2.1.1. A stochastic process $\{N_t, t \in \mathbb{N}_0\}$ is called a **discrete-time counting process** if it N_n represents the total number of events that have occurred up to the discrete time t. It satisfies the following properties:

1. $N_t \in \mathbb{N}_0$ for all t;
2. N_t is non-decreasing in t (i.e., $N_t \leq N_{t+1}$);
3. For all $m < t$, the difference $N_t - N_m$ is the number of events occurring between time $m + 1$ and time t inclusive.

Associated Process in Discrete Time

The **arrival times** T_k (in discrete time) are the instants at which the k-th event occurs. We always have:

$$W_k = T_k - T_{k-1}$$

where W_k is the **discrete interarrival time** (a non-negative integer).

1.2.2 Processes with independent increments

Definition 1.2.2.1. Let $\{X_t, t = 0, 1, 2, \dots\}$ be a discrete-time random process. We say that it has **independent increments** if the random variables

$$X_1 - X_0,\ X_2 - X_1,\ X_3 - X_2,\ \dots$$

They are independent of each other.

That is, the change between each time step and the next is independent of the other changes.

1.2.3 Processes with stationary increments

Definition 1.2.3.1.

Let $\{X_t, t = 0, 1, 2, \dots\}$ be a discrete-time random process. We say that it has **stationary increments** if the probability distribution of the increment $X_t - X_m$ depends only on the length of the interval $t - m$ and not on its location m.

That is, for any m, n, k such that $m < t$ and $k > 0$, we have:

$$X_t - X_m \overset{d}{=} X_{t+k} - X_{m+k}$$

Where $\overset{d}{=}$ means "equal in distribution".

1.2.4 Markov Chains

Definition 1.2.4.1. Consider a stochastic process with values in a finite or countable state space E, $\{X(t); t \geq 0\}$. $\{X(t); t \geq 0\}$ is a Markov process if $\forall s, t \in \mathbb{R}^+$, $\forall i, j, l_u \in E$:

$$\mathbb{P}(X_{t+s} = j \mid X_t = i, X_u = l_u, u < t) = \mathbb{P}(X_{t+s} = j \mid X_t = i). \tag{1.9}$$

In other words, a process is Markovian if the conditional distribution of the future variable X_{t+s} given the present state X_t and the history up to time t focuses solely on the present and does not depend on the past. If the conditional specification $\mathbb{P}(X_{t+s} = j \mid X_t = i)$ is independent of the time values, the Markov process is described as homogeneous, so that $\mathbb{P}(X_{t+s} = j \mid X_t = i) = p_{ij}(s)$.

1.2.5 Discrete-Time Markov Chains

Definition 1.2.5.1 *[20]* A discrete-time Markov chain is a sequence of random variables $(X_t)_{t\geq 0}$ taking values in a countable (finite or infinite) state space (usually represented by the integers 0, 1, 2, ...) such that

$$\mathbb{P}(X_{t+1} = i_{t+1} | X_t = i, ..., X_0 = i_0) = \mathbb{P}(X_{t+1} = i_{t+1} | X_t = i).$$

.

Definition 1.2.5.2. (Transition probability) *[15]* We define the transition probability from state i to state j between times t and t + 1 by the following quantity:

$$P_{ij}(t) = \mathbb{P}(X_{t+1} = j | X_t = i), \quad \forall i, j \in E.$$

Where $P_{ij}(t)$ is the probability that the system is in state j at time t + 1, given that it was in state i at time t.

1.2.6 Transition matrix

Definition 1.2.6.1. For a time-homogeneous Markov chain, the probability

$P(X_{t+1} = y \mid X_t = x)$ does not depend on time. This probability is called the transition probability (from x to y) and is denoted by

$$p_{xy} = P(X_{t+1} = y \mid X_t = x) = P(X_1 = y \mid X_0 = x).$$

The set of these transition probabilities defines the transition matrix denoted by P:

$$P = (p_{xy})_{x,y \in E \times E}.$$

Proposition 1.2.6.1. If X is a Markov chain with transition matrix P, then for all $n \geq 0$ and $(x_0, \ldots, x_t) \in E^{t+1}$, we have

$$P(X_0 = x_0, \ldots, X_t = x_t) = P(X_0 = x_0) \prod_{i=1}^{t} p_{x_{i-1}x_i}.$$

Proof.

We use recurrence on $t \geq 0$. For $t = 0$, there is nothing to prove (with the convention $\prod_{i=1}^{0} = 1$). Suppose $t \geq 1$. Then

$$\begin{aligned} P(X_0^{t+1} = x_0^{t+1}) &= P(X_{t+1} = x_{t+1} \mid X_0^t = x_0^t) P(X_0^t = x_0^t) \\ &= p_{x_t x_{t+1}} \left(P(X_0 = x_0) \prod_{i=1}^{t} p_{x_{i-1} x_i} \right) \end{aligned}$$

where the last equality comes from the Markov property for the term $p_{x_t x_{t+1}}$ and from the recurrence hypothesis for the second term $P(X_0 = x_0) \prod_{i=1}^{t} p_{x_{i-1} x_i}$. The proposition is proven.

Proposition 1.2.6.2. shows that the transition matrix P does not characterize the distribution of (X_t): the initial condition $(P(X_0 = x))_{x \in E}$ must also be specified. Subsequently, we will denote the initial condition by π_0, with $\pi_0 = (\pi_0(x))_{x \in E}$ and $\pi_0(x) = P(X_0 = x)$ and π_t the distribution of X_t, i.e., $\pi_t = (\pi_t(x))_{x \in E}$ with $\pi_t(x) = P(X_t = x)$.

1.2.7 Chapman-Kolmogorov equation

The transition probabilities in t steps are completely determined by the transition probabilities in one step, and therefore by the transition matrix. This is what the Chapman-Kolmogorov equations express.

Proposition 1.2.7.1. *[15]* For all i, j, and for all $0 < k < t$, we have

$$P_{ij}^{(t)} = \sum_{k \in E} P_{ki}^{(t-1)} P_{kj}^{(1)}.$$

Demonstration.

$$\begin{aligned} P_{ij}^{(t)} &= P(X_t = j \mid X_0 = i); \\ &= \sum_{k \in E} P(X_t = j, X_{t-1} = k \mid X_0 = i) \quad (t \geq 2); \\ &= \sum_{k \in E} P(X_t = j \mid X_{t-1} = k, X_0 = i) P(X_{t-1} = k \mid X_0 = i); \\ &= \sum_{k \in E} P_{ik}^{(t-1)} P_{kj}^{(1)}. \end{aligned}$$

In other words, this relationship can be expressed using the following matrix multiplication formula: $P^{(t)} = P^{(t-1)} P^{(1)}$.

1.2.8 Birth and Death Processes

Birth and death processes (N-M) are processes such that a single birth can occur at any time, and the birth rate depends on the number of customers present in the system. Similarly, a single death can occur at any time, and mortality depends on the customer population. These processes are Markovian. We will note later that a birth and death process can only change the population by one unit at a time, so its Markov chain is tridiagonal.

The birth and death process is extremely fundamental in queueing theory. The discrete-time birth-death process has not received much attention in the literature. This is partly because it does not have the simple structure that makes it so general for a class of queues as a counterpart to continuous time. We assume that the birth process follows a state-dependent geometric distribution, and the death process follows a state-dependent geometric distribution. In what follows, we develop the model more explicitly.

1.2.9 Birth and Death Processes In Discrete Time

The birth process is a Bernoulli process, while the death process follows the geometric distribution. Let $p_i = P(\text{a birth occurs})$, $\overline{p_i} = 1 - p_i$, $i \geq 0$, and let $\lambda_i = P(\text{a death occurs})$, $\overline{\lambda_i} = 1 - \lambda_i$, $i \geq 0$.

If we now define a Markov chain with state space i; $i = 0$, where i is the number of customers in the system, the transition matrix P of this chain is given as follows:

$$P = \begin{pmatrix} \bar{p_0} & p_0 & 0 & 0 & \cdots \\ \bar{p_1}\bar{\lambda_1} & \bar{p_1}\lambda_1 + p_1\bar{\lambda_1} & p_1\lambda_1 & 0 & \cdots \\ 0 & \bar{p_2}\bar{\lambda_2} & \bar{p_2}\lambda_2 + p_2\bar{\lambda_2} & p_2\lambda_2 & \cdots \\ 0 & 0 & \bar{p_3}\bar{\lambda_3} & \bar{p_3}\lambda_3 + p_3\bar{\lambda_3} & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

Let $\pi_i(t)$ be the probability that there are i customers in the system at time t, and $\pi(t) = [\pi_0(t); \pi_1(t); \ldots]$, and we have:

$$\pi(t+1) = \pi(t)P; \quad \text{or} \quad \pi(t+1) = \pi(0)P^{t+1}.$$

If P is irreducible and positive recurrent, there exists an invariant vector, which is equivalent to also the limiting distribution, $\lim_{t\to\infty} \pi(t) = \pi$; hence:

$$\pi = \pi P.$$

Let's set $\pi = [\pi_0; \pi_1; \pi_2; \ldots]$, from which the balance equations are given by:

$$\begin{aligned} \pi_0 &= \bar{p_0}\pi_0 + \bar{p_1}\bar{\lambda_1}\pi_1 \\ \pi_1 &= p_0\pi_0 + (\bar{p_1}\lambda_1 + p_1\bar{\lambda_1})\pi_1 + \bar{p_2}\bar{\lambda_2}\pi_2 \\ \pi_i &= \bar{p}_{i-1}\lambda_{i-1}\pi_{i-1} + (\bar{p}_i\lambda_i + p_i\bar{\lambda}_i)\pi_i + \bar{p}_{i+1}\bar{\lambda}_{i+1}\pi_{i+1} \end{aligned}$$

From the first equation, we have:

$$\pi_1 = \frac{p_0}{\bar{p_1}\bar{\lambda_1}}\pi_0;$$

$$\pi_i = \left(\prod_{v=0}^{i-1} \frac{p_v\lambda_v}{\bar{p}_{v+1}\bar{\lambda}_{v+1}} \right) \pi_0; \quad i \geq 1.$$

Using the normalization condition $\sum_{i=0}^{\infty} \pi_i = 1$, we obtain:

$$\pi_0 = \left[1 + \sum_{i=1}^{\infty} \prod_{v=0}^{i-1} \frac{p_v \lambda_v}{\bar{p}_{v+1} \bar{\lambda}_{v+1}} \right]^{-1}.$$

For the stability of the system, the quantity must be:

$$\pi_0 = \sum_{i=1}^{\infty} \prod_{v=0}^{i-1} (p_v \lambda_v)[\bar{p}_{v+1} \bar{\lambda}_{v+1}]^{-1} < 1.$$

Note 1.2.9.1. The birth and death process is a special case of quasi-birth and death (QBD) processes, where the transition matrix is made up of block matrices instead of scalars.

1.2.10 Bernoulli Processes

Definition 1.2.10.1. We will thus define an "arrival process" on a slotted axis. The simplest example is provided by the Bernoulli process, also known as the "geometric arrival process." At each time T_k, the process randomly draws according to a probability p the possible arrival of a customer, independently of previous arrivals (memoryless process). We can give the probability distribution of the number of arrivals during a period of t ticks:

$$P\{A \text{ arrivals in } k \text{ slots}\} = \binom{k}{A} (1-p)^{k-A} p^A \tag{1.10}$$

Similarly, we can give the distribution of the interval between two successive arrivals (from which the name "geometric process" is derived, albeit improperly):

$P(A) = p(1-p)^{A-1}$

The average interval has a duration of $1/p$. If we define K as the duration of the slot, we can interpret the formula as giving the probability of a duration $T = AK$ between arrivals. Then, as K tends towards 0, such that $p/K = \lambda$ remains constant, we can easily see that the previous distribution admits a Poisson process as its limit. The same result would be obtained by starting from formula (1.10).

1.2.11 Renewal Process

Renewal theory plays a fundamental role in many applied fields, particularly in equipment reliability, asset management (especially in risk-based models), and queueing theory. It can be used to model phenomena in which events repeat over time, such as machine breakdowns and replacements, or successive customer arrivals.

1.2.12 Discrete-Time Renewal Process

definition 1.2.12.1. A discrete-time renewal process is a stochastic counting process $\{N(t), t \in \mathbb{Z}_{\geq 0}\}$ that enumerates the number of occurrences of a certain event over discrete time. The inter-arrival times between successive events are assumed to be independent and identically distributed (i.i.d.) positive integer-valued random variables.

Let

$$\{X_t\}_{t\geq 1}$$

be a sequence of i.i.d. positive integer-valued random variables representing the inter-arrival times between consecutive events.

The renewal times are defined by :

$$S_n = \sum_{t=1}^{n} X_t, \qquad n = 1, 2, 3, \ldots$$

where S_n denotes the time of the n-th event occurrence.

The corresponding counting process is given by :

$$N(t) = \max\{n \geq 0 : S_n \leq t\}, \qquad t = 0, 1, 2, \ldots$$

Hence, the process $\{N(t), t \geq 0\}$ counts the total number of renewals up to discrete time t and is called a discrete-time renewal process.

Note 1.2.12.1.

This definition generalizes the classical Bernoulli and geometric renewal processes commonly used in discrete-time queueing models, such as the Geo/Geo/1 queue.

1.3 Introduction to Queueing Systems

Queueing theory is a branch of applied mathematics and probability that studies the behavior of systems where customers wait for service. It is applied in various fields such as banks, hospitals, call centers, and production networks, helping to analyze queue length, waiting times, and system time. Mathematical modeling relies on stochastic processes, such as birth-and-death processes, to accurately evaluate system performance. Key concepts include Little's Law and Kendall's notation for describing different types of queues. The theory began in the early 20th century with Agner Krarup Erlang, who studied telephone traffic, and later evolved to cover more complex models. Today, queueing theory is a crucial tool in operations research and complex system management, improving service quality and reducing waiting times.

Definition 1.3.1. (Queue).

This refers to all customers waiting to be served, but not the customer who is being served.

Definition 1.3.2. (Waiting system).

This refers to all customers in the queue, including the customer who is being served.

The waiting phenomenon affects all customers concerned (in the case, as in looped systems, of a subsequent return to the entrance, for example, of machines breaking down in a workshop, the number of customers is often finite in any case). These terms are becoming more widespread and take on their full meaning when there are several service stations and several queues.

Population: The population is the source of potential customers. It is defined by its cardinality (finite or infinite).

Queue: The queue is defined by its maximum capacity of waiting customers (finite or infinite).

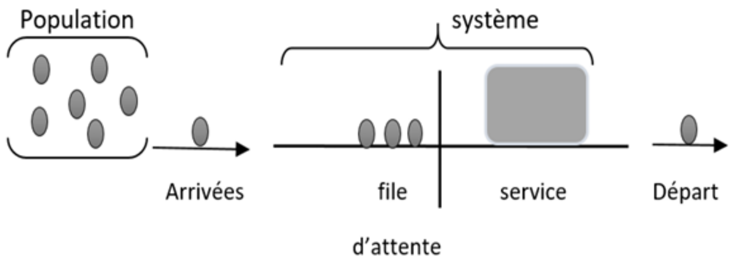

Figure 1.1: General structure of a queueing system

Customers: Customers (from the population) arrive at the system at an average arrival rate.

Service: One or more servers provide the service. The time between the start and end of service for a customer is the service time, which is assumed to be probabilistic. The service rate is another characteristic of the system.

Inter-arrival: This is the time interval between the arrival of successive customers. These inter-arrivals are assumed to be independent and follow the same probability distribution.

Service strategy: The service strategy refers to the order in which customers are served: first come, first served; randomly; according to priority, etc.

1.3.1 Arrival flows

Arrivals may occur regularly (deterministic) or completely randomly, on behalf of individuals or groups, and may come from different populations (in the case of interaction with the environment). The inter-arrival times must therefore be modelled. In certain situations, it will be necessary to take into account the size of the population likely to enter the system (if it is not infinite, the number of entrants will decrease over time). This parameter will therefore need to be taken into account, particularly in the study of emigration-immigration models.

1.3.2 Queue Structure

The service can consist of one or more servers, which can be arranged in various ways:

Parallel Queues

This arrangement applies to queues where the customer can choose the server:

- Queues of people waiting in an administrative office offering several services,
- Queues of shoppers waiting at checkout counters in a hypermarket,
- Queues of cars arriving at a highway toll booth.

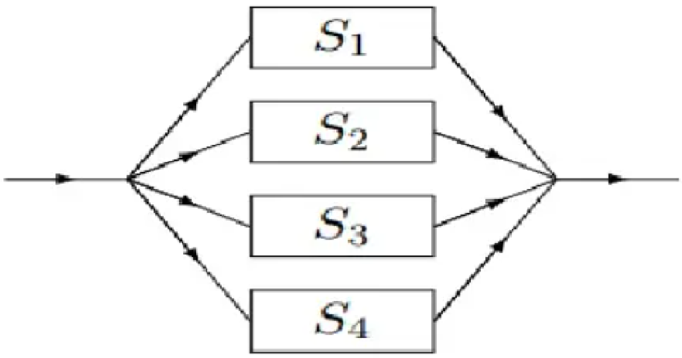

Figure 1.2: parallel servers

Serial Servers

This provision concerns chain services:

- Catering services,
- Vehicle registration services in a prefecture (requiring two steps: registration and then card production),
- Medical examinations in an infirmary (requiring several successive checks),
- Production lines with quality control, etc.

Networked Servers

This arrangement is much more complex and realistic:

- Telecommunication hubs,

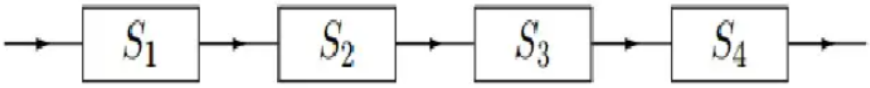

Figure 1.3: serial servers

- Computer networks,
- The internet...

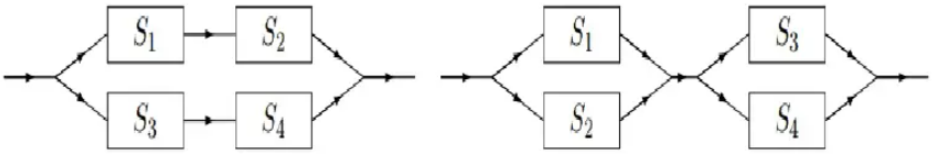

Figure 1.4: network servers

We will also need to model service times.

1.3.3 System Capacity

Many queuing systems have **a waiting room with limited capacity**. When this room is full, people will be turned away at the entrance to the system. This factor is particularly important when studying **tandem queues**.

For example, if **a waiting room with limited capacity is inserted between two consecutive services**, when this room reaches saturation point, a blockage occurs at the first service.

1.3.4 Service disciplines

Service discipline determines the order in which customers are served. Severala single server performs the servicein which people or things arrive, wait, and are served—such as common rules are adopted.

- A rule of courtesy would be to serve people in the order in which they arrive, **FCFS (First Come, First Served)**, which is **FIFO** when the service is performed by a single server.
- A rule that is the opposite of the rule of courtesy is to serve the last customer to arrive first: stock management. This is the **LCFS (Last Come, First Served)** discipline. **LIFO (Last In, First Out)** in cases where the service is performed by a single server.
- In some emergency services, priority rules apply: customers are divided into several classes with different priority orders. A priority person must be served before a non-priority person, even if the latter arrived first.

- Finally, in other situations, **processor sharing** is adopted: this is particularly common in computing, where different tasks are processed simultaneously.

1.4 Kendall Notation

Queuing models help us analyze systems where people or things arrive, wait, and get served—like customers at a bank or data packets in a network. Kendall notation is a shorthand way to describe these models completely. The full version uses six parts:

$$\mathbf{A/B/s/N/K/D}.$$

A: Arrival process
This describes how arrivals happen. Common notations include:

- **Geo (Markovian/Geometric)**: Arrivals follow an geometric distribution (memoryless, Bernoulli process)
- **G (General)**: Any general distribution (no specific assumptions)
- **D (Deterministic)**: Fixed, constant inter-arrival times

B: Service process
This describes the service time distribution. Uses the same symbols as A.

- **Geo**: Geometric service times
- **E**: Erlang service times
- **G**: General service times
- **D**: Constant service times

s: Number of Servers
Number of parallel servers:

$$s = 1, 2, 3, \ldots$$

Examples: $s = 1$ (single cashier) and $s = 2$ (two tellers).

N: System Capacity
Maximum number in system (waiting + being served):

$$N = \text{finite number or } \infty$$

If N is finite, the system can be full and reject arrivals.

K: Population Size
Total potential arrivals in the source:

$$K = \text{finite number or } \infty$$

Finite K = limited pool (e.g., 10 machines in a workshop).

D: Queue Discipline Service order rule:

- **FCFS**: First Come, First Served
- **LCFS**: Last Come, First Served
- **SIRO**: Service In Random Order
- **Priority**: Based on urgency/class

Abbreviated Form: A/B/s
Often, we use a shorter version: **A/B/s**. This assumes the system has infinite capacity ($N = \infty$) and infinite population ($K = \infty$). In other words, there's no limit on how many can wait or how many could arrive—it's an open system with ample capacity and an endless supply of arrivals. This is common for basic models where we don't worry about overcrowding or limited sources.

1.5 Little's Law

Little's law is an empirical relationship formulated by John Little in 1961. It only concerns the steady state of the system. Little's law is crucial for the analysis.

Theorem 1.5.1. (Little's formula) For a steady-state system, the arrival rate p of customers, the mean number N of customers in the system, and the mean dwell time T (mean response time) are related by the equation.

$$\bar{N} = pT.$$

1.6 The Performance Measurement Of a Queueing System

The study of queueing systems focuses on figuring out or predicting how well a system performs under specific conditions. Usually, this is done just for the steady state—when things have settled into a long-term pattern—and the key metrics people look at most often include:

- $\bar{N}$: The mean number of customers in the system.
- Q: The mean number of customers waiting in the queue.
- T: The mean time a customer spends in the system.
- W: The mean time a customer spends just waiting in the queue.

- S: The mean number of customers being served.
- V: The mean number of customers in the orbit.
- Z: The mean wait time for a customer in the orbit.
- C: The number of busy servers.
- U: The number of customers in the orbit.

Chapter 2

Classical Markovian Discrete-Time Queueing System

Introduction

Discrete-time Markov models form one of the cornerstones of queueing theory due to their ease of analysis and their ability to describe a very wide range of real-world systems. Unlike continuous-time models, discrete-time systems evolve at well-defined moments, making them particularly well-suited for modeling many modern systems such as digital communication networks, computer systems, information processing applications, and so on.

Markov models possess the Markov property, which states that the future state of the system depends only on its present state and not on its past, which explains the significant gains in simplicity from the perspective of mathematical analysis and the calculation of probabilistic characteristics.

This chapter addresses classical discrete-time queueing models, specifically the Geo/Geo/1 and Geo/Geo/s models. Their description, basic assumptions, and analytical methods for testing their performance will be presented.

The objective is to provide theoretical tools useful for understanding the more complex model that will be addressed in the next chapter.

2.1 Discrete-Time Markovian Queues

Arrivals and departures in a step-by-step queue are considered as discrete stochastic processes. This corresponds to moments in time when a customer arrives or leaves with a given probability distribution. This presents another important alternative to continuous-time systems, which are most often used in modeling.

Let us now consider the queues already studied as having an exponential service time distribution ($M/M/1; \dots$), which corresponds to a duration that can be used to model the duration of a telephone conversation in a conventional telecommunications system. Basically, as a continuous-time model, this is only an approximation because real events occur in discrete time, so even if we choose at a given moment to use seconds, milliseconds, etc., we are still dealing with a discrete variable, with the duration being quantified using a slot.

In this configuration, arrivals join the system at the beginning of a slot and departures are recorded no later than the second preceding the next time.

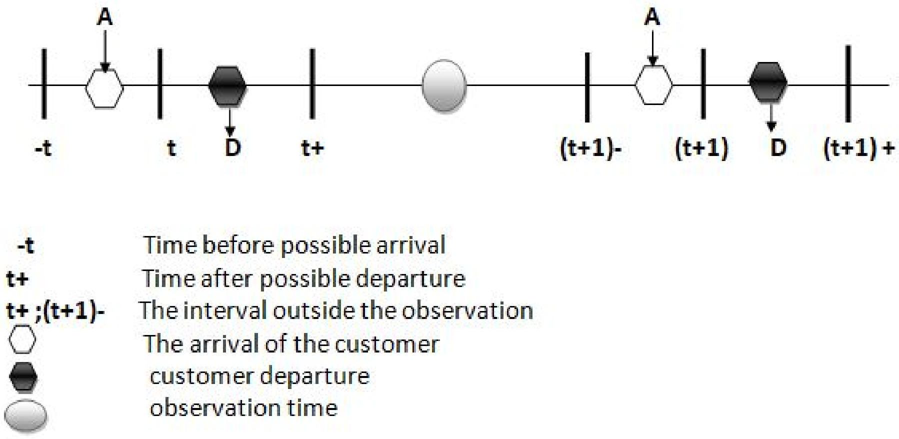

Figure 2.1: Time slot in the system

The customer who finishes their service at the end of slot J leaves behind those who arrived in the second slot and those who are waiting at the beginning of a slot. The service for the customer who arrives in the slot always begins at the start of the slot.

2.2 Some Models Of Discrete-Time Queues

2.2.1 The Geo/Geo/1 Model

Introduction

Queueing systems are used to model many real service systems, such as communication networks and computer systems. In some situations, special events may occur that remove customers from the system, such as negative arrivals or disasters, which help reduce congestion. In this work, we study a Geo/Geo/1 queue, a discrete-time single-server model where arrivals and service times follow geometric distributions. The aim is to analyze the behavior and performance of this system under these conditions.

Consider a classic Geo/Geo/1 discrete-time single-server queueing system, where the time axis is divided into sub-intervals of constant length, called "slots," and where all activities take place at the ends of the "slots."

2.2.2 Model description

The Geo/Geo/1 is the most basic discrete-time queueing model, which can easily be considered as the discrete-time M/M/1 queueing model counterpart to the continuous version.

This model can accommodate an infinite number of customers and operates under a FIFO service system. The primary difference in this model is that time is divided into time intervals (discrete time units), and the interarrival times and service times are all treated as geometric distributions.

2.2.3 Associated Markov Chain

The Geo/Geo/1 queue model can be defined as a special case of a birth-and-death process.

We consider the case where $p_i = p$, $\forall i \geq 0$ and $\lambda_i = \lambda$, $\forall i \geq 0$.

This allows us to explain the generating functions of the inter-arrival and service time laws $p(k) = 1 - p + kp$ and $\lambda(k) = 1 - \lambda + k\lambda$ respectively. We now define the Markov chain $(X_t)_{t\in\mathbb{N}}$ in state space (i, t); $i = 0$ or 1, $t \geq 0$, where i represents the number of customers in the system.

The transition matrix P of this chain is given by

$$P = \begin{pmatrix} \bar{p} & p & 0 & \cdots & \\ \bar{p}\lambda & \bar{p}\bar{\lambda} + p\lambda & p\bar{\lambda} & \cdots & \\ 0 & \bar{p}\lambda & \bar{p}\bar{\lambda} + p\lambda & p\bar{\lambda} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

There are several approaches to analyzing this Markov matrix. Among these, we can mention the standard numerical approach, the approach based on generating functions, and the approach based on geometric matrices.

Recurrence Approach

The chain $(X_t)_{t\in\mathbb{N}}$ is irreducible and positive recurrent and therefore admits a single fixed-point distribution. Let $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$ satisfy $\boldsymbol{\pi} = \boldsymbol{\pi} P$.

From this, we obtain the equations of state:

$$\pi_0 = \pi_0\bar{p} + \pi_1\bar{p}\lambda, \tag{2.1}$$

$$\pi_1 = \pi_0 p + (p\lambda + \bar{p}\bar{\lambda})\pi_1 + \pi_2 \bar{p}\lambda, \tag{2.2}$$

$$\pi_i = \pi_{i-1} p\bar{\lambda} + \pi_i(\bar{p}\bar{\lambda} + p\lambda) + \pi_{i+1}\bar{p}\lambda, \quad i \geq 2. \tag{2.3}$$

If we suppose that $\rho = \frac{p\bar{\lambda}}{\bar{p}\lambda}$, then we have

$$\pi_i = \left(\frac{\rho^i}{\bar{\lambda}}\right)\pi_0, \quad i \geq 1$$

Using the normalization condition, we obtain

$$\sum_{i=0}^{\infty} \pi_i = 1,$$

We will have :

$$\pi_0 = \frac{\lambda - p}{\lambda}, \quad With \quad \rho < 1,$$

$\rho < 1$ **if and only if** $p < \lambda$ **is the stability condition.**

The Generating Function Approach

Let $f(k) = \sum_{i=0}^{\infty} \pi_i k^i, |k| \leq 1$, be the generating function of the stationary distribution π. Multiplying equations (2.1)-(2.3) by k^0, k^1, k^i respectively, we obtain:

$$\pi_0 = \bar{p}\pi_0 + \bar{p}\lambda\pi_1,$$

$$k\pi_1 = pk\pi_0 + (\bar{p}\bar{\lambda} + p\lambda)k\pi_1 + \bar{p}\lambda k\pi_2,$$

$$k^i\pi_i = p\bar{\lambda}k^i\pi_{i-1} + (\bar{p}\bar{\lambda} + p\lambda)k^i\pi_i + \bar{p}\lambda k^i\pi_{i+1}, \quad i \geq 2.$$

Summing over i, we obtain the last equation:

$$f(k) = f(k)\left[\bar{p}\bar{\lambda} + p\lambda + \frac{\bar{p}\lambda}{k} + p\bar{\lambda}k\right] + \left[\bar{p} + pk - \bar{p}\bar{\lambda} - p\lambda - \frac{\bar{p}\lambda}{k} - p\bar{\lambda}k\right]\pi_0.$$

After simplification, we get

$$f(k) = \frac{\lambda(pk + \bar{p})(k-1)}{(\bar{p}\lambda - p\bar{\lambda}k)(k-1)}\pi_0$$

$$= \frac{\lambda(pk+\bar{p})}{(\bar{p}\lambda - p\bar{\lambda}k)}\pi_0.$$

Using $f(k)|_{k=1} = 1$, we obtain:

$$\pi_0 = \frac{\lambda - p}{\lambda}.$$

Since $\pi_0 > 0$, we have the stability condition $p < \lambda$.

The Geometric Matrices Approach.

The transition probabilities of the operation $(X_t)_{t\in\mathbb{N}}$ can be represented in the form of the following mass matrix called the semi-birth matrix and semi-death transition probability matrix (QNM), which is related to the operation level.

The transition matrix is obtained by:

The matrix P in block form is as follows:

$$P = \begin{pmatrix} N_{00} & N_{01} & 0 & 0 & \cdots & & \\ N_{10} & N_{11} & M_0 & 0 & \cdots & & \\ 0 & N_{21} & M_1 & M_0 & 0 & \cdots & \\ 0 & 0 & M_2 & M_1 & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

In matrix form, the system of equations (2.1)-(2.3) above is given by:

$$\pi_0 = \sum_{i=0}^{\infty} \pi_i N_{i0}$$

$$\pi_1 = \sum_{i=0}^{\infty} \pi_i N_{i1}$$

$$\pi_j = \sum_{v=0}^{\infty} \pi_{j+v-1} M_v, \quad j \geq 1$$

In the general case where N_{ij} and M_v are matrices, the transition matrix is called a geometric matrix.

In our case, we have:

$$M_0 = p\bar{\lambda}, \quad M_1 = (\bar{p}\bar{\lambda} + p\lambda), \quad M_2 = \bar{p}\lambda, \quad M_v = 0 \quad \text{for all} \quad v \geq 3$$

$$N_{00} = \bar{p}, \quad N_{10} = \bar{p}\lambda, \quad N_{01} = p, \quad N_{11} = \bar{p}\bar{\lambda} + p\lambda, \quad N_{21} = \bar{p}\lambda,$$
$$N_{ij} = 0 \quad \text{for} \quad i \geq 3, \quad j \geq 2.$$

Since the components of the matrix P are scalars, we are thus led back to the first two approaches.

2.2.4 Performance Measurement

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

$$\bar{N} = \sum_{i=0}^{\infty} i\pi_i$$

and is obtained as follows:

$$\begin{aligned}\bar{N} &= (\bar{\lambda}\lambda)^{-1}(\lambda - p)[\rho + 2\rho^2 + 3\rho^3 + \cdots] \\ &= (\bar{\lambda}\lambda)^{-1}(\lambda - p)\rho[1 + 2\rho + 3\rho^2 + \cdots] \\ &= (\bar{\lambda}\lambda)^{-1}(\lambda - p)\rho\frac{d(1-\rho)^{-1}}{d\rho}.\end{aligned}$$

Hence,

$$\bar{N} = \frac{p\bar{p}}{\lambda - p}.$$

Remark 2.1.

This same result can be determined, in steady state, using the relation:

$$\bar{N} = \left.\frac{df(k)}{dk}\right|_{k\to 1}$$

where $f(k)$ is the generating function.

- **The mean number of customers waiting in the queue** is:

$$Q = \sum_{i=1}^{\infty}(i-1)\pi_i = \bar{N} - \rho. \tag{2.4}$$

Hence:

$$Q = \bar{N} - \frac{p}{\lambda}.$$

The system capacity is unlimited, so $p_e = p$ where p_e is the actual arrival rate into the system.

Using Little's formulas, we obtain:

- **The mean time a customer spends in the system:**

$$T = \frac{\bar{N}}{p} = \frac{\bar{p}}{\lambda - p}.$$

- **The mean time a customer spends just waiting in the queue:**

$$\begin{aligned}W &= \frac{Q}{p} \\ &= \frac{p\bar{\lambda}}{\lambda(\lambda - p)}.\end{aligned}$$

2.3 The Geo/Geo/s

Introduction

Some examples of service systems are telecommunications and computer systems; therefore, queueing systems are used to create models for these types of service systems. When a customer arrives and finds that all servers are currently busy, the customer has three options: Leave, Wait in the Queue, or Wait Until Later to Try Again. Retrial queues are particularly important in communication systems where repeated service attempts occur frequently. In recent years, discrete-time queueing models have received increasing attention due to the development of digital systems. In this work, we study a Geo/Geo/c retrial queue, which is the discrete-time analogue of the M/M/s retrial queue, and analyze its main performance characteristics.

2.3.1 Model description

The Geo/Geo/s queue is a discrete-time queuing system where both arrivals and service completions follow geometric distributions. It is a generalization of the Geo/Geo/1 queue to s parallel servers, allowing the analysis of multi-server systems with discrete-time events.

Key Features

- **Time type:** Discrete.
- **Arrival distribution:** Geometric, with parameter p (probability of arrival in a time slot).
- **Service distribution:** Geometric, with parameter λ (probability of service completion per server per slot).
- **Number of servers:** s.

Model Structure

At each discrete time slot:

- A new customer arrives with probability p.
- Up to s customers are served simultaneously. Each server completes its service with probability λ independently.
- Excess customers wait in the queue according to the service discipline (usually FIFO).

This discrete-time framework is the temporal analogue of the continuous-time M/M/s queue.

2.3.2 Steady-State Distribution

The steady-state distribution of the Geo/Geo/s queue can be computed recursively using the one-step transition probabilities of the discrete-time Markov chain $\{X_t\}$. The transitions depend on the arrival probability p, the service probability λ, and the number of servers s. The steady-state probability π_i represents the long-run probability of having i customers in the system.

Service Completion Probability

The service probability depends on the number of customers in the system. It is given by

$$\lambda_i = \begin{cases} 1-(1-\lambda)^i, & i < s, \\ 1-(1-\lambda)^s, & i \geq s. \end{cases}$$

Balance Equations

For a birth-death process, the steady-state probabilities satisfy.

$$\pi_{s-1} = \pi_{s-1}\bar{p} + \pi_s \bar{p}\lambda, \tag{2.5}$$

$$\pi_s = \pi_{s-1}p + (p\lambda + \bar{p}\bar{\lambda})\pi_s + \pi_{s+1}\bar{p}\lambda, \tag{2.6}$$

$$\pi_i = \pi_{i-1}p\bar{\lambda}_{i-1} + \pi_i(\bar{p}\bar{\lambda}_i + p\lambda_i) + \pi_{i+1}\bar{p}\lambda_{i+1}, \quad i \geq s. \tag{2.7}$$

By recursion, we obtain

$$\pi_i = \prod_{j=0}^{i-1} \frac{p_j(1-\lambda_j)}{\bar{p}_{j+1}(1-\lambda_{j+1})}\pi_0.$$

Steady-State Probabilities

Thus, the steady-state probabilities are given by

$$\pi_i = \begin{cases} \pi_0 \dfrac{\prod_{j=0}^{i-1} p_j \prod_{j=0}^{i-1} \bar{\lambda}_j}{\prod_{j=0}^{i-1} \bar{p}_{j+1} \prod_{j=0}^{i-1} \lambda_{j+1}}, & i < s, \\ \\ \pi_0 \dfrac{\prod_{j=0}^{s} p_j \prod_{j=0}^{s} \bar{\lambda}_j}{\prod_{j=0}^{s} \bar{p}_{j+1} \prod_{j=0}^{s} \lambda_{j+1}}, & i \geq s. \end{cases}$$

$$\pi_i = \begin{cases} \pi_0 \dfrac{(p)^{i-1} \bar{\lambda}^{\sum_{j=0}^{i-1} j}}{(\bar{p})^i \prod_{j=0}^{i-1} \lambda_{j+1}}, & i < s, \\ \\ \pi_0 \dfrac{(p)^s \bar{\lambda}^{\sum_{j=0}^{s} j}}{(\bar{p})^{s+1} \prod_{j=0}^{s} \lambda_{j+1}}, & i \geq s. \end{cases}$$

$$\pi_i = \begin{cases} \pi_0 \dfrac{(p)^{i-1} \bar{\lambda}^{\frac{i(i-1)}{2}}}{(\bar{p})^i \prod_{j=0}^{i-1} (1 - \bar{\lambda}^j)}, & i < s, \\ \\ \pi_0 \dfrac{(p)^s \bar{\lambda}^{\frac{s(s+1)}{2}}}{\bar{p}^{s+1} (1 - \bar{\lambda}^s)^s \prod_{j=0}^{i-1} (1 - \bar{\lambda}^j)}, & i \geq s. \end{cases}$$

Normalization Condition

The constant π_0 is obtained from the normalization condition.

$$\sum_{i=0}^{\infty} \pi_i = 1.$$

Therefore

$$\pi_0 = \left[1 + \sum_{i=1}^{s-1} \frac{p^{i-1} \bar{\lambda}^{\frac{i(i-1)}{2}}}{\bar{p}^i \prod_{j=0}^{i-1} (1 - \bar{\lambda}^j)} + \frac{p^s \bar{\lambda}^{\frac{s(s+1)}{2}}}{\bar{p}^{s+1} (1 - \bar{\lambda}^s)^s} \right]^{-1}$$

Stability Condition

The queueing system is stable if the traffic intensity satisfies.

$$\rho = \frac{p\bar{\lambda}}{\bar{p}\lambda^s} < 1.$$

Under this condition, all performance measures of the system can be derived.

2.3.3 Performance Measurement

In the **Geo/Geo/s** queuing model, performance measures are used to evaluate the efficiency of the system. The mean number of customers in the system is $\bar{N}$, which represents the sum of customers being served and waiting. The mean number of customers being served is $S = 1 - \pi_0$, where π_0 is the probability that the system is empty. The mean number of customers in the queue is $Q = \bar{N} - S$.

From Little's law, we obtain the mean waiting time in the queue $W = \dfrac{Q}{p}$, and the mean time a customer spends in the system is $T = W + \dfrac{1}{\lambda}$.
These measures allow us to analyze the performance of the system when the stability condition is satisfied.

- **The mean number of customers in the system**

  The mean number of customers in the system is:

$$\bar{N} = \sum_{i=1}^{\infty} i\pi_i$$

  which can be written as:

$$\bar{N} = \sum_{i=1}^{s-1} i \frac{p^{i-1}\bar{\lambda}^{\frac{i(i-1)}{2}}}{\bar{p}^i \prod_{j=0}^{i-1}(1-\bar{\lambda}^j)} + \sum_{i=s}^{+\infty} \frac{p^s \bar{\lambda}^{\frac{s(s+1)}{2}}}{\bar{p}^{s+1}(1-\bar{\lambda}^s)^s \prod_{j=0}^{i-1}(1-\bar{\lambda}^j)}$$

$$= \sum_{i=1}^{s-1} i \frac{p^{i-1}\bar{\lambda}^{\frac{i(i-1)}{2}}}{\bar{p}^i \prod_{j=0}^{i-1}(1-\bar{\lambda}^j)} + \frac{sp^s \bar{\lambda}^{\frac{s(s+1)}{2}}}{\bar{p}^{s+1}(1-\bar{\lambda}^s)^s}$$

  This is the mean of the total number of customers in the queue + in service.

- **The mean number of customers in service**

  The mean number of customers being served is:

$$S = 1 - \pi_0$$

  where:

$$\pi_0$$

  is the probability that the system is empty.

- **The mean number of customers in the queue**

  The mean number of customers waiting in the queue is:

$$Q = \bar{N} - S$$

  This corresponds to customers who are waiting without being served.

- **The mean time a customer spends just waiting in the queue**

  The mean waiting time for a customer in the queue is:

$$W = \frac{Q}{p}$$

  This is an application of Little's law.

- **The mean time a customer spends in the system**

  The mean time spent in the system (waiting + service) is:

$$T = W + \frac{1}{\lambda}$$

  where:

$$\frac{1}{\lambda}$$

  is the mean service time.

Chapter 3

Advanced Discrete-Time Queueing Models

Introduction

Classical discrete-time queuing models have been studied in previous chapters, notably the Geo/Geo/1 model, based on simplifying assumptions regarding arrivals, service times, and customer behavior. However, these simplifying assumptions may mask the much more complex reality of modern systems, where phenomena such as customer impatience (reneging), refusal to enter the system (balking), retries, feedback after service, etc., can pose problems. Furthermore, the server's operation itself can be disrupted by breakdowns, vacations, or even disasters.

The objective of this chapter is therefore to explore advanced discrete-time queueing models that account for these various characteristics in order to more accurately model real-world systems and analyze their performance in greater detail. We will present several enhanced models and the analytical methods we used to evaluate key indicators such as average queue length, waiting time, and system stability.

3.1 Queueing Systems with Callbacks (retrial)

Queueing systems with retries, also known as retrial queues, constitute a specific class of stochastic models characterized by a particular congestion management mechanism. When a client enters the system and finds that all servers are busy, it does not necessarily join a traditional queue; it may either leave the system permanently or defer its request by retrying after a random delay. In the latter case, the customer is assumed to enter a virtual entity called an "orbit," from which they subsequently generate attempts to access the service.

These systems occupy an intermediate position between two fundamental models in queueing theory. On the one hand, they generalize the model with rejection, corresponding to low recall intensity, where clients are systematically lost in the event of saturation. On the other hand, they approach the classic FIFO queueing model when the recall rate becomes high, reflecting behavior similar to that of clients willing to wait. Thus, the recall mechanism allows the system's behavior to be modulated between these two extreme regimes.

This type of modeling is particularly useful for analyzing real-world systems characterized by repeated attempts to access the service, thereby enabling a more accurate representation of congestion phenomena and improving our understanding of the system's overall performance.

3.1.1 Description of retrial with a queueing system

A queueing system with retrials is a stochastic service model in which arriving customers who find all servers busy and no available waiting space are not immediately lost from the system. Instead, they are directed to an external orbit, where they remain temporarily inactive. After a random period of time, these customers attempt to access the service again. Each retrial is considered a new attempt for service, and the customer is either served if a server becomes available or returns again to the orbit if the system remains congested. This mechanism is widely used to model systems with repeated access attempts, such as telecommunication networks and call centers.

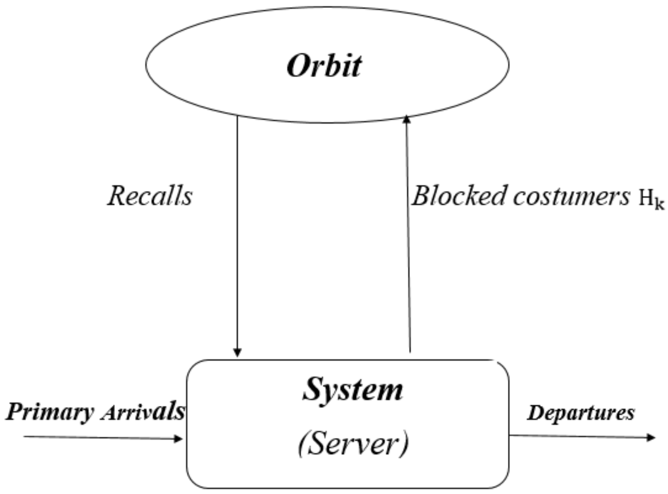

Figure 3.1: The general diagram of a retrial queueing system

3.1.2 Description of a Queueing System with Balking

A queueing system with balking is a service model in which arriving customers make a decision not to enter the system upon observing congestion conditions. Specifically, if all servers are occupied or the waiting line is excessively long, a customer may choose to abandon the system immediately without joining the queue. This decision is typically based on perceived waiting time or system load. Unlike retrial systems, balking customers do not reattempt service at a later time, and their departure is permanent at the moment of arrival.

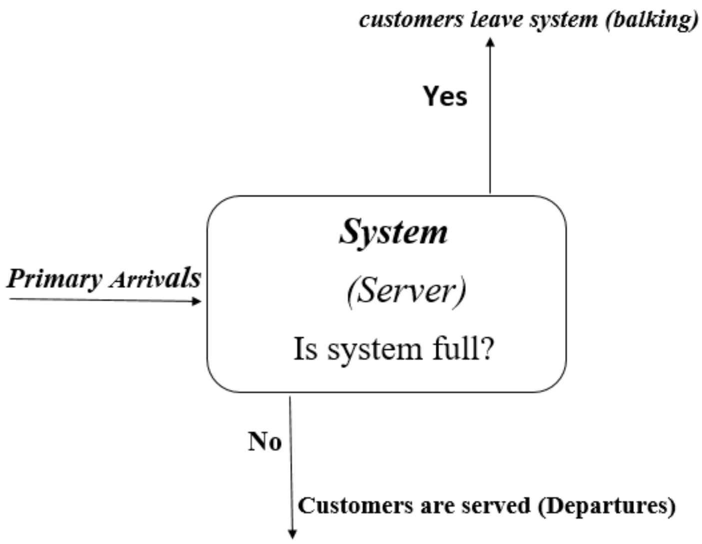

Figure 3.2: The general diagram of a queueing system with Balking

3.1.3 Description of a Queueing System with Retrials, Callbacks, and a Waiting Room

A call-and-return queueing system consists of $s \geq 1$ identical and independent servers, a waiting area with capacity $N - s$ (where $N \geq s$), as well as a service queue of capacity O, which may be finite or infinite.

Upon the arrival of a primary customer, if at least one server is free, the customer is immediately served and leaves the system after the completion of service. Otherwise, if there are available waiting places, the customer joins the queue. When all servers and all waiting slots are occupied, the client is either permanently lost with probability $1 - H$, or admitted into the orbit with probability H, thereby becoming a source of repeated calls.

Customers in the orbit, called secondary customers, attempt to access the service after random time intervals following a given probability distribution. The callback process can be characterized by different intensities, such as constant, classic, or linear callbacks. If a secondary customer retries, it is treated as a primary customer: if it finds an available server, the customer is served immediately; otherwise, if there is space in the queue, the client joins the queue; otherwise, if the system is saturated, the customers permanently leaves the system with probability $1 - H_k$ (where k is the number of failed attempts) or returns to the queue with probability H_k if it is not full; otherwise, the customer permanently leaves the system without service.

The general diagram of a call-back queueing system is shown in Figure 3.3:

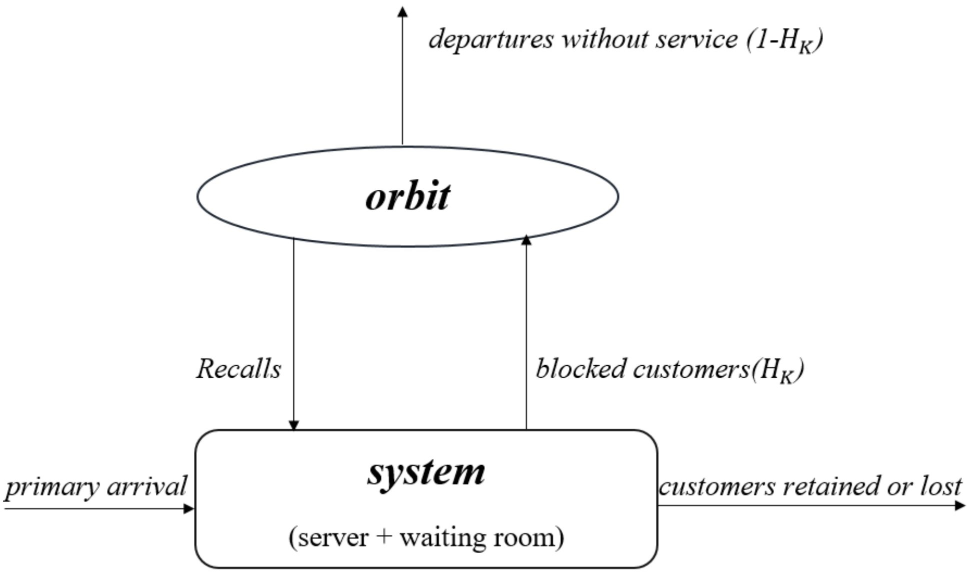

Figure 3.3: The general diagram of a callback queueing system

3.1.4 Terminology and Notation

Referring to the Kendall notation established for the classic queueing model, the notation for the callback queueing model is written by adding two additional symbols to the six symbols previously defined provisionally.

$$A/B/s/N/K/D/O/H$$

Where:

1. O: Orbit capacity.
2. H: The persistence function that defines the customer's behavior when faced with a blocking situation (servers are busy).

Note 3.1.4.1. In the case of infinite orbital capacity, the value of O becomes negligible. Conversely, in a lossless system where $H = 1$, the parameter H becomes redundant, and it is no longer necessary to retain it.

Note 3.1.4.2.

- The time interval between two consecutive calls from the same secondary client is referred to as the call interval.
- The inter-call interval distribution is assumed to be geometric with parameter p.
- As $p \to \infty$, the call-back queueing system is identical to the classical queueing system.
- As $p \to 0$, the call-back queueing system becomes an Erlang system with loss.

3.2 Some Examples of Queueing Systems with Callbacks

Today, there is a wealth of literature on queueing systems with callbacks, illustrated by numerous concrete examples. These examples demonstrate how theoretical concepts translate into real-world situations, particularly in communication networks and computer systems. In this chapter, we present some of these practical examples to better understand how systems with callbacks work and their usefulness in modeling real-world phenomena such as congestion, waiting, and repeated attempts to access a service.

3.2.1 Packet Communication Networks

Consider a computer communication network consisting of a set of interfaces called "Interface Message Processors" (IMPs) (these are the packet-switching nodes used to connect computers to the ARPANET), interconnected by cables. A host computer is connected to one of these interfaces.

If the computer sends a message to another computer, it must first send the message with the destination address to the interface to which it is connected. The interface, in turn, sends the message to the destination computer directly if it is connected to it, or indirectly via other interfaces.

Consider an interface to which a main computer is connected. Messages arrive from outside according to a random process. After receiving the message, the computer immediately sends it to its interface. If there is a free buffer, the message is accepted. Otherwise, the message is rejected, and the computer must retry after a period of time.

If empty buffers are available, the rejected message will be stored in a buffer on the main computer. Otherwise, the message will be considered lost.

This problem can be modeled as a single-server call-back queueing system (IMP interface) with buffers (queue positions). The number of buffers on the main computer constitutes the capacity of the orbit.

3.2.2 Local Area Networks: CSMA

In local area networks based on a single shared bus, one of the communication protocols used is non-persistent CSMA (Carrier Sense Multiple Access). This is a method for accessing a local area network. Suppose this network consists of n stations (or processors) connected by this bus. The stations communicate using this bus.

Messages of varying lengths arrive from the station in the outside world. Upon receiving a message, the station divides it into a finite number of fixed-length packets and immediately checks the bus availability. If the bus is free, one of these packets is transmitted via the bus to the destination station, while the other packets are stored in the buffer for future transmission. However, if the bus is busy, all packets must be stored in the buffer, and the station can check the bus after a certain random interval.

This problem can be modeled as a single-server system, where the bus is the server, and the stations' memory represents the queue.

This problem can be modeled as a single-server system, where the bus is the server, and the stations' memory represents the queue.

This system is illustrated in Figure 3.2

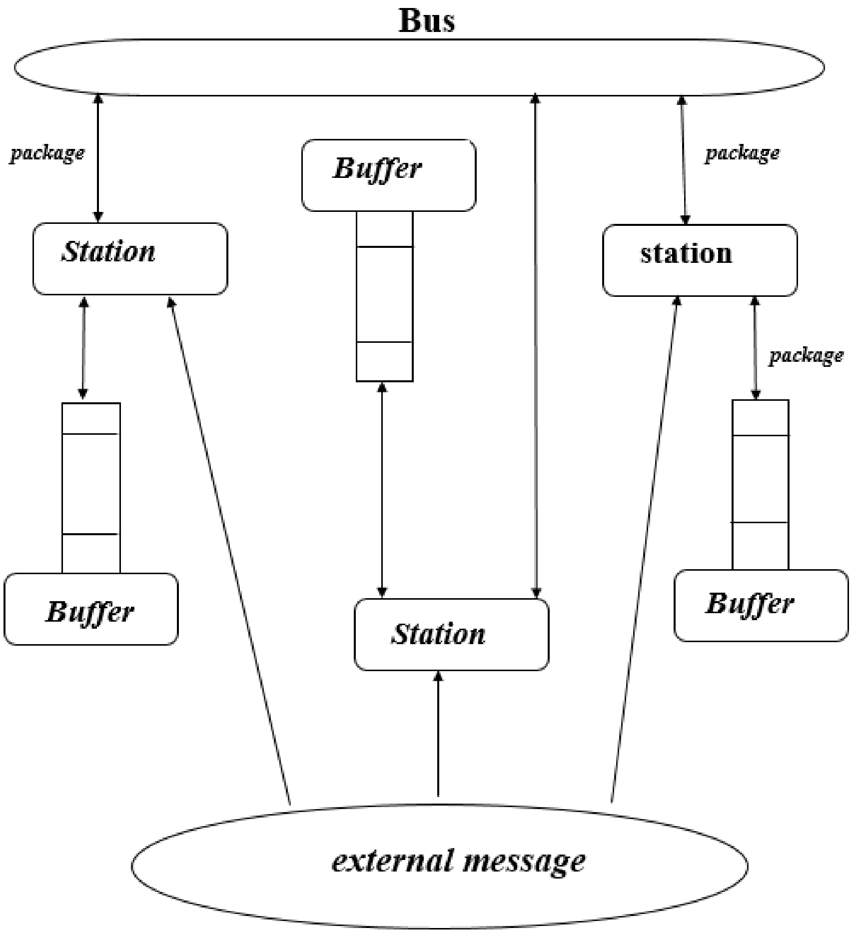

Figure 3.4: Diagram of a local area network

3.3 Geo/Geo/s Queueing System with Callbacks

The present section aims to investigate the Geo/Geo/s queue with geometric retrials; that is, the discrete-time analogue of the M/M/s retrial queue. We propose several algorithmic procedures for the efficient computation of the main performance measures. More specifically, we focus on the stationary distribution of the system state, the busy period, and the waiting time.

Although the efficient computational analysis of non-homogeneous GI/M/1-type processes remains an open problem, we shall show that it is possible to exploit the matrix structure and sparsity of the underlying blocks to proceed algorithmically. There are several methods for an efficient solution of the stationary probabilities and the rate matrix. For the sake of brevity, we present in detail only those methods that we use for obtaining the numerical results.

3.3.1 System Description

We assume that the time axis has been divided into equal intervals of unit length called slots. In the Geo/Geo/s retrial queue, primary customers arrive at the system according to a Bernoulli process with probability p. Equivalently, the arrival process is a discrete renewal process with geometric interarrival times with probability mass function (p.m.f.) $p\bar{p}^{t-1}$, for $t \geq 1$; where $\bar{p} = 1 - p$.

Service is rendered by s identical servers with service times geometrically distributed with p.m.f. $\lambda\bar{\lambda}^{t-1}$; for $t \geq 1$; where $\bar{\lambda} = 1 - \lambda$. This means that for any busy server, the probability that it will finish the ongoing service in the next slot is λ.

An arriving customer who finds a free server upon his arrival begins to be served immediately. Otherwise, he joins the retrial orbit and reapplies for service later. Customers in the retrial orbit behave independently of each other and retry with probability s at every time slot, i.e., their retrial times are independent geometric random variables with p.m.f. $r\bar{r}^{t-1}$; for $t \geq 1$, where $\bar{r} = 1 - r$.

We suppose that primary arrivals occur first, so they have priority over retrials. Among retrying customers, the ones who occupy the servers are selected randomly, i.e., we assume the random order (RO) discipline.

The state of the system at time t can be described by a process $\{(C_t, U_t), t \geq 0\}$, where C_t denotes the number of busy servers and U_t records the number of customers in the retrial orbit. The process $\{(C_t, U_t), t \geq 0\}$ is an irreducible Markov chain with state space

$$\mathcal{S} = \{(i, j) \mid 0 \leq i \leq s,\ j \geq 0\}$$

and one-step transition probabilities $P_{(i,j)(m,n)}$ given as in the following theorem.

Theorem 3.3.1.1. *[8]* The non-zero transition probabilities $P(i,j)(m,n)$ of the process $\{X_t; t \geq 0\}$ are as follows:

(i) If $n = j + 1$ and $i = m = s$, then

$$P(s, j)(s, j + 1) = p\bar{\lambda}^s.$$

(ii) If $\max(0, j-s) \leq n \leq j$, $j-n \leq m \leq s$ and $\max(0, m+n-j-1) \leq i \leq s$, then

$$\begin{aligned} P(i,j)(m,n) = &\Big[(1-\delta_{m,i+j-n+1})\bar{p}\binom{i}{i+j-m-n}\lambda^{i+j-m-n}\bar{\lambda}^{m+n-j} \\ &+ (1-\delta_{j-n,m})p\binom{i}{i+j+1-m-n}\lambda^{i+j+1-m-n}\bar{\lambda}^{m+n-j-1}\Big] \\ &\times\left(\binom{j}{j-n}r^{j-n}\bar{r}^{n} + \delta_{ms}\sum_{h=j-n+1}^{j}\binom{j}{h}r^{h}\bar{r}^{j-h}\right). \end{aligned} \tag{3.1}$$

where the symbol δ_{ab} denotes Kronecker's function defined by

$$\delta_{ab} = \begin{cases} 1, & \text{if } a = b; \\ 0, & \text{if } a \neq b. \end{cases}$$

3.3.2 The stationary distribution of the system state

In this subsection, we study the stationary behavior of the system state. As usual, we begin by investigating the system stability. We expect to have the same situation for the *Geo/Geo/s* retrial queue because, as the number of retrying customers grows, the system approaches the standard *Geo/Geo/s* queue.

Theorem 3.3.2.1. The Markov chain $\{(C_t, U_t); t \geqslant 0\}$ is positive recurrent if and only if

$$p < s\lambda.$$

3.3.3 Busy period analysis

We present some algorithms for the recursive computation of the busy period characteristics. Let $L_{(i,j)}$ be a random variable representing the first passage time to $(0,0)$, starting from a state (i,j). Then, the busy period is defined as $L = L_{(1,0)}$. We are interested in obtaining the probabilities $P\{L_{(i,j)} = t\}$, for $t \geqslant 1$, and the moments of L. Conditioning on the first transition of the process, we have that

$$P\{L_{(i,j)} = 1\} = P_{(i,j)(0,0)}, \quad 0 \leqslant i \leqslant s,\ j \geqslant 0,\ (i,j) \neq (0,0), \tag{3.2}$$

$$P\{L_{(i,j)} = t\} = \sum_{(m,n)\in S_{(i,j)}-\{(0,0)\}} P_{(i,j)(m,n)}P\{L_{(m,n)} = t-1\}, \quad 0 \leqslant i \leqslant s,\ j \geqslant 0,\ (i,j) \neq (0,0),\ t \geqslant 2. \tag{3.3}$$

For every $t \geqslant 1$, we define the infinite column vector $\mathbf{l}^t = (\mathbf{l}^t(0), \mathbf{l}^t(1), \ldots)'$ where

$$\mathbf{l}^t(0) = (P\{L_{(1,0)} = t\}, \ldots, P\{L_{(s,0)} = t\})',$$

$$\mathbf{l}^t(j) = (P\{L_{(0,j)} = t\}, \ldots, P\{L_{(s,j)} = t\})', \quad j \geqslant 1.$$

We denote by $\mathbf{P}_L$ the matrix that results from removing the first row and the first column of the transition matrix $\mathbf{P}$. Let it also $\mathbf{P}_0$ be the column vector of dimension $s + (s+1)t$ created by removing the first element of the first column of the matrix $\mathbf{P}$. Then Equations (3.2) and (3.3) are written in matrix form as

$$\begin{aligned} \mathbf{l}^1 &= \mathbf{P}_0, \\ \mathbf{l}^t &= \mathbf{P}_L \mathbf{l}^{t-1}, \quad t \geqslant 2. \end{aligned} \tag{3.4}$$

The probability $P\{L = t\}$ equals the first element of $\mathbf{l}^t(0)$, i.e., $P\{L = t\} = \mathbf{e}_1(s)\mathbf{l}^t(0)$.

For any fixed t, the computation of $\mathbf{l}^t(0)$ can be done recursively by exploiting (3.4). Indeed, because of the lower Hessenberg matrix structure of $\mathbf{P}_L$, (3.4) assumes the block form:

$$\mathbf{l}^t(0) = \mathbf{A}^*_{00}\mathbf{l}^{t-1}(0) + \mathbf{A}^*_{01}\mathbf{l}^{t-1}(1), \tag{3.5}$$

$$\mathbf{l}^t(j) = \mathbf{A}^*_{j0}\mathbf{l}^{t-1}(0) + \sum_{i=1}^{j+1} \mathbf{A}_{ji}\mathbf{l}^{t-1}(i), \quad 1 \leqslant j \leqslant s, \tag{3.6}$$

$$\mathbf{l}^t(j) = \sum_{i=j-s}^{j+1} \mathbf{A}_{ji}\mathbf{l}^{t-1}(i), \quad j \geqslant s+1, \tag{3.7}$$

where $\mathbf{A}^*_{00}$ is obtained from $\mathbf{A}_{00}$ by deleting the first row and column. Analogously, $\mathbf{A}^*_{01}$ follows from $\mathbf{A}_{01}$ by deleting the first row, while the matrices $\mathbf{A}^*_{j0}$, for $1 \leqslant j \leqslant s$, are obtained from $\mathbf{A}_{j0}$ by removing the first column.

Observe that to compute $\mathbf{l}^t(0)$ we need $\mathbf{l}^{t-1}(0)$ and $\mathbf{l}^{t-1}(1)$ so we need $\mathbf{l}^{t-2}(0)$, $\mathbf{l}^{t-2}(1)$ and $\mathbf{l}^{t-2}(2)$, etc. Hence, for the computation, $P\{L = t\}$ we need to begin by computing the vectors $\mathbf{l}^1(0), \mathbf{l}^1(1), \ldots, \mathbf{l}^1(t-1)$. It should be pointed out that it is impossible to move in one transition from states with at least one customer in orbit to the state $(0,0)$. Hence $\mathbf{l}^1(j) = 0$, for $j \geqslant 1$, and using (3.7), we find that $\mathbf{l}^i(j) = 0$, for $i \geqslant 1$ and $j \geqslant (i-1)s + 1$.

For a given level of accuracy $\epsilon > 0$, we may compute $P\{L = t\}$ until the $(1-\epsilon)$ percentile of L has been achieved, i.e., we stop the computations at a level $\mathbb{K}$ such that $P\{L \leqslant \mathbb{K} - 1\} \leqslant 1 - \epsilon < P\{L \leqslant \mathbb{K}\}$. The idea now is to approximate the moment $E[L^n]$ by the truncated moment $\sum_{t=0}^{\mathbb{K}} t^n P\{L = t\}$. However, this method neglects small probabilities, $P\{L = t\}$, which are multiplied by large numbers t^n. An alternative approach may be attained by considering the direct truncation model with a transition matrix $\overline{\mathbf{P}}(\mathbb{K})$. If we denote by $L^{(K)}$ the length of a busy period in the truncated model, then we have

$$P\{L = t\} = P\{L^{(\mathbb{K})} = t\}, \quad 0 \leqslant t \leqslant \mathbb{K}.$$

This happens because the event $\{L = t\}$ with $t \leqslant \mathbb{K}$ implies that the process $\{(C_t, U_t); t \geqslant 0\}$ has not visited $l(\mathbb{K})$ during the busy period under consideration. Then, the transition probabilities of the paths that are counted in $\{L = t\}$ are equal whether we use the matrix $\mathbf{P}$ or its truncation $\overline{\mathbf{P}}(t)$. It seems numerically advantageous to approximate the moments of L by the moments of $L^{(\mathbb{K})}$. Note that we can think of it $L^{(\mathbb{K})}$ as a discrete

PH distribution with representation $PH_d(\mathbf{e}_1(s+(s+1)\mathbb{K}), \overline{\mathbf{P}}_L(\mathbb{K}))$, where it $\overline{\mathbf{P}}_L(\mathbb{K})$ is obtained by neglecting the first row and column of $\overline{\mathbf{P}}(\mathbb{K})$. Using standard results for the moments of a discrete PH distribution (see e.g. Latouche and Ramaswami, 1999), we have that the factorial moments of $L^{(\mathbb{K})}$ are given by

$$E[L^{(\mathbb{K})}(L^{(\mathbb{K})}-1)\dots(L^{(\mathbb{K})}-t+1)] = t!\,\mathbf{e}_1(s+(s+1)\mathbb{K})(\mathbf{I}-\overline{\mathbf{P}}_L(\mathbb{K}))^{-t}\overline{\mathbf{P}}_L(\mathbb{K})^{t-1}\mathbf{e}(s+(s+1)\mathbb{K}), \quad t \geqslant 1. \tag{3.8}$$

For using (3.8), we need to compute the column vectors $(\mathbf{I}-\overline{\mathbf{P}}_L(\mathbb{K}))^{-t}\overline{\mathbf{P}}_L(\mathbb{K})^{t-1}\mathbf{e}(s+(s+1)\mathbb{K})$. It amounts to successfully solving linear systems of the form $(\mathbf{I}-\overline{\mathbf{P}}_L(\mathbb{K}))\mathbf{x} = \mathbf{a}$. This can be done at low computational cost by exploiting the lower block Hessenberg form of the coefficient matrix $\mathbf{I}-\overline{\mathbf{P}}_L(\mathbb{K})$.

3.3.4 Waiting time analysis

Let Z be a random variable representing the waiting time that a tagged customer spends in orbit. According to this definition, Z excludes the service time. We remark that a primary arrival has priority over customers in the retrial orbit and assume the RO policy among retrying customers. In what follows, we approximate the analysis of Z, the intractable $Geo/Geo/s$ retrial queue, by the parallel study for the truncated model with sufficiently large $\mathbb{K}$.

Since we deal with the G-EAS policy, the probability that a primary arrival customer finds the state (s,j) is $\bar{\lambda}^s\pi_{sj}$. Let $E(Z)_{(i,j)}$ be the conditional waiting time for a customer in orbit, given that the current system state is (i,j). We now may condition on the state viewed by the tagged customer. This yields

$$P\{Z=t\} = \bar{\lambda}^s \sum_{j=0}^{\mathbb{K}-1} \pi_{sj} P\{Z_{(s,j+1)} = t\}, \quad t \geqslant 1.$$

Moreover, we have

$$P\{Z=0\} = 1 - \bar{\lambda}^s \sum_{j=0}^{\mathbb{K}-1} \pi_{sj}.$$

The probabilities $P\{E(W)_{(i,j)} = t\}$ can be obtained recursively by conditioning on the state of the next transition. We have the first-step analysis equations.

$$P\{Z_{(i,j)} = 1\} = \sum_{(m,n)\in S^*_{(i,j)}} P_{(i,j)(m,n)} \frac{j-n}{j}, \quad 0 \leqslant i \leqslant s,\ 1 \leqslant j \leqslant \mathbb{K}, \tag{3.9}$$

$$\begin{aligned} P\{Z_{(i,j)} = t\} = &\sum_{(m,n)\in S^*_{(i,j)}} P_{(i,j)(m,n)} \frac{n}{j} P\{Z_{(m,n)} = t-1\} \\ &+\delta_{is}(1-\delta_{j\mathbb{K}})P_{(s,j)(s,j+1)}P\{Z_{(s,j+1)} = t-1\} \\ &+\delta_{is}\delta_{j\mathbb{K}}P_{(s,\mathbb{K})(s,\mathbb{K}+1)}P\{Z_{(s,\mathbb{K})} = t-1\}, \quad 0 \leqslant i \leqslant s,\ 1 \leqslant j \leqslant \mathbb{K},\ t \geqslant 2, \end{aligned} \tag{3.10}$$

where $S^*_{(i,j)} = S_{(i,j)}$, for $0 \leqslant i \leqslant s-1$; and $S^*_{(s,j)} = S_{(s,j)} - \{(s, j+1)\}$.

The above equations (3.9) and (3.10) provide a simple recursive scheme to obtain $P\{E(Z) = t\}$ up to any given level of accuracy. In particular, they can be expressed in matrix form as in the case of (3.2) and (3.3) for $P\{L = t\}$. Thus, a similar algorithm can be developed.

By multiplying (3.9) by z and (3.10) by z^t, and adding for all t, we obtain recursive relations for the corresponding probability generating functions $Z_{(i,j)}(z) = \sum_{t=1}^{\infty} P\{Z_{(i,j)} = t\}z^t$:

$$\begin{aligned} Z_{(i,j)}(z) = \sum_{(m,n)\in S^*_{(i,j)}} P_{(i,j)(m,n)} \left(\frac{j-n}{j} z + \frac{n}{j} z Z_{(m,n)}(z)\right) \\ +\delta_{is}(1-\delta_{jK})P_{(s,j)(s,j+1)} z Z_{(s,j+1)}(z) \\ +\delta_{is}\delta_{j\mathbb{K}} P_{(s,\mathbb{K})(s,\mathbb{K}+1)} z Z_{(s,\mathbb{K})}(z), \quad 0 \leqslant i \leqslant s,\ 1 \leqslant j \leqslant \mathbb{K}. \end{aligned}$$

By differentiating at $z = 1$, we get recursive relations for the factorial conditional waiting time moments $Z^{(t)}_{(i,j)} = E[Z_{(i,j)} \ldots (Z_{(i,j)} - t + 1)]$ for $t \geqslant 1$. In particular, we have

$$\begin{aligned} Z^{(t)}_{(i,j)} = \sum_{(m,n)\in S^*_{(i,j)}} P_{(i,j)(m,n)} \left(\frac{j-n}{j}\delta_{t1} + \frac{n}{j}\left(Z^{(t)}_{(m,n)} + tZ^{(t-1)}_{(m,n)}\right)\right) \\ +\delta_{is}(1-\delta_{j\mathbb{K}})P_{(s,j)(s,j+1)}\left(Z^{(t)}_{(s,j+1)} + tZ^{(t-1)}_{(s,j+1)}\right) \\ +\delta_{is}\delta_{j\mathbb{K}} P_{(s,\mathbb{K})(s,\mathbb{K}+1)}\left(Z^{(t)}_{(s,\mathbb{K})} + tZ^{(t-1)}_{(s,t)}\right), \quad 0 \leqslant i \leqslant s,\ 1 \leqslant j \leqslant \mathbb{K},\ t \geqslant 1, \end{aligned}$$

and consequently

$$E[Z] = \bar{\lambda}^s \sum_{j=0}^{\mathbb{K}-1} \pi_{sj} Z^{(1)}_{(s,j+1)}.$$

3.4 Geo/G/1 Queueing System with Callbacks

In this section, we focus on the analysis of a discrete-time Geo/G/1 retrial queue. In this model, arrivals follow a geometric distribution, while service times are generally distributed. To simplify the analysis and avoid tracking multiple retrial attempts, we assume that only one customer from the orbit is allowed to retry at a time. This assumption makes the model analytically tractable and allows the derivation of key performance measures.

The study is based on the Markov chain formulation of the system, through which we derive the stability condition, the stationary distribution, and several important performance characteristics of the queue.

3.4.1 System Description

Let us consider a discrete-time single-server retrial queue where the time axis is segmented into a sequence of equal time intervals (called slots). Further, let the time axis be marked by $0, 1, \ldots, m, \ldots$. It is assumed that all queueing activities (arrivals, departures, and retrials) take place around the slot boundaries. For mathematical clarity, we assume that a potential departure occurs in the interval (m^-, m), and a potential arrival or retrial occurs in the interval (m, m^+); that is, arrivals and retrials occur at the moment immediately after the slot boundaries, and the departures occur at the moment immediately before the slot boundaries.

Customers arrive according to a geometrical arrival process with rate p, i.e., p is the probability that an arrival occurs in a slot. If an arriving customer finds the server free, he immediately begins his service. Otherwise, the customer leaves the service area and enters a group of blocked customers called "orbit" in accordance with an FCFS discipline. We will assume that only the customer at the head of the orbit is allowed access to the server. It is always supposed that retrials and services can be started only at slot boundaries, and their durations are integral multiples of a slot duration. Successive inter-retrial times of any customer are governed by an arbitrary distribution $\{r_i\}_{i=0}^{\infty}$ with a generating function $A(x) = \sum_{i=0}^{\infty} r_i x^i$. Service times are independent and identically distributed with a general distribution $\{s_i\}_{i=1}^{\infty}$, generating function $S(x) = \sum_{i=1}^{\infty} s_i x^i$ and nth factorial moments β_n. After service completion, the served customer leaves the system forever and will have no further effect on the system. In order to avoid trivial cases, we assume $0 < p < 1$. The traffic intensity is given by $\rho = p\beta_1$.

3.4.2 Markov Chain

At time m^+ (the instant immediately after time slot m), the system can be described by the process

$$Y_m = (C_m, \xi_{0,m}, \xi_{1,m}, U_m),$$

where C_m represents the server state (0 or 1 according to whether the server is free or busy, respectively) and U_m is the number of customers in the retrial group. If $C_m = 0$ and $U_m > 0$, $\xi_{0,m}$ denotes the remaining retrial time. If $C_m = 1$, $\xi_{1,m}$ corresponds to the remaining service time of the customer being served.

It can be shown that $\{Y_m, m \in \mathbb{N}\}$ is the Markov chain of our queueing system, whose state space is

$$\{(0,0); (0,i,k) : i \geq 1, k \geq 1; (1,i,k) : i \geq 1, k \geq 0\}.$$

Our objective is to find the stationary distribution.

$$\begin{aligned}
\pi_{0,0} &= \lim_{m\to\infty} P[C_m = 0, U_m = 0], \\
\pi_{0,i,k} &= \lim_{m\to\infty} P[C_m = 0, \xi_{0,m} = i, U_m = k]; \quad i \geq 1, k \geq 1, \\
\pi_{1,i,k} &= \lim_{m\to\infty} P[C_m = 1, \xi_{1,m} = i, U_m = k]; \quad i \geq 1, k \geq 0,
\end{aligned}$$

of the Markov chain $\{Y_m, m \in \mathbb{N}\}$.

The one-step transition probabilities $p_{yy'} = P[Y_{m+1} = y'|Y_m = y]$ are given by the formulae

$$\begin{aligned} p_{(0,0)(0,0)} &= \bar{p}, \\ p_{(1,1,0)(0,0)} &= \bar{p}. \end{aligned}$$

if $i \geq 1, k \geq 1$:

$$\begin{aligned} p_{(0,i+1,k)(0,i,k)} &= \bar{p}, \\ p_{(1,1,k)(0,i,k)} &= \bar{p}r_i, \end{aligned}$$

if $i \geq 1, k \geq 0$:

$$\begin{aligned} P_{(0,0)(1,i,k)} &= ps_i, \quad k = 0 \\ P_{(0,1,k+1)(1,i,k)} &= \bar{p}s_i, \end{aligned}$$

if $i \geq 1, j \geq 1$ et $k \geq 1$:

$$\begin{aligned} P_{(0,j,k)(1,i,k)} &= ps_i, \\ P_{(1,1,k)(1,i,k)} &= ps_i, \\ P_{(1,1,k+1)(1,i,k)} &= \bar{p}s_i r_0, \end{aligned}$$

if $i \geq 1, k \geq 1$:

$$\begin{aligned} P_{(1,i+1,k-1)(1,i,k)} &= p, \\ P_{(1,i+1,k)(1,i,k)} &= \bar{p}, \end{aligned}$$

Where $\bar{p} = (1 - p)$.

The Kolmogorov equations for the stationary distribution are

$$\pi_{0,0} = \bar{p}\pi_{0,0} + \bar{p}\pi_{1,1,0}, \tag{3.11}$$

$$\pi_{0,i,k} = \bar{p}\pi_{0,i+1,k} + \bar{p}r_i\pi_{1,1,k}; \quad i \geq 1, k \geq 1, \tag{3.12}$$

$$\begin{aligned} \pi_{1,i,k} = {} & \delta_{0k}ps_i\pi_{0,0} + \bar{p}s_i\pi_{0,1,k+1} + (1 - \delta_{0k})ps_i \sum_{j=1}^{\infty} \pi_{0,j,k} \\ & + ps_i\pi_{1,1,k} + \bar{p}r_0s_i\pi_{1,1,k+1} + (1 - \delta_{0k})p\pi_{1,i+1,k-1} + \bar{p}\pi_{1,i+1,k}; \\ & i \geq 1, k \geq 0, \end{aligned} \tag{3.13}$$

and the normalization condition is

$$\pi_{0,0} + \sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\pi_{0,i,k} + \sum_{i=1}^{\infty}\sum_{k=0}^{\infty}\pi_{1,i,k} = 1.$$

In order to solve (3.11)–(3.13), we introduce the generating functions

$$\phi_0(x,z) = \sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\pi_{0,i,k}x^i z^k,$$
$$\phi_1(x,z) = \sum_{i=1}^{\infty}\sum_{k=0}^{\infty}\pi_{1,i,k}x^i z^k$$

and the auxiliary generating functions

$$\phi_{0,i}(z) = \sum_{k=1}^{\infty}\pi_{0,i,k}z^k, \quad i \geq 1,$$
$$\phi_{1,i}(z) = \sum_{k=0}^{\infty}\pi_{1,i,k}z^k, \quad i \geq 1.$$

Multiplying (3.12) and, (3.13) by z^k and summing over k, these equations become

$$\phi_{0,i}(z) = \bar{p}\phi_{0,i+1}(z) + \bar{p}r_i\phi_{1,1}(z) - \bar{p}r_i\pi_{1,1,0}, \quad i \geq 1, \tag{3.14}$$

$$\begin{aligned}\phi_{1,i}(z) = (\bar{p}+pz)\phi_{1,i+1}(z) + ps_i\phi_0(1,z) + \frac{\bar{p}r_0 + pz}{z}s_i\phi_{1,1}(z)\\ + \frac{\bar{p}}{z}s_i\phi_{0,1}(z) - \frac{\bar{p}r_0}{z}s_i\pi_{1,1,0} + ps_i\pi_{0,0}, \quad i \geq 1.\end{aligned} \tag{3.15}$$

By substituting equation (3.11) into equations (3.14) and (3.15), we get

$$\phi_{0,i}(z) = \bar{p}\phi_{0,i+1}(z) + \bar{p}r_i\phi_{1,1}(z) - pr_i\pi_{0,0}, \quad i \geq 1, \tag{3.16}$$

$$\begin{aligned}\phi_{1,i}(z) = (\bar{p}+pz)\phi_{1,i+1}(z) + ps_i\phi_0(1,z) + \frac{\bar{p}r_0 + pz}{z}s_i\phi_{1,1}(z)\\ + \frac{\bar{p}}{z}s_i\phi_{0,1}(z) + \frac{p(z-r_0)}{z}s_i\pi_{0,0}, \quad i \geq 1.\end{aligned} \tag{3.17}$$

Then, multiplying (3.16) and (3.17) by x^i and summing over i leads to

$$\frac{x-\bar{p}}{x}\phi_0(x,z) = \bar{p}[A(x)-r_0]\phi_{1,1}(z) - \bar{p}\phi_{0,1}(z) - p[A(x)-r_0]\pi_{0,0}, \tag{3.18}$$

$$\begin{aligned}\frac{x-(\bar{p}+pz)}{x}\phi_1(x,z) = \left[\frac{\bar{p}r_0+pz}{z}S(x) - (\bar{p}+pz)\right]\phi_{1,1}(z)\\ + \frac{\bar{p}}{z}S(x)\phi_{0,1}(z) + pS(x)\phi_0(1,z) + \frac{p(z-r_0)}{z}S(x)\pi_{0,0}.\end{aligned} \tag{3.19}$$

Choosing $x = 1$ in (3.18) yields

$$p\phi_0(1, z) = \bar{p}(1 - r_0)\phi_{1,1}(z) - \bar{p}\phi_{0,1}(z) - p(1 - r_0)\pi_{0,0}$$

and substituting the above equation into (3.19), we get

$$\begin{aligned}\frac{x - (\bar{p} + pz)}{x}\phi_1(x, z) &= \left[\frac{z + \bar{p}r_0(1 - z)}{z}S(x) - (\bar{p} + pz)\right]\phi_{1,1}(z) \\ &\quad + \frac{\bar{p}(1 - z)}{z}S(x)\phi_{0,1}(z) - \frac{pr_0(1 - z)}{z}S(x)\pi_{0,0}.\end{aligned} \tag{3.20}$$

Setting $x = \bar{p}$ in (3.18) and $x = \bar{p} + pz$ in (3.20), we obtain

$$\begin{aligned}p[A(\bar{p}) - r_0]\pi_{0,0} &= \bar{p}[A(\bar{p}) - r_0]\phi_{1,1}(z) - \bar{p}\phi_{0,1}(z), \\ \frac{pr_0(1 - z)}{z}S(\bar{p} + pz)\pi_{0,0} &= \left[\frac{z + \bar{p}r_0(1 - z)}{z}S(\bar{p} + pz) - (\bar{p} + pz)\right]\phi_{1,1}(z) \\ &\quad + \frac{\bar{p}(1 - z)}{z}S(\bar{p} + pz)\phi_{0,1}(z).\end{aligned}$$

And from this system of equations, we find the generating functions $\phi_{1,1}(z)$ and $\phi_{0,1}(z)$:

$$\phi_{1,1}(z) = \frac{pA(\bar{p})(1 - z)S(\bar{p} + pz)}{\bar{p}A(\bar{p})(1 - z)S(\bar{p} + pz) - z[(\bar{p} + pz) - S(\bar{p} + pz)]}\pi_{0,0}, \tag{3.21}$$

$$\phi_{0,1}(z) = \frac{pz[A(\bar{p}) - r_0][(\bar{p} + pz) - S(\bar{p} + pz)]}{\bar{p}A(\bar{p})(1 - z)S(\bar{p} + pz) - z[(\bar{p} + pz) - S(\bar{p} + pz)]}\frac{\pi_{0,0}}{\bar{p}}. \tag{3.22}$$

Substituting (3.21) and (3.22) into (3.18) and (3.20), respectively, yields the following generating functions:

$$\phi_0(x, z) = \frac{A(x) - A(\bar{p})}{x - \bar{p}} \cdot \frac{pxz[(\bar{p} + pz) - S(\bar{p} + pz)]}{\bar{p}A(\bar{p})(1 - z)S(\bar{p} + pz) - z[(\bar{p} + pz) - S(\bar{p} + pz)]}\pi_{0,0},$$

$$\phi_1(x, z) = \frac{S(x) - S(\bar{p} + pz)}{x - (\bar{p} + pz)} \cdot \frac{pxA(\bar{p})(1 - z)(\bar{p} + pz)}{\bar{p}A(\bar{p})(1 - z)S(\bar{p} + pz) - z[(\bar{p} + pz) - S(\bar{p} + pz)]}\pi_{0,0}.$$

Using the normalization condition $\pi_{0,0} + \phi_0(1, 1) + \phi_1(1, 1) = 1$, we obtain:

$$\pi_{0,0} = \frac{p + \bar{p}A(\bar{p}) - \rho}{A(\bar{p})}. \tag{3.23}$$

According to (3.23), since $\pi_{0,0} > 0$, we obtain the ergodicity condition for the Markov chain:

$$p\beta_1 < \bar{p}A(\bar{p}).$$

This condition can be written as $p(1-\bar{p}) < \bar{p}A(\bar{p})$, where the first term is the average number of primary customers arriving during the service interval and the second term represents the mean number of secondary customers (callbacks) entering service at the start of service.

Note 3.4.2.1. This condition states that primary clients must arrive more slowly than clients who make follow-up requests (on average). If the condition $p\beta_1 < \bar{p}A(\bar{p})$ is satisfied, the system is stable.

The results above are summarized in the following theorem:

Theorem 3.4.2.1. *[11]*

The Markov chain $\{Y_m, m \in \mathbb{N}\}$ is ergodic if and only if

$$\rho < p + \bar{p}A(\bar{p}).$$

The generating functions of the stationary distribution of the chain are given by

$$\phi_0(x,z) = \frac{A(x) - A(\bar{p})}{x - \bar{p}} \cdot \frac{pxz[(\bar{p}+pz) - S(\bar{p}+pz)]\pi_{0,0}}{\bar{p}A(\bar{p})(1-z)S(\bar{p}+pz) - z[(\bar{p}+pz) - S(\bar{p}+pz)]},$$

$$\phi_1(x,z) = \frac{S(x) - S(\bar{p}+pz)}{x - (\bar{p}+pz)} \cdot \frac{pxA(\bar{p})(1-z)(\bar{p}+pz)\pi_{0,0}}{\bar{p}A(\bar{p})(1-z)S(\bar{p}+pz) - z[(\bar{p}+pz) - S(\bar{p}+pz)]},$$

where

$$\pi_{0,0} = \frac{p + \bar{p}A(\bar{p}) - \rho}{A(\bar{p})}.$$

corollary 3.4.2.2. *[11]*

1. The marginal generating function of the number of customers in the retrial group when the server is idle is given by

$$\pi_{0,0} + \phi_0(1,z) = \frac{A(\bar{p})(\bar{p}+pz)[S(\bar{p}+pz) - z]\pi_{0,0}}{\bar{p}A(\bar{p})(1-z)S(\bar{p}+pz) - z[(\bar{p}+pz) - S(\bar{p}+pz)]}.$$

2. The marginal generating function of the number of customers in the retrial group when the server is busy is given by

$$\phi_1(1,z) = \frac{A(\bar{p})(\bar{p}+pz)[1 - S(\bar{p}+pz)]\pi_{0,0}}{\bar{p}A(\bar{p})(1-z)S(\bar{p}+pz) - z[(\bar{p}+pz) - S(\bar{p}+pz)]}.$$

3. The probability generating function of the number of customers in the retrial group is given by

$$\begin{aligned}\Phi(z) &= \pi_{0,0} + \phi_0(1,z) + \phi_1(1,z) \\ &= \frac{A(\bar{p})(1-z)(\bar{p}+pz)\pi_{0,0}}{\bar{p}A(\bar{p})(1-z)S(\bar{p}+pz) - z[(\bar{p}+pz) - S(\bar{p}+pz)]}.\end{aligned} \tag{3.24}$$

4. The probability generating function of the number of customers in the system is given by

$$\begin{aligned}\Psi(z) &= \pi_{0,0} + \phi_0(1,z) + z\phi_1(1,z) \\ &= \frac{A(\bar{p})(1-z)(\bar{p}+pz)S(\bar{p}+pz)\pi_{0,0}}{\bar{p}A(\bar{p})(1-z)S(\bar{p}+pz) - z[(\bar{p}+pz) - S(\bar{p}+pz)]}.\end{aligned} \tag{3.25}$$

3.4.3 System Characteristics

The Geo/G/1 queuing system with callbacks can be characterized by the following performance measures:

1. The system is free with probability

$$\pi_{0,0} = \frac{p + \bar{p}A(\bar{p}) - \rho}{A(\bar{p})}.$$

2. The system is occupied with probability

$$\phi_0(1,1) + \phi_1(1,1) = \frac{\rho - p[1 - A(\bar{p})]}{A(\bar{p})}.$$

3. The server is idle with probability

$$\pi_{0,0} + \phi_0(1,1) = 1 - \rho.$$

4. The server is busy with probability

$$\phi_1(1,1) = \rho.$$

5. The mean number of customers in the orbit is

$$V = \Phi'(1) = \frac{2\bar{p}(\rho - p)[1 - A(\bar{p})] + p^2\beta_2}{2[p + \bar{p}A(\bar{p}) - \rho]}.$$

6. The mean number of customers in the system is

$$\bar{N} = \Psi'(1) = \rho + \frac{2\bar{p}(\rho - p)[1 - A(\bar{p})] + p^2\beta_2}{2[p + \bar{p}A(\bar{p}) - \rho]}.$$

7. The mean time a customer spends in the system is given by

$$T = \frac{\bar{N}}{p} = \beta_1 + \frac{2\bar{p}(\beta_1 - 1)[1 - A(\bar{p})] + p\beta_2}{2[p + \bar{p}A(\bar{p}) - \rho]}.$$

8. The mean waiting time for a customer in the orbit:

$$Z = \frac{V}{p} = \frac{2\bar{p}(\beta_1 - p)[1 - A(\bar{p})] + p\beta_2}{2[p + \bar{p}A(\bar{p}) - \rho]}.$$

Note 3.4.3.1 The stationary distribution of the server state is given by:

$$\pi_{0,0} + \phi_0(1,1) = 1 - \rho; \quad \phi_1(1,1) = \rho,$$

depends only on the distribution of the service time through its mean β_1 and not on the distribution of the inter-arrival times.

Note 3.4.3.2. (Special case) When $r_0 = 1$, $\Phi(z)$ is reduced to:

$$\Phi(z) = \frac{(1-\rho)(1-z)S(\bar{p}+pz)}{S(\bar{p}+pz)-z},$$

which is the probability generating function of the number of customers in the $Geo/G/1/\infty$ queueing system *[7]*. This result is not surprising because when $r_0 = 1$, the customer is at the head of the orbit, and service begins immediately when the server is free.

Chapter 4

Model Simulations

4.1 Introduction

In this chapter, we will simulate classical discrete-time queueing models **(Geo/Geo/1 and Geo/Geo/s)** and advanced discrete-time queueing models **(Geo/G/1 with call-backs)**.

4.2 The Geo/Geo/1 Queueing Model

4.2.1 System characteristics as a function of parameter variation p

Based on this table, we can find numerical results for different components, such as traffic intensity ρ, the mean number of customers in the system $\bar{N}$, the mean number of customers waiting in the queue Q, the mean time a customer spends in the system T, and the mean time a customer spends just waiting in the queue W. According to the variation in arrival probability p in the Geo/Geo/1 model.

We fixed $\lambda = 0.9$

p	ρ	$\overline{N}$	Q	T	W
0.1	0.01234568	0.1125000	0.001388889	1.125000	0.01388889
0.2	0.02777778	0.2285714	0.006349206	1.142857	0.03174603
0.3	0.04761905	0.3500000	0.016666667	1.166667	0.05555556
0.4	0.07407407	0.4800000	0.035555556	1.200000	0.08888889
0.5	0.11111111	0.6250000	0.069444444	1.250000	0.13888889
0.6	0.16666667	0.8000000	0.133333333	1.333333	0.22222222
0.7	0.25925926	1.0500000	0.272222222	1.500000	0.38888889
0.8	0.44444444	1.6000000	0.711111111	2.000000	0.88888889

Table 4.1: Variation of Geo/Geo/1 system characteristics as a function of p

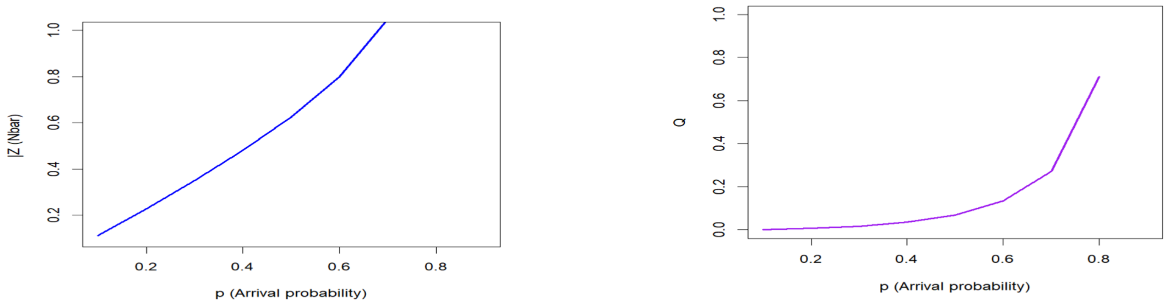

Figure 4.1: The mean number of customers in the system $\bar{N}$ and in the queue Q

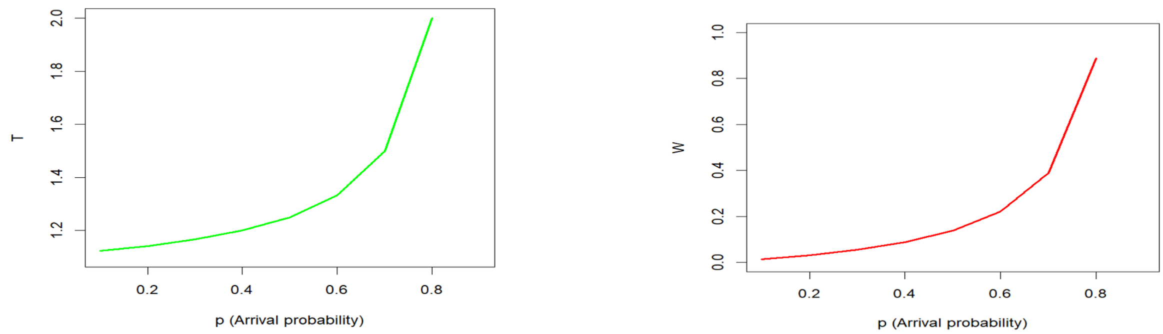

Figure 4.2: The mean time spent in the system T and in the queue W

Interpretation of Figures 4.1 and 4.2: ***Influence of Arrival Probability*** p
These two figures analyze the degradation of system performance as the frequency of arrivals increases.

Correlation between Number and Time: As the arrival probability p increases, we observe a simultaneous growth in the mean number of customers in the system $\bar{N}$ and in the queue Q, and the mean time spent in the system T and in the queue W. This is a direct illustration of **Little's Law**, which links the mean number of customers to the mean waiting time.

4.2.2 System characteristics as a function of parameter variation λ

On this table, we can find numerical results for different parameters, according to the variation in service probability λ in the Geo/Geo/1 model.

We fixed $p = 0.2$

λ	ρ	$\overline{N}$	Q	T	W
0.3	0.58333333	1.6000000	0.933333333	8.000000	4.66666667
0.4	0.37500000	0.8000000	0.300000000	4.000000	1.50000000
0.5	0.25000000	0.5333333	0.133333333	2.666667	0.66666667
0.6	0.16666667	0.4000000	0.066666667	2.000000	0.33333333
0.7	0.10714286	0.3200000	0.034285714	1.600000	0.17142857
0.8	0.06250000	0.2666667	0.016666667	1.333333	0.08333333
0.9	0.02777778	0.2285714	0.006349206	1.142857	0.03174603

Table 4.2: Variation of Geo/Geo/1 system characteristics as a function of λ

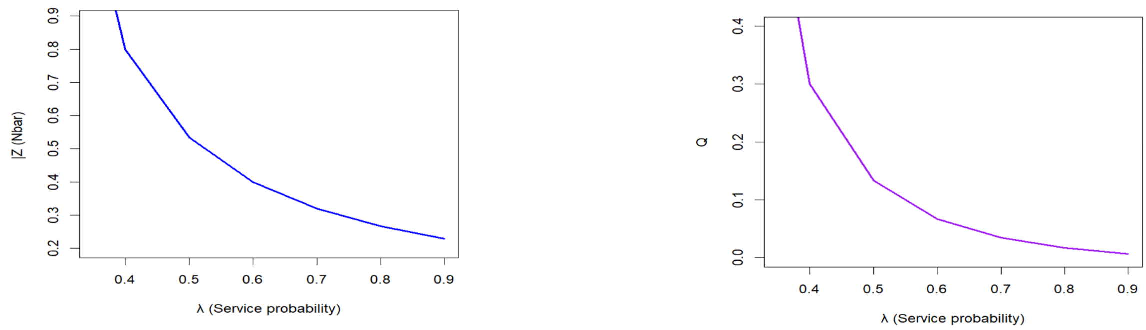

Figure 4.3: The mean number of customers in the system $\bar{N}$ and in the queue Q

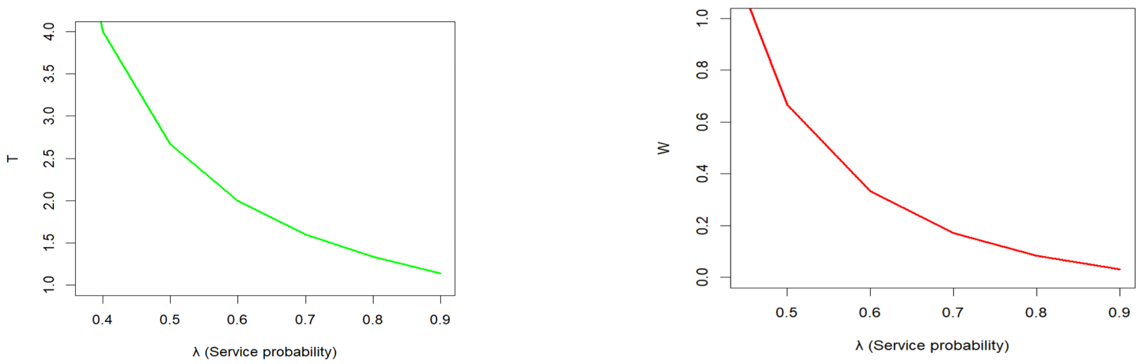

Figure 4.4: The mean time spent in the system T and in the queue W

Interpretation of Figures 4.3 and 4.4: ***Influence of Service Probability*** λ
These two figures show the improvement in system performance as server efficiency or speed increases.

Transition from Instability to Stability: For low service probability values (λ near 0.4), the system is near a breaking point: the mean number of customers in the system $\bar{N}$ and in the queue Q, and the mean time spent in the system T and in the queue W are at their maximum (the curves are very high on the left).

The Effect of Efficiency: As soon as the service probability increases (towards 0.6 and above), there is a sharp and radical drop in all indicators. The mean number of customers in the system $\bar{N}$ and in the queue Q, and the mean time spent in the system T and in the queue W, both approach their minimum values.

4.3 The Geo/Geo/s Queueing Model

4.3.1 The Geo/Geo/s Queueing Model as a function of p

From the following graphs, we can see that the parameter values, the mean number of customers in the system $\bar{N}$ and in the queue Q, and the mean time spent in the system T and in the queue W vary as the probability of arrivals $p$increases.
There is an arrival in the queue as p increases. This results in an increase in the number of customers in the system, and as shown in the graphs, we observe that the number of servers increases, with some servers handling more customers than others.

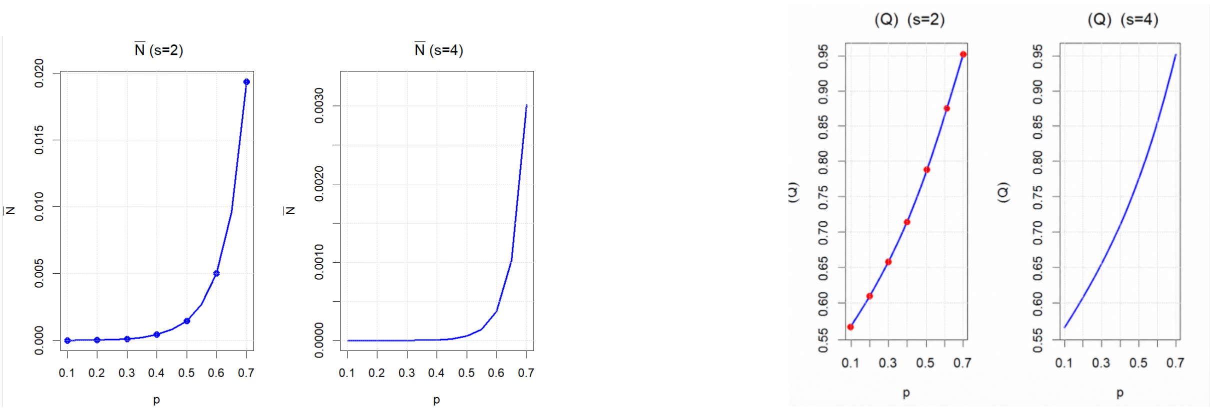

Figure 4.5: The mean number of customers in the system $\bar{N}$ and in the queue Q as a function of p

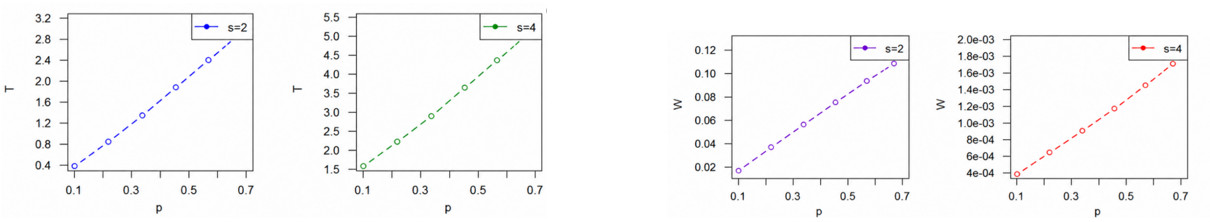

Figure 4.6: The mean time spent in the system T and in the queue W as a function of p

4.4 The Geo/G/1 Queueing Model with Callbacks

4.4.1 The Geo/Geo/1 Queueing Model with Callbacks

Consider the Geo/Geo/1 system with discrete-time callbacks. Seriously, this third moment in the stem has been described through geometric access to the parameter p. We assume that it's geometric with parameter r, and that its generating function is $A(x) = (1-r)/(1-rx)$. The service time must be independent and distributed equally according to the law with geometrical parameter s, first, first and third factorial moment β_1, β_2. The 4.5. You know what? Three graphs (a), (b), and (c) are presented in Fig. Seriously, in this case we set $\beta_1 = 2$ and $p = 0.2, 0.3, 0.4$.

- Graph (a) shows the change in the probability of the system, system not free, $(1 - \pi_{0,0})$ as a function of r. Guess what? We see that the probability $(1 - \pi_{0.0})$ increases towards 1. Each time the value of r approaches 1, we also observe that at every time p increases, the probability $(1 - \pi_{0.0})$ tends towards 1 more quickly.

- Graph (b) shows the mean number of customers in the system $\bar{N}$ as a function of r. You know what? Same situation. Other Observations: Note that as r approaches 1, the mean number of customers in the system, $\bar{N}$, tends toward infinity.

- Graph (c) shows the mean number of customers in the orbit V by changing the ratio of reminders r, we see that V increases and goes to infinity as a function of p, every time r approaches 1, and for low values, the orbit r is empty.

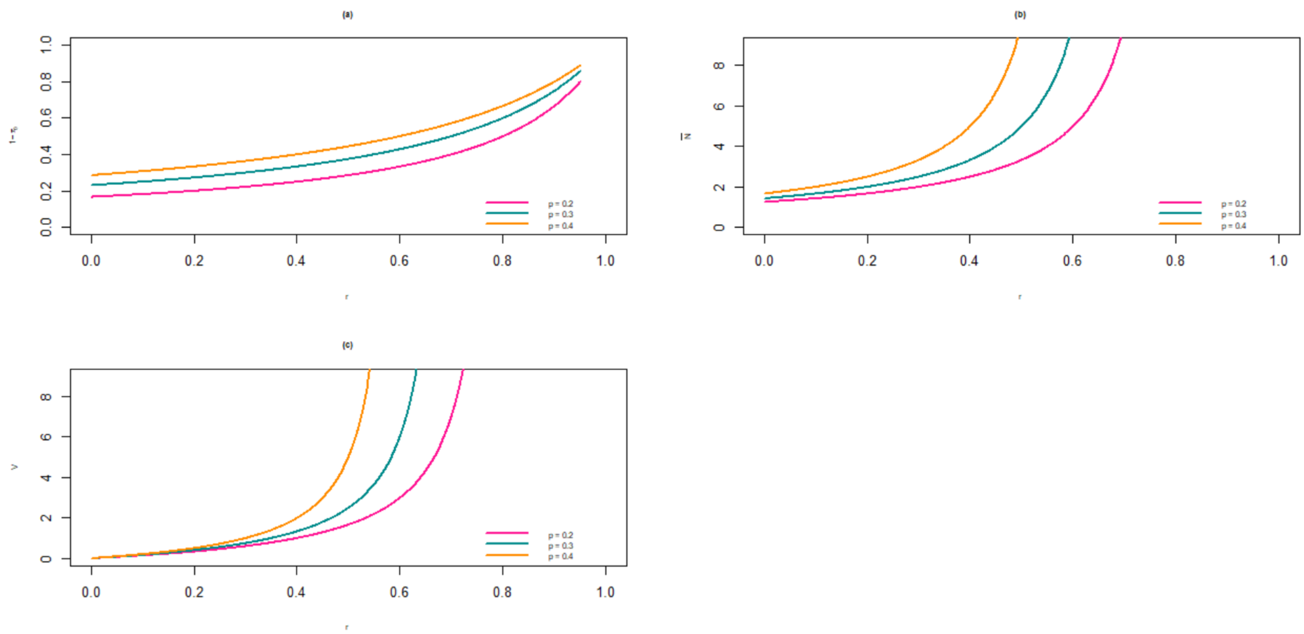

Figure 4.7: (a) The probability that the system is occupied as a function of r. (b) The mean number of customers in the system. (c) The mean number of customers in the orbit.

Conclusion

In conclusion, this memory, which focuses on the study of advanced discrete-time queueing models, demonstrates the high effectiveness of this approach, particularly in representing modern technological systems that rely on the division of time into time slots, such as digital communication networks and computer systems.
In this memory, our approach was to progress linearly from theoretical fundamentals to increasingly complex and realistic models.

- First Chapter One: We laid the groundwork for discrete-time queueing theory. We introduced key concepts, such as the Bernoulli distribution and the geometric distribution, which allowed us to formalize arrival and service processes. We also discussed the queuing system, including system capacity, service discipline, and performance measures.

- Second chapter: We then explored the classical Markovian discrete-time approach (Classical Markovian Discrete-Time Queues). Through the study of the fundamental models $Geo/Geo/1$ and $Geo/Geo/s$, we described their mathematical behavior, constructed the Markov chains associated with their states, and derived their key performance indicators (mean waiting time, queue length, and utilization rate).

- Third chapter: We have broadened our scope to include more complex and realistic models (Advanced Discrete-Time Queueing Models). These extensions allow us to move beyond the simplifications of classical models to accommodate more irregular data flows, correlated arrival processes, or more sophisticated service strategies.

Finally, the last chapter presents simulations of the $Geo/Geo/1$, $Geo/Geo/s$, and $Geo/G/1$ with recall models. This study allowed us to analyze the performance of discrete-time queuing systems and to confirm the theoretical results obtained in the previous chapters.

Bibliography

1. Aissani, A. and J.R. Artalejo, On the single server retrial queue subject to breakdown, Queueing Sys., 30 (1998), 309-321.

2. Aya. BOUDALLI, différents types de files d'attente à temps discret, Mémoire de Master, Année universitaire 2022–2023.

3. Agner Krarup Erlang, Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges, The Post Office Electrical Engineers' Journal, 1917.

4. Abdel-Karim Aboul-Hassan, Sherif I. Rabia, and Ahmed Kadry, Analytical study of a discrete time retrial queue with balking customers and early arrival scheme, Alexandria Engineering Journal, 44 (2005), No. 6, 911-917

5. Alfa, A.S.: Applied Discrete-Time Queues. Springer, Berlin (2016)

6. Bruneel, H., Kim, B. G. (1993). Discrete-Time Models for Communication Systems Including ATM. Kluwer Academic Publishers.

7. C. CHABRIAC. *Stochastic processes and modeling.* ISMAG MASTER 2, 2012-2013.

8. Frederick Solomon, Probability and Stochastic Processes, Prentice Hall, 2007.

9. J. JACDUES et R.M.L. CHABNOL. *CHAÎNES DE MARKOV*. Préparation à l'Agrégation Bordeaux 1, 2012 - 2013

10. J. Kim. TRANSIENT ANALYSIS OF THE GEO/GEO/1 QUEUE, *Journal of the Changcheng Mathematical Society,* Volume 21, No. 3, September 2008.

11. J. J. Hunter, Mathematical Techniques of Applied Probability, *Vol. 2, Discrete-Time Models: Techniques and Applications* (Academic Press, New York, 1983).

12. J.R. Artalejo, Algorithmic analysis of the Geo/Geo/c retrial queue. European Journal of Operational Research 189 (2008) 1042–1056.

13. J.R. Artalejo, Performance analysis and optimal control of the Geo/Geo/c queue, Performance Evaluation 52 (2003) 15–39.

14. J. W. Cohen, Basic problem of telephone and in uence of repeated calls. Telecommunication review, (2), 49100, 1957.

15. J. Wang, Discrete-time Geo/G/1 retrial queues with general retrial time and Bernoulli vacation, J Syst Sci Complex, 25, 504513, 2012.

16. I. Atencia and P. Moreno, A discrete-time Geo/G/1 retrial queue with general retrial times. *Queueing Syst.* 48 (2004) 5–21.

17. I. Atencia, A Geo/G/1 retrial queueing system with priority services. European Journal of Operational Research 256 (2017) 178–186.

18. L. BOURAINE, Analysis of Queuing Systems with Callbacks and Negative Arrivals, *Ph.D. thesis,* Mouloud Mammeri University, Tizi Ouzou, 2007.

19. N. Pukazhenthi, S. Ramki, Discrete- Time Queueing Model Geo[X]/G/∞ with Bulk Arrival Rule, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020.

20. T. Nassim. Chaînes de Markov à temps discret. Cours de troisième année, mathématiques appliquées, Université Abderrahmane Mira de Béjaia, Année Universitaire 2017-2018.

21. T. Yang and H. Li (1995), On the steady-state queue size distribution of the discrete-time Geo/G/1 queue with repeated customers, *Queueing Systems* 21, 199-215.

22. U. N. Bhat. *An Introduction to Queueing Theory*, Modeling and Analysis in Applications, Second Edition, Birkhäuser.