SCHOOL OF APPLIED MATHEMATICAL
& PHYSICAL SCIENCES
DEPARTMENT OF MATHEMATICS

MASTER OF SCIENCE (M.Sc)
«APPLIED MATHEMATICAL SCIENCES»
FLOW: PROBABILITIES & STATISTICS

TITLE:

QUEUING THEORY & MARKOV CHAINS

Author : TZIMI AIKATERINI
Supervisor : Associate Proffessor Coletsos John

A thesis Submitted to the Department of Mathematics of the National Technical University of Athens in partial fulfilment of the requirements for the degree of Applied Mathematical Sciences

Examination Committee
Coletsos J. Associate Professor
Kokkinis B. Assistant Professor
Stefaneas P. Assistant Professor

Athens, 2021

Abstract

Operations Research (OR) has offered invaluable knowledge and methodologies so as various fields of everyday life, such as businesses for instance, have the opportunity to ameliorate their operations and management techniques. One fundamental concern of businesses is the fact that customers are satisfied and it is this satisfaction which comes strictly in accordance with the time spent on queues. The queuing theory has been developed in order to study the performance measures of businesses so as to give indicators, as far as the number of customers and the time spent on waiting lines are concerned. In order to face the uncertainty of the number of customers within a certain period of time, Markov chains, as a great tool of OR, are used in order to predict efficiently the arrival and service rates. By depicting transitions from state to state in matrices, Markov chains are used to prove evidence for the evolution of a process over time. Queuing theory and Markov chains have led to the construction of a plethora of Queuing models. Each queuing model corresponds to different cases of businesses queuing systems.

Περίληψη

Η επιστήμη της Επιχειρησιακής έρευνας προσφέρει πολύτιμες γνώσεις και μεθοδολογίες που αφορούν διάφορους τομείς της καθημερινότητας. Ο πιο βασικός τομέας στον οποίο συνεισφέρει, είναι αυτός των επιχειρήσεων με στόχο τη βελτίωση της λειτουργίας, της οργάνωσης και της απόδοσης τους. Βασικότερος στόχος κάθε επιχείρησης είναι η ικανοποίηση των πελατών της, μια παράμετρος η οποία επηρεάζεται σημαντικά από το χρόνο αναμονής για απόκτηση μιας υπηρεσίας. Η θεωρία των Ουρών Αναμονής έχει αναπτυχθεί με σκοπό να βελτιώσει τις παραμέτρους απόδοσης μιας επιχείρησης, όπως είναι η εκτίμηση του αναμενόμενου χρόνου αλλά και του αριθμού των πελατών σε μια ουρά αναμονής. Η αντιμετώπιση της αβεβαιότητας ύπαρξης συγκεκριμένου αιρθμού πελατών στο σύστημα για συγκεκριμένο χρονικό διάστημα έγκειται στην χρήση των Μαρκοβιανών αλυσίδων. Οι Μαρκοβιανές αλυσίδες, ως βασικό εργαλείο της επιχειρησιακής έρευνας, δίνουν τη δυνατότητα αποτελεσματικής πρόβλεψης των ρυθμών άφιξης και εξυπηρέτησης ενός συστήματος. Οι μεταβάσεις από μια κατάσταση σε μια άλλη απεικονίζονται με τη βοήθεια πινάκων, μέσω των οποίων οι Μαρκοβινές αλυσίδες δίνουν πληροφορίες για την εξέλιξη μιας διαδικασίας μέσα στο χρόνο. Ο συνδυασμός της θεωρίας των Ουρών Αναμονής με τις Μαρκοβιανές αλυσίδες οδήγησε στη δημιουργία μοντέλων ουρών αναμονής. Κάθε μοντέλο ουρών αναμονής αποσκοπεί στην μοντελοποίηση διαφορετικών περιπτώσεων συστημάτων ουρών αναμονής για τις επιχειρήσεις.

To my family

Acknowledgements

I would like to express my gratitude to all those people who have supported me during this journey.
First and foremost, I would like to express my deepest appreciation to my supervisor Mr. Coletsos Jonh for his expertise, patience and support when it was most needed.
I would also like to express my sincere thanks to my fellow classmate, Matsavelas Nikos for always offering me his help through the writing of this dissertation.
Last but not least, I am extremely gratefull to having always my family by my side. They always tirelessly support me and believe in me in every decision I get to take in my life.

Contents

Introduction

Queuing theory is widely used in many aspects of our every day life. From waiting lines in supermarkets and banks to queues in manufacturing procedures, queues are embedded in performance and time management as far as daily schedule is concerned. It is therefore crucial that a queuing theory be studied, analyzed and improved continuously. This thesis constitutes a representation of the basic knowledge needed in order to comprehend how queuing systems are formed and applied. In the first chapter, an introduction to Operations Research (OR) is presented. Not only the origin and the definition of this special scientific field be illustrated, but also the definition of the basic steps of a successful implementation of OR. In the second chapter, a great tool of OR is introduced, the Markov chains. Through Markov chains an evolution of a process can be predicted by taking into consideration only the present state of the process and by ignoring the past events. Transitions from state to state are depicted as transition matrices in order to calculate the probability a specific transition may occur within a specific period of time, a methodology which can be applied in various fields. Queuing theory is one of this fields, whose a great insight of the basics are illustrated extensively in chapters 3 and 4. The core of queuing models is being analyzed as well as the most popular tools of queuing theory, exponential distribution and Poisson process. Next, in Chapter 5, Birth and Death process is introduced as a special case of Markov chains with continuous time parameter. Then on Chapter 6 some basic queuing models are presented such as $M/M/1, M/M/c$ and $E_r/M/1$ queuing model. Finally, the thesis is completed with an application of the theory presented in the previous chapters. A case study is formed regarding a bakery's queuing system where $M/M/1$ and $E_r/M/1$ queuing models are applied and then compared in terms of efficiency.

Chapter 1

Introduction to Operations Research

1.1 Operations Research: Origin

The industrial revolution was the initial cause for remarkable changes in the way of how organizations operate. The management responsibilities have been more and more focusing on the division of labor and segmentation, although certain problems have occurred.

One drawback lies on the fact that “the components of an organization grow into relatively autonomous” [14, p.1] sub organizations which set their own goals. This fact has shown that there has been some incapacity to mesh with the overall organization, its activities and objectives. In order to deal with this complexity and specialization in an organization, a need for the emergence of a new field has occurred. The new field is known as Operations Research or commonly referred to as OR.

OR holds its origin to the military services early in World War II. Scientists were asked at that time to invent new methods so as to deal with strategic and tactical methods. These methods successfully applied to fields outside military after the end of the war.

By the early 1950’s, the bloom of OR was evident. Linear programming, dynamic programming, queuing theory as well as inventory theory evolved adequately during that era. Last but not least, the computer revolution has been a crucial element taken into consideration so as to solve complex issues efficiently and rapidly. The tremendous rise of computer technology and especially the use of personal computers after the 1980's have enhanced the implementation of OR methods in dealing with large amount of computation data.

1.2 Operations Research: Definition

The term "Operations Research" refers to how operations (i.e. activities) within an organization can be managed and organized. Several areas such as financial planning or telecommunications are based on OR. The term "research" itself applies to scientific areas where certain steps are followed. From collecting data and setting the hypothesis of the solution that a certain problem might have to conducting experiments so as to test this hypothesis, all this process follows a scientific model (typically mathematical).
Moreover, OR focuses on finding solutions in order to encounter possible opposing obligations occurring among components of the organization. It must be pointed out that the main goal of the OR is to attempt to find a best solution to occurring problems among various optimal solutions. This is the reason why a team approach is necessary for OR. A number of highly trained individuals in various fields such as Mathematics, Statistics and Probability theory or Economics can work together combining the necessary experience and the variety of skills suitable so as to deal with the certain consequences the organization might have to face.

1.3 Operations Research: Construction

The success of OR in practice is based on the combination of the mathematical techniques with the creativity and experiences the OR members share. The value of teamwork is evident. As Willemain (1994) points out and is mentioned in Taha's work [25, p.40], "effective [OR] practice requires more than analytical competence: it also requires among other attributes, technical judgement (e.g. when and how to use a given technique) and skills in communication and organizational survival". The steps of a successful implementation of OR in practice are listed as follows:

1. Definition of the problem
2. Construction of the model
3. Solution of the model
4. Validation of the model
5. Implementation of the solution

Definition of the problem
In order to define the problem, the team has to:

i. Present the problem after a thorough technical analysis and make suggestions to management in order to propose associated alternatives

ii. Set the objectives the management needs to accomplish and

iii. Set the limitations the modeled system needs to apply.

It must be pointed out that the success of OR system also depends on the quality of the data that the management will provide.

Construction of the model
This stage entails the modification of the problem into mathematical relationships. If a standard mathematical model can be applied, then a solution is achieved by using available algorithms. For more complex models, a combination of simplified models is required or the use of simulation or any heuristic models can be effective, as well.

Solution of the model
The outcome of well-defined optimization algorithms entails the model solution. It is crucial that the solution be accompanied with the sensitivity analysis, a term which refers to the behavior of the optimal solution if the parameters have to undergo any changes.

Validation of the model

Validation refers to the valid results which occur after the process of testing and improving a model in order for the model to be reliable used.

Implementation of the solution

This is the final step after the team has developed a certain model and reached an optimal solution. The model, the solution procedure and operating procedures for implementation will follow a system which is usually computer-based. The operating instructions extracted through this system are to be issued to the people who will put them into use under real conditions.

Chapter 2

Markov Chains

Operations Research (OR) is all about finding a way to make the right decisions. The decisions made within an environment of uncertainty concerning a future event follow probabilistic models for processes that evolve over time and which are called stochastic processes. One of the most known is the Markov Chains due to their special characteristics, which make them a great tool of OR. Markov Chains are capable of calculating probabilities of how a process will evolve in the future taking into consideration only the present state of the process, ignoring the past events.

2.1 Discrete time Markov Chains

A stochastic process describes the relation between random variables X$_t$ where the index t runs through a given set T. T is often a period of time where a system operating is being observed and X$_t$ represents the *state* of the system at time t. A discrete-time stochastic process is a Markov Chain if, for t = 0, 1, 2, ... and every sequence i, j, k$_0$, k$_1$, ..., k$_{t-1}$:

$$\begin{aligned} P\{X_{t+1} = j | X_0 = k_0, X_1 = k_1, ..., X_{t-1} = k_{t-1}, X_t = i\} \\ = P\{X_{t+1} = j | X_t = i\} \end{aligned}$$

The above mathematical representation is known as *Markovian property* and indicates that conditional probability of any future "event" is independent

of a given past "event" and depends only upon the present state $X_t = i$.
In this special case conditional probabilities, $P\{X_{t+1} = j|X_t = i\}$, are known as *transition probabilities*. If for each i and j,

$$P\{X_{t+1} = j|X_t = i\} = P\{X_1 = j|X_0 = i\}$$

for all $t = 1, 2, ...$, it is conclude that the probability law relating to the next period's state to the current state does not change over time. For this reason, they are called stationary transition probabilities.
To simplify notation with transition probabilities, let for each i, j, n :

$$P_{ij} = P\{X_{t+1} = j|X_t = i\}$$
$$P_{ij}^{(n)} = P\{X_{t+n} = j|X_t = i\}$$

The $P_{ij}^{(n)}$ represents the n-step transition probability which gives us the probability of the system's state j after n steps (units of time), given that it starts in state i at any time t. It is also important to underline some crucial properties as the $P_{ij}^{(n)}$ are conditional probabilities. These are:

- $P_{ij}^{(n)} \geq 0$ for all i, j, n = 0,1,2,...
- $\sum_{j=0}^{M} P_{ij}^{(n)} = 1$ for all i, j, n = 0,1,2,...
- $\sum_{j=1}^{j=s} P(X_{t+n} = j|P(X_t = i)) = 1$ for all i, j, n = 0,1,2,...

It is usually preferable to use matrix form so as to present all the n-step transition probabilities as they are shown bellow:

State	0	1	$\cdots$	M
0	$p_{00}^{(n)}$	$p_{01}^{(n)}$	$\cdots$	$p_{0M}^{(n)}$
1	$p_{10}^{(n)}$	$p_{11}^{(n)}$	$\cdots$	$p_{1M}^{(n)}$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
M	$p_{M0}^{(n)}$	$p_{M1}^{(n)}$	$\cdots$	$p_{MM}^{(n)}$

where the horizontal line refers to *present* state 'i' and the vertical line refers to the *future* state 'j'.

A great example of application of Markov chains is proposed by M. Grinstead and J. Snell in their book Introduction to Probability [2, ex.13, p.424]

Example 2.1.1. Suppose Smith is in jail and has 3 dollars. Given the bail out fee is 8 dollars, a guard agrees to make a series of bets with him. If Smith bets A dollars, he wins A dollars with probability 0.4 and loses A dollars with probability 0.6.[12, ex.13, p.9]

(a) The probability that he wins 8 dollars before losing all of his money if he bets 1 dollar each time (timid strategy) is:
The Markov chain $(X_n, n = 0, 1, ...)$ representing the evolution of Smith's money has diagram

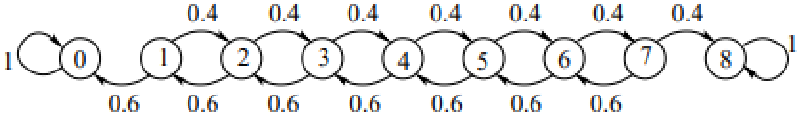

Figure 2.1: The transition diagram for Example 2.1.1

Let $\phi(i)$ be the probability that the chain reaches state 8 before reaching state 0, starting from state i. In other words, if S_j is the first $n \geq 0$ such that $X_n = j$ then

$$\phi(i) = P_i(S_8 < S_0) = P(S_8 < S_0 | X_0 = i)$$

Using Markov property at time $n = 1$:

$$\begin{aligned}
&\phi(i) = 0.4\phi(i+1) + 0.6\phi(i-1), \quad i = 1, 2, 3, 4, 5, 6, 7 \\
&\phi(0) = 0 \\
&\phi(8) = 1
\end{aligned}$$

By solving the system of the previous linear equations it is concluded that

$$\begin{aligned}
\phi &= (\phi(1), \phi(2), \phi(3), \phi(4), \phi(5), \phi(6), \phi(7)) \\
&= (0.0203, 0.0508, 0.0964, 0.1649, 0.2677, 0.4219, 0.6531, 1)
\end{aligned}$$

Thus, the probability of reaching state 8 before reaching state 0 starting from state 4 is only 16.49%.

(b) The probability that he wins 8 dollars before losing all of his money if he bets, each time, as much as possible but not more than necessary to bring his fortune up to 8 dollars (bold strategy) is:
The transition diagram takes a new form

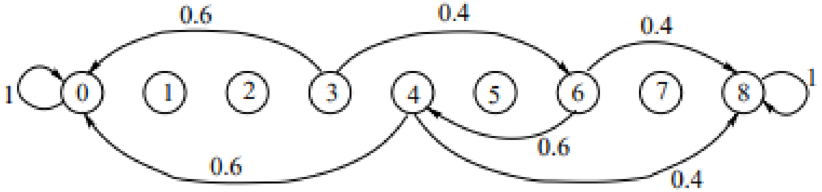

Figure 2.2: The transition diagram for question b

The new equation are:

$$\begin{aligned}
\phi(3) &= 0.4\phi(6)\\
\phi(6) &= 0.4\phi(8) + 0.4\phi(4)\\
\phi(4) &= 0.4\phi(8)\\
\phi(0) &= 0\\
\phi(8) &= 1
\end{aligned}$$

By solving these equations the results are

$$\phi(3) = 0.256,\ \phi(4) = 0.4,\ \phi(6) = 0.64,$$

(c) Which strategy gives Smith the better chance of getting out of jail?
By comparing the fourth components of the vector ϕ it is concluded that the bold strategy gives Smith a better chance to get out jail.

2.2 Chapman - Kolmogorov Equations

A great method for computing these n-step transition probabilities is by using the Chapman-Kolomogorov equations:

$$P_{ij}^{(n)} = \sum_{j=0}^{M} P_{ik}^{(m)} P_{kj}^{(n-m)}$$

for all $i = 0, 1, ..., M, j = 0, 1, ..., M$ and
any $m = 1, 2, ..., n-1, n = m+1, m+2, ...$
These equations rely on the fact that between i and j states there is a state k after exactly m states. Thus, the conditional probability $P_{ik}^{(m)} P_{kj}^{(n-m)}$ states that, given a starting point i, the process goes to state k after m steps and then to state j after $n-m$ steps. In this way, $P_{ij}^{(n)}$ can be computed by summing conditional probabilities over all possible k. The expressions for $m = 1$ and $m = n-1$ are presented below:

$$P_{ij}^{(n)} = \sum_{j=0}^{M} P_{ik} p_{kj}^{(n-1)}$$

and

$$P_{ij}^{(n)} = \sum_{j=0}^{M} P_{ik}^{(n-1)} P_{kj}$$

for all states i and j.
It is evident that ,by using Chapman-Kolomogorov equations, the n-step probabilities can be obtained from the one-step ones recursively.
For a homogeneous discrete-time Markov chain these equations have the following form:

$$\begin{aligned} P_{ij}^{(n)} &= \mathbb{P}\{X_n = j | X_0 = i\} \\ &= \sum_{\text{all k}} \mathbb{P}\{X_n = j, X_m = k | X_0 = i\} \quad \text{for 0<m<n} \\ &= \sum_{\text{all k}} \mathbb{P}\{X_n = j | X_m = k, X_0 = i\} \mathbb{P}\{X_m = k | X_0 = i\} \end{aligned}$$

By applying the Markov property,

$$\begin{aligned} P_{ij}^{(n)} &= \sum_{\text{all k}} \mathbb{P}\{X_n = j|X_m = k, X_0 = i\}\mathbb{P}\{X_m = k|X_0 = i\} \\ &= \sum_{\text{all k}} P_{kj}^{(n-m)} P_{ik}^{(m)} \quad \text{ for 0<m<n} \end{aligned}$$

In matrix notation the Chapman-Kolmogorov equations are written as

$$P^{(n)} = P^{(m)} P^{(n-m)}$$

where, by definition, $P^{(0)} = I$ the identity matrix.
It is also given that

$$P^{(n)} = PP^{(n-m)} = P^{(n-m)}P$$

Regarding the previous property , the matrix of n-step transition probabilities is calculated by multiplying the matrix of one-step transition by itself $(n-1)$ times. In other words, $P^{(m)} = P^m$.
For a nonhomogenous discrete time Markov chain, the matrices $P(n)$ may depend on the particular time step n. In this case the product P^2 will be equal to $P(n)P(n+1)$, P_3 will be equal to $P(n)P(n+1)P(n+2)$ and so on. As a result a new matrix arises:

$$P^{(n)}(m, m+1, \cdots, m+n-1) = P(m)P(m+1)\cdots P(m+n-1)$$

whose ij element refers to $\mathbb{P}\{X_{n+m} = j|X_m = i\}$.
The notation used for the probability a Markon chain begins in state i is $\pi_i^{(0)}$ and $\pi^{(0)}$ is the row vector whose i^{th} element is $\pi_i^{(0)}$. The probability of being in state j after the first step is given by:

$$\pi^{(1)} = \pi^{(0)}P(0)$$

For a homogeneous Markov chain

$$\pi^{(1)} = \pi^{(0)}P$$

The elements of the vector $\pi^{(1)}$ give the probability of being in the various states of the Markov chain after the first step.
The probability of being in state j after two time steps is given by:

$$\pi^{(2)} = \pi^{(1)}P(1) = \pi^{(0)}P(0)P(1)$$

And for a homogeneous Markov chain

$$\pi^{(2)} = \pi^{(1)} P = \pi^{(0)} P^2$$

In general, the probability distribution after n steps is given by

$$\pi^{(n)} = \pi^{(n-1)} P(n-1) = \pi^{(0)} P(0) \cdots P(n-1)$$

Accordingly, for a homogeneous Markov chain

$$\pi^{(n)} = \pi^{(n-1)} P = \pi^{(0)} P^n$$

Letting $n \to \infty$ then

$$\lim_{n\to\infty} \pi^{(n)} = \pi^{(0)} \lim_{n\to\infty} P^n$$

where the limit does not necessarily exists for all Markov chains, not even for all finite-state ones.

2.3 Classification of states of a Markov Chain

It is evident that the relation between transition probabilities and states plays an important role in understanding the Markov Chains. The states of a Markov Chain can be classified based on the transition probability P_{ij} of $\mathbf{P}$ [25, p.633].

1. A state j is **absorbing** if it is certain to return to itself in one transition. The transition probability of an absorbing state is $P_{jj} = 1$.
2. A state j is **transient** if it can reach another state but cannot be reached back from another state. The transition probability in this case will follow the condition $\lim_{n\to\infty} P_{ij}^{(n)} = 0$, for all i.
3. A state j is **recurrent** if the probability of being revisited from other states is 1. This is possible only if the state is not transient.
4. A state j is **periodic** with period $t > 1$ if a return is possible only in t,2t,3t,...steps. The transition probability of a periodic state is $P_{jj}^{(n)} = 0$ when n is not divisible by t.

For a better comprehension regarding the classification of states the following illustration is being explained.

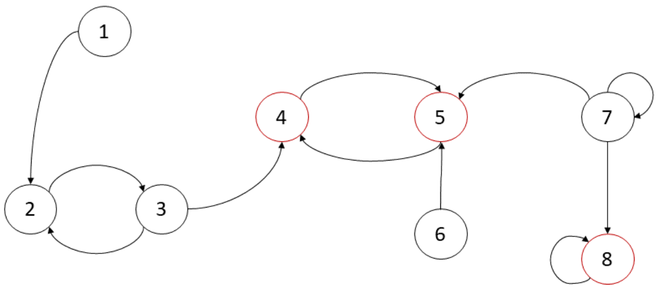

Figure 2.3: Transient & Recurrent states

States 1 and 6 are defined as transient. It is also observed that the Markov chain can be in state 1 or 6 only for the first time step thus they are also characterized as **ephemeral** states according to William j. Stewart [22, p.207]. States 2 and 3 are also transient states as the Markov chain can enter from state 2 and move to one from the other for a number of time steps but it will eventually exit from state 3 to 4.
States 4 and 5 are recurrent states . Once the Markov chain reaches state 4 then all subsequent transitions will take it from one to the other. It is also worth mentioning that each state 4 or 5 is reached after two time steps thus these states are defined also as **periodic** with period 2. Positive recurrent states are the states whose mean recurrence time is finite so states 4 and 5 are characterized **positive recurrent** as well. Recurrent states with infinite recurrent time are known as *null recurrent states* and this is feasible when the Markov chain has a infinite number of states. Furthermore, a state whose period p=1 is defined as **aperiodic**. A state that is positive recurrent and aperiodic is said to be **ergodic**.

State 7 is a transient state. Once the Markov chain enters state 7 may, for some finite number of time steps, remain in state 7, but eventually it will move on to either state 5 or state 8.
Finally state 8 is a recurrent state but also it is defined as **absorbing**. It is evident that if a Markov chain enters state 8 it will remain there forever and thus $P_{ii} = 1$. If $P_{ii} < 1$ then the state will be defined either as recurrent or as transient.

It is also required to pinpoint the need for **a new quantity** f_{jj} which is going to give the probability of the first *return* to state j after leaving it before exactly *n steps* [22, p.207].This quantity is defined as:

$$f_{jj}^{(n)} = \mathbb{P}[X_n = j, X_{n-1} \neq j, ..., X_1 \neq j | X_0 = j\}]$$

for $n = 1, 2, ..$
The probability $P_{jj}^{(n)}$ is *not the same* with the probability $f_{jj}^{(n)}$, as the first one calculates the probability of returning to state j, without taking into consideration if the state j was visited again at one or more intermediate steps.
Based on the definition of $f_{jj}^{(n)}$ it is evident that $f_{jj}^{(1)} = P_{jj}^{(1)} = P_{jj}$ as the probability that the first return to state j demands one step as the single step transition probability. The relation between these two probabilities is given by the following form, using $P_{jj}^{(0)} = 1$:

$$P_{jj}^{(n)} = \sum_{l=1}^{n} f_{jj}^{(l)} P_{jj}^{(n-l)}$$

for $n \geq 1$.
or respectively

$$f_{jj}^{(n)} = P_{jj}^{(n)} - \sum_{l=1}^{n-1} f_{jj}^{(l)} P_{jj}^{(n-l)}$$

for $n \geq 1$.
For example,

$$P_{jj}^{(1)} = f_{jj}^{(1)} P_{jj}^{(0)} \text{ for } n = 1,$$
$$P_{jj}^{(3)} = f_{jj}^{(1)} P_{jj}^{(2)} + f_{jj}^{(2)} P_{jj}^{(1)} + f_{jj}^{(3)} P_{jj}^{(0)} \text{ for } n = 3.$$

The second example gives the probability a process is in state j three steps after leaving it. The first term gives the probability of the first return to state j after just one step and the probability it returns to state j by the end of two steps remaining. The second term gives the probability the process returns to state j after exactly two steps for the first time and in the third step remains to state j and the third term gives the probability the process returns to state j after exactly three steps for the first time.
The probability of returning to state j is given by:

$$f_{jj} = \sum_{n=1}^{\infty} f_{jj}^{(n)}$$

- If $f_{jj} = 1$, the state is denoted as *recurrent* as it will inevitably return to this state in the future. The expected returns to state j of a Markov chain is equal to $\sum_{n=0}^{\infty} P_{jj}^{(n)} = \infty$.
  Let $I_n = 1$ if the Markov chain is in state j at step n and $I_n = 0$ otherwise. Then the total time steps that state j is occupied is $\sum_{n=0}^{\infty} I_n$. Given the Markov chain starts in state j, the expected number of visits the Markov chain makes to state j is [22, p.209]:

  $$\begin{aligned} E[\sum_{n=0}^{\infty} I_n | X_0 = j] &= \sum_{n=0}^{\infty} E[I_n | X_0 = j] \\ &= \sum_{n=0}^{\infty} Prob\{X_n = j | X_0 = j\} \\ &= \sum_{n=0}^{\infty} P_{jj}^{(n)} \end{aligned}$$

  Thus, when state j is recurrent,

  $$\sum_{n=0}^{(n)} P_{jj}^{(n)} = \infty$$

- If $f_{jj} < 1$, the state is denoted as *transient*. The probability of never returning to this state is nonzero so the expected number of returns to state j is finite. The probability that the Markov chain is in state j and it will never return is $1 - f_{jj}$ and it can be described with a sequence of Bernoulli trials, where "success" is considered as the never return of the Markov chain to state j. The probability the Markov chain returns to state j, $n-1$ times after leaving it and never return ($n \geq 1$), is equal to $(1-f_{jj})f_{jj}^{n-1}$, as the geometric probability mass function. Thus, the mean number of returns to state j is equal to $1/(1-f_{jj})$,which is finite and as a result [22, p.209]:

$$\sum_{n=0}^{\infty} Pjj^{(n)} < \infty$$

When state j is recurrent, the *mean recurrence time* M_{jj} of state j, in other words the mean time steps it takes to return to state j for the first time after leaving it, is defined as

$$M_{jj} = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$$

If M_{jj} is finite, is called *positive recurrent state* and if $M_{jj} = \infty$ is called *null recurrent state.*

Theorem. *In a finite Markov chain*

- *No state is null recurrent*
- *At least one state must be positive recurrent, in other words not all states can be transient.*

Assuming a Markov chain has all its states transient then it would spend an finite amount of time in each of its states. This assumption though is impossible as after a finite period of time it will have nowhere to go.

2.3.1 Stationary Distribution

Let X denote a Markov chain with state space E and π a measure on E. If $\mathbb{P}(X_n = i) = \mathbb{P}(X_0 = i) = \pi_i$ for all $n \in \mathbb{N}$ and $i \in \mathbb{E}$, then X^π is called stationary and π is called a stationary measure for X. In case π is a probability measure then π is the stationary distribution for X [13, p.21].

Theorem. *Let X denote a Markov chain with state space E and transition matrix P. Furthermore, let π denotes a probability distribution on E with $\pi P = \pi$, i.e.*

$$\pi_i = \sum_{j \in E} \pi_j P_{ji} \quad \textit{and} \quad \sum_{j \in E} \pi_j = 1$$

for all $i \in E$. Then π is a stationary distribution for X. If π is a stationary distribution for X, then $\pi P = \pi$ holds.

Example 2.3.1. A four-state Markov chain has the transition matrix :

$$T = \begin{bmatrix} 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Show that all states have period 3.
Regarding the transformation of a transition matrix to a transition diagram, it is important to mention that rows in a transition matrix represent inputs and columns represent outputs. So the transition diagram depicts the probabilities transitions are made between states. For example, probability a transition occurs from state E_1 to E_1 is 0, from state E_1 to E_2 is 1/3, from state E_1 t state E_3 is 0 and from state E_1 to E_4 is 1/3 . In this case, if the chain starts in E_1, then returns to state E_1 are only possible at steps $3, 6, 9, \cdots$ either through E_2 or E_3.

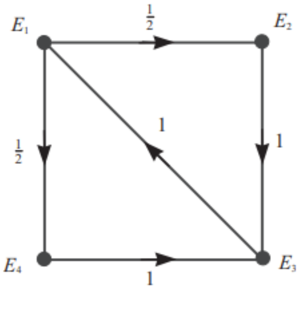

Figure 2.4: The transition diagram for Example 2.3.1

A suspected periodicity can be checked by direct computation as follows:

$$S = T^3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

In this example,

$$S^2 = T^6 = SS = S$$

so that

$$S^r = T^{3r} = S, \quad (\mathrm{r} = 1,\, 2,\, ...)$$

which always has non zero diagonal elements. On the other hand,

$$S^{r+1} = S^r S = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad S^{r+2} = S^r S^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Both of these matrices have zero diagonal elements for r=1,2,3,... . Therefore, for i = 1, 2, 3, 4, ...

$$P_{ii}^{(n)} = 0, \text{ for n} \neq 3,\, 6,\, 9,\, ...$$
$$P_{ii}^{(n)} \neq 0, \text{ for n} = 3,\, 6,\, 9,\, ...$$

which means that all states are period 3.

Example 2.3.2. A three-state Markov chain has the transition matrix

$$T = \begin{bmatrix} p & 1-p & 0 \\ 0 & 0 & 1 \\ 1-q & 0 & q \end{bmatrix}$$

where $0 < p < 1$, $0 < q < 1$. Show that the state E_1 is recurrent.[8, p.86] The transition diagram of this example is illustrated below:

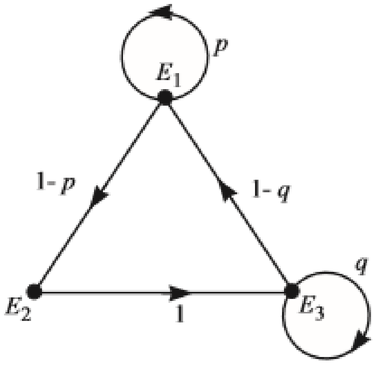

Figure 2.5: The transition diagram for Example 2.3.2

If a sequence begins at state E_1 it can return to state E_1 at every step except for n=2 since after two steps the chain must be in state E_3. From the figure it can be argued that

$$f_1^{(1)} = p, \quad f_1^{(2)} = 0, \quad f_1^{(3)} = (1-p)\cdot 1 \cdot (1-q),$$
$$f_1^{(n)} = (1-p)\cdot 1 \cdot q^{n-3} \cdot (1-q), \quad (n \geq 3)$$

The last result is derived from the following sequence of transitions:

$$E_1 \quad E_2 \quad \overbrace{E_3 \quad E_3 \quad E_3 \cdots E_3}^{\text{(n-3) times}} \quad E_1.$$

The probability f_1 that the state E_1 is reached at least once after the sequence begins is:

$$\begin{aligned} f_1 = \sum_{n=1}^{\infty} f_1^n = p + \sum_{n=3}^{\infty}(1-p)(1-q)q^{n-3} \\ = p + (1-p)(1-q)\sum_{s=0}^{\infty} q^s, \quad \text{(s=n-3)} \\ = p + (1-p)\frac{(1-q)}{(1-q)} = 1 \end{aligned}$$

Therefore, the $f_1 = 1$ and the state E_1 is characterized as recurrent. The mean recurrence time is given by

$$\mu_1 = \sum_{n=1}^{\infty} n f_1^{(n)} = p + (1-p)(1-q)\sum_{n=3}^{\infty} n q^{n-3}$$
$$= p + (1-p)(1-q)\left[\frac{3-2q}{(1-q)^2}\right] = \frac{3-2p-2q+pq}{1-q}$$

which is finite, so the E_1 state is called positive recurrent state.

Example 2.3.3. A three-state inhomogenous Markov chain is described by the transition matrix presented below:

$$T_n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 1/(n+1) & 0 & n/(n+1) \end{bmatrix}$$

where T_n is the transition matrix at step n. Show that E_1 is a null recurrent state [8, p.87].
From the figure 2.6 it can be argued that

$$f_1^{(1)} = \frac{1}{2}, \quad f_1^{(2)} = 0, \quad f_1^{(3)} = \frac{1}{2}\cdot 1 \cdot \frac{1}{4},$$
$$f_1^{(n)} = \frac{1}{2}\cdot 1\cdot \frac{3}{4}\cdot\frac{4}{5}\cdots\frac{n-1}{n}\cdot\frac{1}{n+1} = \frac{3}{2n(n+1)}, \quad (n \geq 4)$$

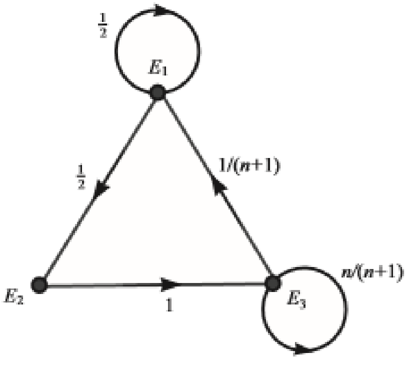

Figure 2.6: The transition diagram for Example 2.3.3

Thus,

$$f_1 = \frac{1}{2} + \frac{1}{8} + \frac{3}{2}\sum_{n=4}^{\infty}\frac{1}{n(n+1)}$$

But since,

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$$

then

$$\sum_{n=4}^{\infty}\frac{1}{n(n+1)} = \lim_{N\to\infty}\sum_{n=4}^{N}\left(\frac{1}{n} - \frac{1}{n+1}\right) = \lim_{N\to\infty}\left(\frac{1}{4} - \frac{1}{N+1}\right) = \frac{1}{4}$$

So it is concluded that

$$f_1 = \frac{5}{8} + \frac{3}{8} = 1$$

which means that the state E_1 is recurrent.
Furthermore, the mean recurrence time

$$\begin{aligned}\mu_1 = \sum_{n=1}^{\infty} n f_1^{(n)} &= \frac{7}{8} + \frac{3}{2}\sum_{n=4}^{\infty}\frac{n}{n(n+1)} \\ &= \frac{7}{8} + \frac{3}{2}\left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \cdots\right) \\ &= \frac{7}{8} + \frac{3}{2}\sum_{n=5}^{\infty}\frac{1}{n}\end{aligned}$$

The last series in the previous equation are known as Harmonic series which are divergent. As a result the mean recurrence time is $\mu_1 = \infty$ and the state E_1 is characterized as null recurrent.

Example 2.3.4. A four-state Markov chain is described by the transition matrix

$$T = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Show that E_1 is a transient state [8, p.89].
The transition diagram of Example 2.3.4 is illustrated below:

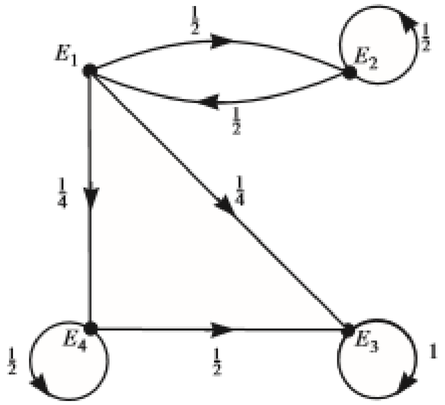

Figure 2.7: The transition diagram for Example 2.3.4

From the figure it can be argued that

$$f_1^{(1)} = 0, \quad f_1^{(2)} = \frac{1}{2} \cdot \frac{1}{2} = \big(\frac{1}{2}\big)^2, \quad f_1^{(3)} = \big(\frac{1}{2}\big)^3, \quad f_1^{(n)} = \big(\frac{1}{2}\big)^n$$

Thus,

$$f_1 = \sum_{n=1}^{\infty} f_1^{(n)} = \sum_{n=2}^{\infty} \big(\frac{1}{2}\big)^n = \frac{1}{2} < 1$$

Regarding the previous results the state state E_1 is characterized as transient. The transience of E_1 is also evdent from the figure as transitions from states E_3 or E_4 to states E_1 or E_2 are not feasible.

2.4 Renewal Processes

A renewal process is a special case of counting process. A counting process, $\{N(t), t \geq 0\}$, as Stewart W. defines in his book, is a stochastic process which counts the number of events that occur up to (and including) time t. Thus N(t) is expected to be integer valued with the properties that $N(t) \geq 0$ and

$N(t_1) \leq N(t_2)$ if $t_1 \leq t_2$ [22, p.267].

Renewal process
Let X_n, $n \geq 1$ be non negative random variables that represent the time between successive events. If a sequence consists of X_n which are independent and identically distributed then the counting process $\{N(t), t \geq 0\}$ is defined as renewal. Sometimes recurrent process and renewal process are considered as identical.
It is also important to underline that although the option $X_n = 0$ is feasible, events which occur simultaneously will not be examined.

Example 2.4.1. Suppose there is an infinite supply of lightbulbs whose lifetimes are independent and identically distributed. Given that only one lightbulb is used at a time when one fails then it is replaced by a new one immediately. Under these conditions, $\{N(t),\, t \geq 0\}$ is a renewal process when $N(t)$ represents the number of lightbulbs that have failed by time t. [20, p.417]

Given X_1,X_2,X_3,... are interarrival times, let

$$S_0 = 0,\ S_n = \sum_{i=1}^{n} X_i,\ n \geq 1$$

The illustration presented below, suggested by Sheldon M. Ross [20, p.418], depicts renewal and interarrival times

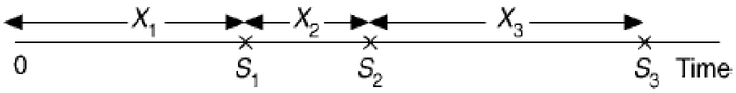

Figure 2.8: Renewal & Interarrival times

The time of the first renewal is $S_1 = X_1$. The time of the second renewal is $S_2 = X_1 + X_2$, in other words, the time until the first renewal plus the time between the first and the second renewal. In general S_n denotes the time of the n-nth renewal.
Given that F denotes the interarrival distribution, it is assumed that $F(0) =$

$\mathbb{P}\{X_n = 0\} < 1$. Furthermore, the mean time between successive renewals is given by

$$\mu = E[X_n], \quad n \geq 1$$

It also important to pinpoint that an infinite number of renewals can not occur in a finite amount of time. Given that S_n denotes the time of the n-nth renewal then

$$N(t) = \max\{n : S_n \leq t\}$$

Supposing that $S_4 \leq t$ and $S_5 > t$, it concluded that the fourth renewal had occurred by time t but the fifth renewal occurred after time t. By using the strong law of large numbers it follows that:

$$\mathbb{P}\left[\lim_{N\to\infty} \frac{f(X_1) + f(X_2) + \cdots + f(X_N)}{N} = \mathbb{E}^{\pi}[f]\right] = 1$$

The non negativity of X_n and the fact that X_n is not identically 0 follows that $\mu > 0$ and as result S_n must be going to infinity as n goes to infinity. Therefore, S_n can be less than or equal to t for a finite number of values of n and as a result $N(t)$ must be finite.

On the other hand, even though $N(t) < \infty$ for each t, it is a fact that, with probability 1,

$$N(\infty) \equiv \lim_{t\to\infty} N(t) = \infty$$

In other words, the only way to achieve $N(\infty)$ is for one of the interarrival times to be infinite.

Consequently,

$$\begin{aligned} P\{N(\infty) < \infty\} &= P\{X_n = \infty \quad \text{for some n}\} \\ &= P\{\bigcup_{n=1}^{\infty}\{X_n = \infty\}\} \\ &\leq \sum_{n=1}^{\infty} P\{X_n = \infty\} \\ &= 0 \end{aligned}$$

2.5 Continuous time Markov chains

In a discrete time Markov chain, there is an infinite sequence of time steps where a change of state may occur or the chain may remain in its current state. There are situations though, that a continuous time parameter is required so as to observe possible changes of state at any point of time.

A stochastic process {X(t), $t \geq 0$} is a *continuous time Markov chain*, if for all states n (n$\in$ $\mathbb{Z}$) and for any sequence $t_0, t_1, t_2, .., t_n, t_{n+1}$ where $t_0 < t_1 < ... < t_{n+1}$:

$$\begin{aligned}\mathbb{P}\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, ..., X(t_0) = x_0\} \\ = \mathbb{P}\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\}\end{aligned}$$

It is important to mention that not only does this definition not affect the future evolution of the chain but also does not take into consideration the time spent in the current state.
An alternative definition could be according to W.J. Stewart [22, p.253]:
The stochastic process {X(t), $t \geq 0$} is a continuous Markon chain if for states i, j, k and for all instants s, t, u with $t, s \geq 0$ and $0 \leq u \leq s$:

$$\mathbb{P}\{X(s+t) = k | X(s) = j, X(u) = i\} = \mathbb{P}\{X(s+t) = k | X(s) = j\}$$

where i represents the state at time u (past time)
j represents the state at time s (current time)
k represents the state at time s+t (future time)
A continuous time Markov chain is called *nonhomogeneous* when, given $t \geq s$:

$$P_{ij}(s,t) = \mathbb{P}\{X(t) = j | X(s) = i\}$$

On the other hand, a continuous time Markov chain is *homogeneous* when:

$$P_{ij}(\tau) = \mathbb{P}\{X(s+\tau) = j | X(s) = i\} \text{ for all } s \geq 0$$

It is important to underline that in a homogeneous Markov chain, the difference $\tau = t - s$ affects the transition probabilities, not the values s, t.
A transition probability is defined as *stationary transition probability* if the probability is independent of s so that:

$$P_{ij}(t) = P\{X(t) = j | X(0) = i\}$$

and as a result:

$$\lim_{t \to\ 0} P_{ij}(t) = \begin{cases} 1 & \text{if i=j} \\ 0 & \text{if } i \neq j \end{cases}$$

It is assumed that the continuous time Markov chains are referred to have a finite number of states and the transition probabilities are considered as stationary.

Let the random variable T_i denote the amount of time a process spends in state i before moving to another, where i = 0, 1, 2, ..., M. Supposing the process enters a state i at time s, then according to Markovian property:

$$P\{T_i > t + s | T_i > s\} = P\{T_i > t\}$$

In other words the remaining time until the process leaves from the current state is the same as the process is not affected by the time already spent, due to its memoryless property. The only continuous probability distribution that follows this property is the exponential, whose parameter will be denoted as q. So a continuous time Markov chain can have a new form ,where the random variable T_i follows an exponential distribution with a mean $1/q_i$.

In a continuous time Markov chain the interactions between the states are not described by transition probabilities as in a discrete time Markov chain, but in terms of the rates at which transitions occur [22, p.254]. The transition of a process from state i to state j at time t is described from the rate $q_{ij}(t)$ per unit time. The transition rates are defined as:

$$q_i = -\frac{d}{dt} P_{ii}(0) = \lim_{t \to\ 0} \frac{1 - P_{ii}(t)}{t} \qquad \text{for i = 0, 1, 2, ..., M}$$

and

$$q_{ij} = -\frac{d}{dt} P_{ij}(0) = \lim_{t \to\ 0} \frac{P_{ij}(t)}{t} = q_i P_{ij} \qquad \text{for all } i \neq j$$

where q_i is the transition rate out of state i and refers to the expected time that the process spends in state i per visit to state i ($q_i = 1/E[T_i]$). The transition rate q_{ij} refers to the expected number of times there is a transition from state i to state j per unit of time spent in state i. It is evident that $q_i = \sum_{j \neq i} q_{ij}$. The transition rate equivalent to the system remaining in place is defined by

$$q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t)$$

When state i is an absorbing state, $q_{ii}(t) = 0$. The reason why $q_{ii}(t)$ is negative is that this quantity denotes a transition rate and as such is defined as a derivative. "Given that the system is in state i at time t, the probability that it will transfer to a different state j increases with time, whereas the probability that it remains in state i must decrease with time. It is appropriate in the first case that the derivative at time t be positive, and in the second that it be negative" [22, p.255].
The matrix $Q(t)$ is called the infinitesimal generator or transition-rate matrix and its ij^{th} element is $q_{ij}(t)$. The matrix form is calculated by

$$Q(t) = \lim_{\Delta t \to 0} \{ \frac{P(t, t + \Delta t) - I}{\Delta t} \}$$

where P(t,t+Δt) is the transition probability matrix, its ij^{th} element is $q_{ij}(t, t + \Delta t)$, and I is the identity matrix. The sum of all elements in any row of Q(t) must be zero. Furthermore, given a homogeneous continuous-time Markov chain , the transition rates q_{ij} are independent of time and the matrix of transition rates is denoted as Q.

Example 2.5.1. Cars arrive at a service center at an average rate of five per hour and it takes on average ten minutes to service each car. To represent this situation as a homogeneous continuous-time Markov chain, the state space of the model must be specified. It will be assumed that the non-negative integers 0, 1, 2, 3,... will represent the situation in which there are 0, 1, 2, 3,... cars in the center. Additionally, it will be assumed that no more than one car can arrive at any moment, no more than one car can exit from service at any moment and that cars do not arrive or depart simultaneously. Taking into consideration that transitions can only be made among nearby states, an illustration of possible transitions is depicted below:

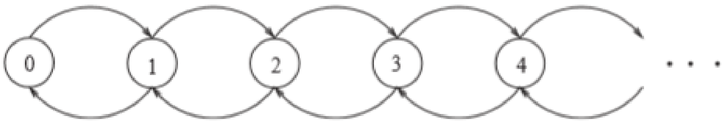

Figure 2.9: Transitions diagram

When a car arrives then a transition is made from state i to the next highest neighbor i+1, given that the mean arrival time is 1/5 hours. In terms of satisfying the Markov property (exponentially distributed interarrival time with mean 1/5), the rate of transition from any state i to state i+1 is $q_{i,i+1}$ =5/hour for $i \geq 0$. Additionally, based on the Markov property again, the mean service time will be equal to 10 minutes (exponential distribution). Thus the rate of transition from state i+1 to state i is $q_{i+1,i}$ =6/hour for $i \geq 0$. The transition rate matrix Q must be tridiagonal with superdiagonal elements all equal to 5 and subdiagonal elements all equal to 6. Its form is

$$Q = \begin{bmatrix} -5 & 5 & 0 & 0 & 0 & \cdots \\ 6 & -11 & 5 & 0 & 0 & \cdots \\ 0 & 6 & -11 & 5 & 0 & \cdots \\ 0 & 0 & 6 & -11 & 5 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

Chapter 3

Queueing Theory

Waiting for service is part of the everyday life. Not only are queues being observed among human services but also in machine, vehicle, or even airport services. The aim of queueing theory is finding a balance between the cost of offering a service and the cost of waiting experienced by customers. Daily queueing systems have led to the creation and evolution of various queueing models so as to enhance their operation. Too much service capacity as well as excessive waiting lines can prove both negative as far as costs are concerned. The goal is to make as less as possible the effects of the cost of service and waiting.

3.1 Structure of a Queueing model

In order to create a queueing model the definition of several terms must be specified. It is important to understand how a queueing system works. For the characterization of a queueing situation, it is necessary to define **customers** and **servers**. Customers arriving at a queueing system are generated by an **input source** and the size of this source defines the total number of customers that will require potential service. The **interarrival time** is the time between successive arrivals of customers and the offered service is measured by the **service time** per customer. Generally, the interarrival and service times are probabilistic or deterministic.

Once the customer enters the system, the service itself will either start right away (no queue) or the customer will have to wait in **queue**. A queue can be finite or infinite regarding the maximum permissible number of customers that it can contain. Usually the selection of a customer to be served from a queue follows a **queue discipline**, which defines the selection order of customers from a queue. The most common disciplines are *first-in, first-out* (FIFO), *last-in, first-out* (LIFO) and *service in random order* (SIrO). Customers may also be selected from the queue based on some order of priority [23]. Then the required service is performed for the selected customer by the service mechanism, after which the customer leaves the queueing system.

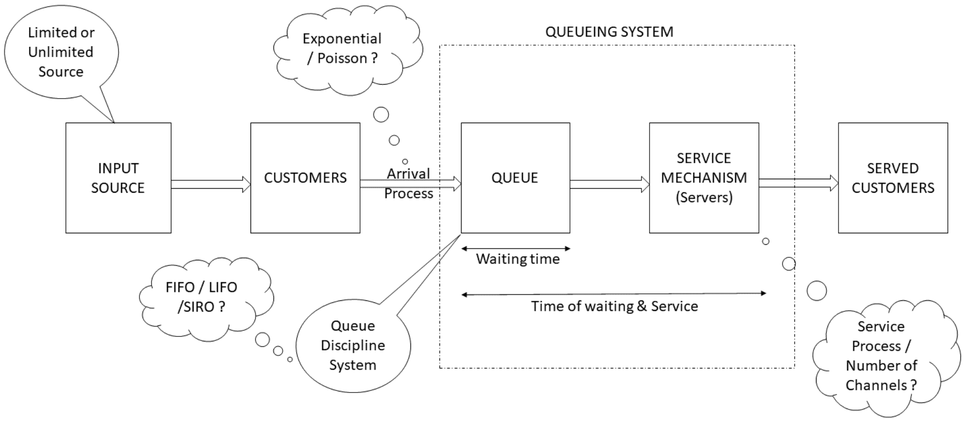

Figure 3.1: The basic Queueing Process

There are several assumptions that need to be taken into consideration for a better understanding of the Queueing theory, which are listed below:

- □ If the server is free, an arriving customer goes immediately into service without waiting in the queue.
- □ If the server is busy, then the arriving customer joins a queue and stays there until entering service.

- □ The time, between a customer leaves the service and eventually the system and a new customer entering a service, is considered as zero. The customers are ordered according to a queue discipline and they are distributed to the servers immediately.
- □ The system does not take into consideration human parameters, such as impatience. Customers remain in the system until they receive service.

A cost based queuing decision model is also presented below according to Hamdy A. Taha [25, p.654]:

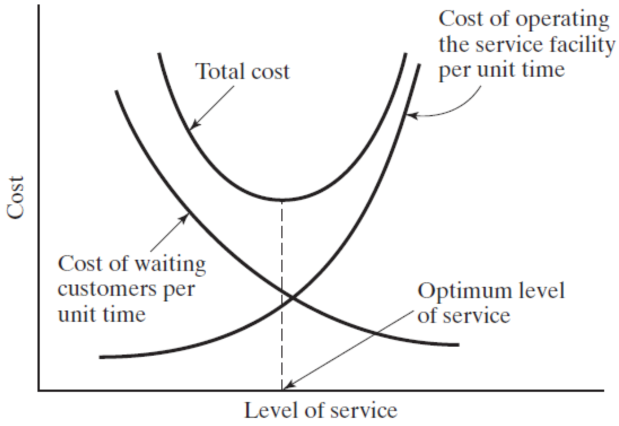

Figure 3.2: Cost-Based Queueing decision model

3.2 Terminology & Notation

The structure of queueing models uses a standard terminology and notation which is presented below:

State of system = n
= number of customers in queueing system (both in queues and servers)
Queue length =
number of customers in waiting lines

N(t) =
number of customers in queueing system at time t ($t \geq 0$)
$P_n(t)$ =
probability n customers are in queueing system at time t, given number at time 0
c =
number of servers (service channels) in queueing system
λ_n =
mean arrival rate (arrivals per unit time) of new customers entering the system, given n customers are already in it
μ_n =
mean service rate (served customers per unit time) given n customers are in system
λ =
λ_n, when the mean arrival rate is constant for all n
μ=
$\mu_n = c\mu$ (when $n \geq c$), when the mean rate service per busy server is constant for all $n \geq 1$
α_n = time of arrival of the n^{th} customer
d_n = time of departure of the n^{th} customer
$A_n = \alpha_n - \alpha_{n-1}$ = time between two successive arrivals of the $n - 1^{th}$ and the n^{th} customer
μ_A = mean time of arrivals (random variable A) = $\frac{1}{\lambda}$
σ_A = standard deviation of random variable A (arrivals)
$CV_A = \frac{\sigma_A}{\mu_A}$ = coefficient of variation of A
S_n = time of service required for the n_{th} customer
μ_s = mean time of required service (S random variable) = $\frac{1}{\mu}$
σ_S = standard deviation of random variable S (service time)
$CV_S = \frac{\sigma_S}{\mu_S}$ = coefficient of variation of S
WTM = waiting time multiple = $\frac{W_q}{\mu_S}$
$\rho = \lambda/(c\mu)$
is the **load factor**, which gives the system's service capacity (cμ) that is being utilized on the average by arriving customers (λ). [23]

Furthermore, in order to describe state results the following notations are used:

P_n=
probability of exactly n customers in the system
L=
expected number of customers in queueing system
L_n=
expected queue length excluding the customers being served
W =
waiting time in system, both in queue and in service, for each customer
W_q =
waiting time in queue (ignoring service time) for each customer

3.3 Arrival & Service Processes

In a queueing system, the queue length depends on the arrival and service processes. If the rate of arriving customers is greater than the channels offered for service, the system will eventually break down, as unbounded queues will form. On the other hand, if there is a low arrival rate then the number of channels will be reduced in order to avoid further costs, thus queues will form again. In order to regulate queues in a manner that there won't be a problem neither for customers nor for systems, probability distributions are being used in order to predict average waiting times, average queue length or even expected service times. It is important to identify:

- the arrival rate of customers to the system
- the service pattern of customers
- the way customers are being selected and distributed to service channels

in order to make valuable predictions. A queue is considered "full" when all channels are occupied and the customers arriving in this situation are identified as "lost".

The arrival process
The customer arrival process can be defined either as the number of arrivals

per unit time, known as *arrival rate*, or as the time between successive arrivals, known as the *interarrival time*. Given that λ is the mean arrival rate, the mean time between arrivals will be defined as $1/\lambda$. In case the arrival process is stochastic, the probability distribution of the interarrival time is given by:

$$A(t) = \mathbb{P}[\text{ time between arrivals} \leq t]$$

and

$$\frac{1}{\lambda} = \int_0^\infty t dA(t)$$

where dA(t) is the probability that the interval time is between t and t+dt and assuming that interval times are independent and identically distributed. An arrival process is denoted as *homogeneous* when A(t) does not change over time.

The service process

The service process is defined by the number of customers served per unit time from the available channels of the system or by the time required to serve a customer. Given that μ is the mean service rate, the *mean service time* is denoted as $1/\mu$. If B(x) denotes the probability of the service time required then:

$$B(x) = \mathbb{P}[\text{ Service time} \leq x]$$

and

$$\frac{1}{\mu} = \int_0^\infty x dB(x)$$

where dB(x) is the probability that the service time is between x and x+dx. It is important to pinpoint that the service time does not include the time spent in waiting lines. Furthermore, the service rates are calculated based on the time spent for a customer to be served, given that the channel is not idle and the channel not empty. The channels may be *batch* or *single*. A batch channel could be a train station, as multiple citizens are being served when the train arrives. The service rate may also be affected from everyday life factors such as:

- The number of customers in the system. There are cases where a server will slow down when a queue starts to empty or speed up the process when the arriving rate is increasing.

- The time when customers arrive to the system in order to receive service. For example, the service channels of a supermarket (cashiers) may start slowly in the morning and speed up gradually during the day due to the increasing number of customers especially during noontime.

The total number of servers or channels is denoted by c. When c>1 then there are two possibilities:

1. Each server has its own queue but that is not always fixed. There are several situations such as supermarkets where, when a line is more empty than another, the customers change their waiting lines in order to leave the system as soon as possible. In this case, a model computing waiting time for a single queue could be more accurate.
2. There are many examples in the everyday life where a single queue will be formed for distribution to multiple queues. A classic example could be the waiting line in a bank or at a mall.

3.4 Forms of Disciplines

In chapter 3.1, Queue discipline is mentioned as one of the crucial structural parts of a Queueing model. The way customers are allocated in a queue, in order to have a desired service, plays an important role in the accurate operation of a queueing system. Scheduling disciplines are categorized based on the preemptive or nonpreemptive policies. Preemptive policies refer to the priority a customer has. If a high priority customer enters a system then the service of a low priority customer is postponed and priority is given to the first one. The preempted customer is reinserted into the queue and when he returns to a service channel, he either returns to the previous service point (preempt - resume) or his service has to begin all over again (preempt - restart).

On the other hand, in a nonpreemptive system, service channels are committed to serve the selected customer, regardless his priority type and a possible arrival of a high priority customer during his service. The most common scheduling disciplines are:

- FIFO (first-in, first-out) or FCFS (first come, first served)
  The customer that has waited the longest in the queue is the next one to be served. (nonpreemptive system)

- LIFO (last-in, first-out) or LCFS (last come, first served)
  The last customer to arrive in the queue is the next one to be served. (nonpreemptive system)

- SIRO (service in random order) or ROS (random order of service)
  The next customer to be served is selected from the queue in a random order. (nonpreemptive system)

- RR (round robin)
  The service channels of this category provide a fixed service time, known as time slice. If a customer completes his service within the fixed duration then he immediately leaves the system. If a customer does not manage to complete his service at any point during the time slice then he is reinserted back in the queue as many times as the service needs to be completed. (nonpreemptive system)

- GD (General discipline)
  The customers waiting to receive service are allocated to service channels according to general discipline. (nonpreemptive system)

- TQ (truncated queues)
  The queue of this category can include a fixed number of customers waiting for service.(nonpreemptive system)

- SPTF (shortest processing time first)
  The customers that require minimum service time are given priority.(preemptive system)

- PRI (priority scheduling)
  The first customer to be selected from the queue is the one with the highest priority service. In case there are multiple high priority customers then the queue follows FCFS discipline. (preemptive system)

3.5 Kendall's Notation

In order to define a queueing system, Kendall's notation is used. D. G. Kendall in 1953 devised the first three elements of the notation (a/b/c), in 1966 A. M. Lee added the symbols d and e and in 1968 Hamdi A. Taha added the last element f. This special notation sums up the main characteristics of a queueing system such as interarrival and service time distribution, number of servers and queue discipline. Kendall's notation format is:

$$(a/b/c) : (d/e/f)$$

where:
a = arrivals distribution
b = service time distribution
c = number of channels available for service
d = queue discipline
e = maximum number of customers allowed in the system
f = size of the calling population where customers come from

The arrivals and service time distributions (symbols a, b) may be represented by one of the following notations:
M = Markovian or Poisson distributions regarding arrivals or departures (in other words exponential interarrival or service time distribution)
D = constant (deterministic) time
E_k = Erlang or Gamma distribution of time (in other words the sum of independent exponential distributions)
H_k = Hyperexponential distribution of time
GI = General distribution of interarrival time
G = General distribution of service time

Symbol d refers to the queue discipline in a queueing system and it can be represented by one of the eight notations of scheduled disciplines which were mentioned in 3.4.

For example, the model (M/E_k/2) : (FIFO/100/∞) refers to a queueing system whose arrivals follow a Poisson distribution or exponential interarrival time, service times follow an Erlang distribution and the availability of parallel service channels is 2. Furthermore, the queue discipline is FIFO, the maximum capacity of the system is 100 customers and the size of the calling population is infinite.

3.6 Performance measures at steady state

In order to analyze the performance of a queueing system, the definition of the *measures of effectivness* is important. These measures of performance are:
L = expected number of customers in system
L_q = exxpected number of customers in queue
W = expected waiting time in system, known as *response* or *sojourn time*
W_q = expected waiting time in queue
$\overline{c}$ = expected number of busy servers

Let N (random variable) be the number of customers in the system. The probability there are n customers in the system at equilibrium is:

$$P_n = \mathbb{P}\{N = n\}$$

The average number of customers in the system is:

$$L = E[N] = \sum_{i=0}^{\infty} nP_n$$

and the average number of customers in the waiting line is:

$$L_q = \sum_{n=c+1}^{\infty} (n-c)P_n$$

The relationship between L,W,L_q,W_q is defined by **Little's formula** :

$$L = \lambda_{eff} W$$
$$L_q = \lambda_{eff} f W_q$$

where λ_{eff} represents the effective arrival rate at the system. If all arriving customers can join the system then $\lambda_{eff} = \lambda$, otherwise $\lambda_{eff} < \lambda$.

A relationship between W and W_q can also be defined as:

$$W = W_q + \frac{1}{\mu}$$

where the sum of expected waiting time in queue (W_q) and expected service time ($1/\mu$) gives the total expected waiting time in the system (W).
By multiplying the previous equation by λ_{eff} a new relationship arises:

$$W\lambda_{eff} = W_q\lambda_{eff} + \frac{\lambda_{eff}}{\mu}$$

a relationship between L and L_q can be defined, taking also into consideration Little's formula:

$$L = L_q + \frac{\lambda_{eff}}{\mu}$$

Expected number of busy servers must be the difference between the average number in system (L) and the average number in the queue (L_q). Thus,

$$\overline{c} = L - L_q = \frac{\lambda_{eff}}{\mu}$$

3.7 Little's Law

W. S. Jewel in his research "A simple proof of L = λW" stated that J.D.C. Little's proof of "L = λW" ranks as one of the most important unifying results of queueing theory. According to Jewel, as Little himself has remarked, in a private communication: "the author must be congratulated for the rigor of his presentation, but he might have explained the ideas a little more"[7, p.1].

A proof is represented according to J.D. Little and W.S. Jewel.
Given that queue discipline is FIFO, let:
a(t) = number of arrivals during [0,t]
d(t) = number of departures during [0,t]
$\eta(t)$ = number of customers in the system at time t
τ_i = interval time between successive arrivals of the $(i-1)^{st}$ and i^{th} customer
ω_i = waiting time in the system of the i_{th} customer

Supposing that the queueing system is idle at t=0, then the customers in the system derive from:

$$N(\tau) = a(\tau) - d(\tau)$$

Let γ_τ be the area which is defined from the variables a_τ and d_τ in [0,t], as it is represented in figure 3.3. The integral of this area gives the total waiting time of the i_{th} customer in the system:

$$\gamma_\tau = \int_0^t N(t) d\tau$$

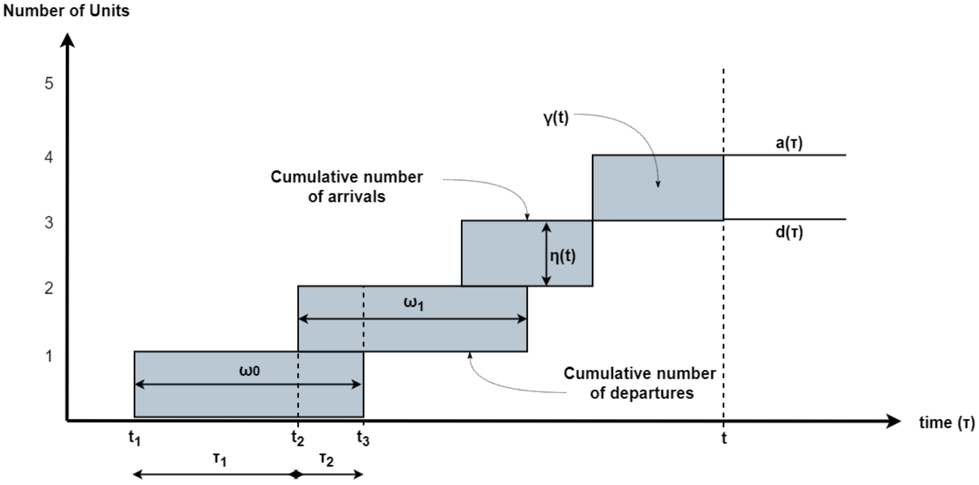
Number of Units
5
4
3
2
1
γ(t)
a(τ)
Cumulative number of arrivals
d(τ)
η(t)
ω1
ω0
Cumulative number of departures
time (τ)
t1
t2
t3
t
τ1
τ2

Figure 3.3: Representation of a busy system

Supposing λ_t represents the mean of arrivals in the system in [0,t], then:

$$\lambda_t = \frac{a(t)}{t}$$

Supposing W_t is the mean waiting time of customers in the system in [0,t], then:

$$W_t = \frac{\gamma_t}{a(t)}$$

Combining the relations above, N_t which represents the mean number of customers in the system in [0,t] can be defined as:

$$N_t = \frac{1}{t} \int_0^t N(t) d\tau = \frac{1}{t}\gamma(t) = \frac{\lambda_t}{a(t)} \cdot \gamma(t) = \lambda_t \cdot W_t$$

Let $t \to \infty$, then the following limits are assumed:

$$\lim_{t\to\infty} N_t = \lim_{t\to\infty}(\lambda_t \cdot W_t), \quad lim_{t\to\infty}\lambda_t = \lambda, \quad lim_{t\to\infty}W_t = W$$

Thus, Little's law formula is deduced:

$$L = \lambda W$$

where the number of customers in the system can be defined by multiplying the average arrival rate of the customers to the system with the average system time per customer.
A great characteristic of Little's law is that it may be applied individually regarding the queue and the service channels. Supposing L_q and L_s represent the average number of customers in the waiting line and in the service channels, respectively and W_q and W_s the average time spent in the queue and for receiving service, then according to Little's formula:

$$L_q = \lambda W_q \quad \text{and} \quad L_s = \lambda W_s$$

and

$$L = L_q + L_s = \lambda W_q + \lambda W_s = \lambda(W_q + W_s) = \lambda W$$

It is also important to mention that the proof of Little's law is independent of [22, p.401]:

- specific asumptions regarding the arrival distribution
- specific asumptions regarding the service time distribution
- the number of servers
- the particular queueing discipline

Chapter 4

The Exponential Distribution and the Poisson Process on Queues

The arrivals in a queueing system are considered to occur randomly. The arrival of a customer or the completion of a service channel are not influenced by the period of time that has elapsed since the last event. Thus, a probability distribution of *interarrival times* and a probability distribution of *service times* are essential in order to determine a queueing system.
In order to formulate a valid queueing theory model, not only should its form be *realistically applicable* but also *mathematically tractable*.

4.1 The Exponential Distribution

Exponential distributions are considered ideal for the description of totally random phenomena and as a result for a plethora of queueing systems.

Supposing that a random variable T represents either interarrival or service times. This random variable is defined, by the exponential distribution, as

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & otherwise \end{cases}$$

where $f(t)$ is the probability density function with parameter $\lambda > 0$. The respective cumulative distribution function is

$$P\{t \leq T\} = F(t) = \int_0^t f(t)dt = \begin{cases} 1 - e^{-\lambda t}, & t \geq 0 \\ 0, & otherwise \end{cases}$$

The mean of the exponential distribution, $E[T]$, is given by

$$\begin{aligned} E[T] &= \int_{-\infty}^{\infty} tf(t)dt \\ &= \int_0^{\infty} \lambda t e^{-\lambda t} dt \\ &= -te^{-\lambda t}|_0^{\infty} + \int_0^{\infty} e^{-\lambda t} dt \\ &= \frac{1}{\lambda} \end{aligned}$$

The moment generating function $\phi(t)$ is given by

$$\begin{aligned} \phi(t) &= E[e^{tX}] \\ &= \int_0^{\infty} e^{tx} e^{-\lambda x} dx \\ &= \frac{\lambda}{\lambda - t} \quad \text{for t} < \lambda \end{aligned}$$

By using the previous equation, $E[X^2]$ can be calculated as

$$\begin{aligned} E[T^2] &= \frac{d^2}{dt^2}\phi(t)|_{t=0} \\ &= \frac{2\lambda}{(\lambda - t)^3}|_{t=0} \\ &= \frac{2}{\lambda^2} \end{aligned}$$

Consequently,

$$\begin{aligned} Var(T) &= E[T^2] - (E[T])^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2} \end{aligned}$$

4.1.1 Exponential Distribution: Ideal for Queueing theory models

The exponential distribution is best suitable for a queueing theory model because of its six characteristic properties as Hillier F. and Lieberman G. state in their book "Introduction to Operations Research" [14].

1. The probability density function $f_T(t)$ is a strictly decreasing function of t (t$\geq$0). This, in terms of probability is presented as :

   $$P\{0 \leq T \leq \Delta t\} > P\{t \leq T \leq t + \Delta t\}$$

   for any positive values of Δt and t. If the service time required is essentially the same for each customer then the actual service times are expected to take values near the expected service time. A special case where the service time is far lower than the mean time is considered impossible as even a top speed service channel needs time to complete a required service operation. This special situation can not be predicted by using exponential distribution.
   On the other hand, in situations where the nature of service may be the same there are infrequent cases where the type and amount of service may differ. A real-life example could be bank tellers, as most of the time the required service is rather brief, there are situations where extensive service is required.
   If T represents interarrival times, the case where potential customers postpone their entry to a queueing system if they see another customer precede can not be predicted by applying the exponential distribution. On the other hand, it is important to mention the plethora of common phenomena of arrivals which occur randomly.

2. Forgetfulness or lack of memory
   In mathematical terms this property is stated as

   $$P\{T > t + \Delta t | T > \Delta t\} = P\{T > t\}$$

   for any positive quantities t and Δt.
   In other words, the probability distribution of the remaining arrival or service time T does not depend on how much time (Δt) has passed until

the event occurs, thus the probability is going to be always the same. The process seems to forget its past and with the help of exponential distribution is stated that

$$\begin{aligned}
P\{T > t + \Delta t | T > \Delta t\} &= \frac{P\{T > \Delta t, T > t + \Delta t\}}{P\{T > \Delta t\}} \\
&= \frac{P\{T > t + \Delta t\}}{P\{T > \Delta t\}} \\
&= \frac{e^{-\lambda(t+\Delta t)}}{e^{-\lambda \Delta t}} \\
&= e^{-\lambda t} \\
&= P\{T > t\}
\end{aligned}$$

If T represents interarrival time then this property states that the next arrival does not depend on the last arrival occurred. Furthermore, if T represents the service time it is important to underline that the type of service varies for each customer and as a result the service operations are not fixed. In case a long duration of a service occurs then the only implication is that the specific customer requires extensive service and it will not affect or be affected from service times of other customers.

3. The minimum of several independent exponential random variables has an exponential distribution.
Let $T_1, T_2, ..., T_n$ be independent exponential random variables with parameters $\lambda_1, \lambda_2, ..., \lambda_n$ respectively and U be a random variable that takes the minimum value of $T_1, T_2, ..., T_n$.That is
$U = min\{T_1, T_2, ..., T_n\}$
Given that T_i represents the time an event occurs then U represents the time until the first of the n events occurs.Thus, for any $t \geq 0$,

$$\begin{aligned}
P\{U > t\} &= P\{T_1 > t, T_2 > t, ..., T_n > t\} \\
&= P\{T_1 > t\}P\{T_2 > t\}...P\{T_n > t\} \\
&= e^{-\lambda_1 t}e^{-\lambda_2 t}...e^{-\lambda_n t} \\
&= exp(-\sum_{i=1}^{n} \lambda_i t)
\end{aligned}$$

Regarding the results, it is deduced that U indeed has an exponential distribution with parameter $\lambda = \sum_{i=1}^{n} \lambda_i t$.

In the matter of queueing models, suppose that there are n different types of customers and the interarrival time of each type follows an exponential distribution with parameter λ_i. Taking into consideration Property 2, the remaining time from any specified moment until the next arrival of a customer has the same distribution. Thus, T_i can be the remaining time measured from the instant a customer of any type arrives. Taking a step further according to Property 3 U, which is represented as the interarrival times of a queueing system as a whole, follows an exponential distribution with parameter $\lambda = \sum_{i=1}^{n} \lambda_i t$. As a result, this property gives the opportunity to ignore the distinction between customers and still have exponential interarrival times. On the other hand, regarding the service times in multiple servers, let c be the number of channels that currently provide service and let T_i be the remaining service time for server i, which follows an exponential distribution with parameter $\lambda_i = \mu$. By using Property 3, U has an exponential distribution with parameter $\lambda = c\mu$, as the queueing system is currently performing as a single server system. Variable U can also be used to determine the probability an exponential random variable will turn out to be the one with the minimum value. This probability can be calculated by:

$$P\{T_j = U\} = \lambda_j / \sum_{i=1}^{n} \lambda_i$$

for j = 1,2,...,n.

4. Relationship between Exponential and Poisson distribution
Suppose that the time between consecutive arrivals or service completions follows an exponential distribution with parameter λ. Let N(t) be the number of events by time t ($t \geq 0$), then the probability distribution of the number of times an event occurs within a specific period of time is given by:

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

for n = 0,1,2,...
Variable $N(t)$ is defined as a renewal process whose distributions of time are exponential. This special renewal process which is discrete-state and continuous parameter is known as **Poisson process**.

The mean of Poisson distribution is

$$E\{X(t)\} = \lambda t$$

which states that the expected number of events per unit time is λ. If N(t) represents the number of service operations completed by a busy channel by time t then $\lambda = \mu$ and if completed by multiple channels $\lambda = c\mu$, where c = continuously busy channels.
If N(t) represents the number of arrivals until time t, where interarrival times have an exponential distribution with parameter λ then this queueing model is said to have a Poisson input.

5. For t>0, $P\{T \leq t + \Delta t | T > t\} \approx \lambda \Delta t$, where Δt is considered small. In other words, the probability an event will occur within time of fixed length Δt ($\Delta t > 0$) is a constant. Given λ is the mean rate an event occurs, the expected number of events within time of length Δt is $\lambda \Delta t$.

$$\begin{aligned} P\{T \leq t + \Delta t | T > t\} &= P\{T \leq \Delta t\} \\ &= 1 - e^{-\lambda \Delta t} \end{aligned}$$

for any $t \geq 0$.
Using the series expansion of e^x for any exponent x:

$$e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}$$

The probability will now be equal to:

$$\begin{aligned} P\{T \leq t + \Delta t | T > t\} &= 1 - 1 + \lambda \Delta t - \sum_{n=2}^{\infty} \frac{(\lambda \Delta t)^n}{n!} \\ &\approx \lambda \Delta t \end{aligned}$$

as for small values of $\lambda \Delta t$ the summation terms becomes insignificant. Regarding queueing models, this property can be used for predicting if an event will occur in the next small interval of time Δt.

6. Cases of aggregation or disaggregation leave it unaffected
Let n different types of customers (customer of type i) arrive in a queueing system according to a Poisson input process with parameter λ_i.

Given that the customers arrivals are independent Poisson processes then *the aggregate* input process must be Poisson with parameter the arrival rate $\lambda = \sum_{i=1}^{n} \lambda_i$. The aggregate process is eventually Poisson as it is deduced from Properties 3,4 & 5.
On the other hand, supposing the aggregate input process is Poisson with parameter λ then the disaggregated input processes must be Poisson according to this property. Given that each arriving customer of type i (i =1,2,..., n) has a fixed probability P_i with

$$\lambda_i = P_i \lambda \text{ and } \sum_{i=1}^{n} P_i = 1$$

then each customers input process is Poisson with parameter λ_i and thus the Poisson process is unaffected by disaggregation.
William J. Stewart also refers to cases of aggregation and disaggregation as superposition (pooled stream) and decomposition of Poisson streams respectively [22, p.393-4].

Example 4.1.1. Suppose that the mean time a customer spends in a bank is ten minutes and the time follows the exponential distribution. What is the probability that a customer will spent more than fifteen minutes in the bank? What is the probability that a customer will spend more than fifteen minutes in the bank given that he is already in the bank for ten minutes? [20, p.285]
Let X represent the amount of time a customer spends in the bank, then the probability he spends more than fifteen minutes in the bank is given by

$$P\{X > 15\} = e^{-15\lambda} = e^{-3/2} \approx 0.220, \quad \text{where } \lambda = 1/10$$

Regarding the second question, due to the memoryless property of the exponential distribution, the probability will be equal to the probability that a new customer will spend at least five minutes in the bank. Thus the required probability is

$$P\{X > 5\} = e^{-5\lambda} = e^{-1/2} \approx 0.604$$

4.2 Poisson Queueing model

The general queueing model refers to both interarrival and service times that follow the exponential distribution. In order to develop a general model, a queueing system in equilibrium is assumed, where the system is said to achieve a steady state behaviour after it operates for a long duration. Furthermore, arrivals and departures are assumed to be state dependent as the number of customers in the system affects them. Let
n = number of customers in the system
λ_n = arrival rate, given n customers in the system
μ_n = departure rate, given n customers in the system
P_n = Probability there are already n customers in the system (in equilibrium)
In order to make it more clear, a transition-rate diagram is used in order to explain the probabilities P_n, where n represents the state of the system in other words, its customers.

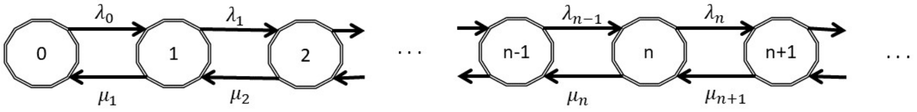

Figure 4.1: Transistion-rate diagram

For example, in state n, a step forward can be made to state n+1 where an arrival occurs at rate λ_n or a step back to state n-1 where a departure occurs at rate μ_n. The term λ_0 refers to the transition from state 0 to state 1 when an arrival at rate λ_0 occurs. On the other hand, μ_n stops at n=1 as μ_0 can not be defined if the system is empty and as a result no departures can occur.
If a queueing system is at equilibrium (n>0), it is expected that the arrival and departure rates are going to be equal.

The expected rate for transition to state n regardless the way it reaches state n is:

$$\text{(Rate for transition into state n)} = \lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1}$$

Accordingly, the expected rate for transition out of state n regardless the way it leaves state n is:

$$\text{(Rate for transition out of state n)} = (\lambda_n + \mu_n)P_n$$

The previous equations are more understandable by observing the Figure 4.1. Given that the queueing system is in steady-state then the balance equation will be:

$$\lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1} = (\lambda_n + \mu_n)P_n \text{ for n=1, 2, ...}$$

For n=0:

$$\lambda_0 P_0 = \mu_1 P_1$$

Thus, given P_0, for n=0:

$$P_1 = (\frac{\lambda_0}{\mu_1})P_0$$

By induction, a general form of P_n is achieved:

$$P_n = (\frac{\lambda_{n-1}\lambda_{n-2}...\lambda_0}{\mu_n\mu_{n-1}...\mu_1})P_0 \text{ for n=1, 2, ...}$$

where the value of P_0 is determined by using the property $\sum_{n=0}^{\infty} P_n = 1$.

4.2.1 Counting Process

A stochastic process $\{N(t), t \geq 0\}$ is denoted as a counting process if $N(t)$ represents the total number of events that have occurred by time t. For example, supposing $N(t)$ is the number of customers who enter a store at or prior to time t, then $\{N(t), t \geq 0\}$ is a counting process and a customer entering the store is denoted as an event. It is important to pinpoint that if $N(t)$ was set equal to the number of customers in the store at time t, then $N(t)$ could not be characterized as a counting process.

A counting process $N(t)$ must satisfy the properties presented below:

i. $N(t) \geq 0$.

ii. $N(t)$ is integer valued.

iii. If s<t, then $N(s) \leq N(t)$.

iv. For s<t, $N(t)-N(s)$ gives the number of events that occur in the interval (s,t].

A counting process is said to posses *independent increments* if the number of events that occur in disjoint time intervals are independent [20, p.303]. For example, the number of events that have occurred by time $t = 10, N(10)$, must be independent of the number of events that have occurred between times 10 and 15, in other words, $N(15) - N(10)$.
A counting process is said to posses *stationary increments*, if the number of events that occur in any interval of time follows a distribution that depends only in the length of the time interval. In other words, if the number of events in the interval $(s, s + t)$ follows the same distribution for all s.

4.2.2 Poisson Process

A Poisson process is a counting process $\{N(t), t \geq 0\}$ with rate λ, $\lambda > 0$, if [20, p.305]

i. N(0) = 0.

ii. The process has independent and stationary increments.

iii. The number of events in any interval of length t follows a Poisson distribution with mean λt. In other words, for all s, $t \geq 0$

$$P\{N(t+s) - N(s) = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad \text{n=0,1,...}$$

and as result

$$E[N(t)] = \lambda t$$

An alternative of Property iii is:

- $P\{N(h) = 1\} = \lambda h + o(h)$
- $P\{N(h) \geq 2\} = o(h)$

The following graph shows the probabilities of $P_n(t)$ for various values of n.

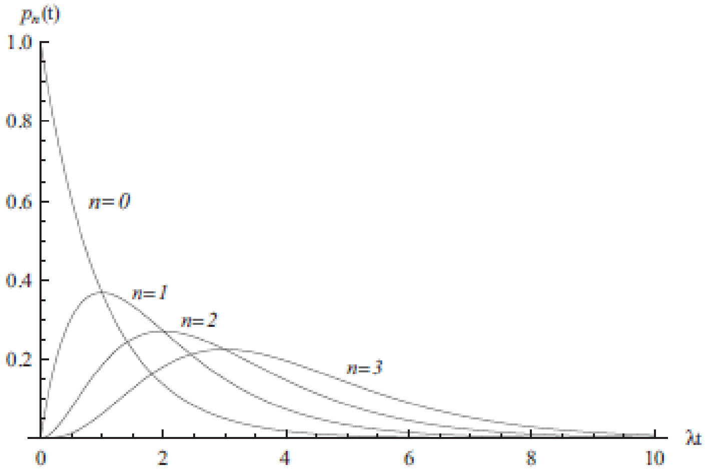

Figure 4.2: Probabilities $P_n(t)$ for n=0,1,2,3

4.2.3 Distributions of Interarrival and Waiting time

For $n > 1$, let T_n denote the elapsed time between the $(n-1)$st and the n^{th}. The sequence of interarrival times is denoted as $\{T_n, n = 1, 2, ..\}$. For instance, if $T_1 = 2$ and $T_2 = 4$, then the first event would have occurred at time 2 and the second at time 6.
Given the event $\{T_1 > t\}$ takes place if and only if no events of the Poisson process occurs in the interval [0,t] then

$$P\{T_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}$$

which means that T_1 follows the exponential distribution with mean $1/\lambda$.
For event $T_2 > t$,

$$P\{T_2 > t\} = E[P\{T_2 > t|T_1\}]$$

However,

$$\begin{aligned} P\{T_2 > t | T_1 = s\} &= P\{0 \text{ events in } (s, s+t] | T_1 = s\} \\ &= P\{0 \text{ events in } (s, s+t]\} \\ &= e^{-\lambda t} \end{aligned}$$

Therefore, it is concluded that T_2 also follows the exponential distribution with mean $1/\lambda$. As a result, T_2 is independent of T_1. In general, it is concluded that $T_n, n = 1, 2, 3, ...$ are independent identically distributed exponential random variables with mean $1/\lambda$.
The arrival time of the nth event, also known as the waiting time, is denoted as S_n and is calculated by

$$S_n = \sum_{i=1}^{n} T_i, \quad \text{n} \geq 1$$

The quantity S_n follows a gamma distribution with parameters n and λ. Thus the probabaility density of S_n is given by

$$f_{S_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \geq 0$$

The equation above can also be derived by noting that the n^{th} event will occur prior to or at time t if and only if the number of events occurring by time t is at least n [20, p.308]. In other words, the following relation must be applicable

$$N(t) \geq n \Leftrightarrow S_n \leq t$$

Hence,

$$F_{S_n}(t) = P\{S_n \leq t\} = P\{N(t) \leq n\} = \sum_{j=n}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!}$$

Consequently, by differentiating,

$$\begin{aligned} f_{S_n}(t) &= -\sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} + \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} + \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} \\ &= \lambda e^{\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \end{aligned}$$

Proposition. *Given $\{N_1(t), t \geq 0\}$ is a Poisson process where each event occurs with probability p and $\{N_2(t), t \geq 0\}$ is a Poisson process where each event occurs with probability (1-p) then these two processes have rates λp and $\lambda(1-p)$ respectively. They will also be independent.*

4.2.4 Conditional distribution of Arrival times

Due to the characteristic of Poisson process of having independent and stationary increments, it can be concluded that each interval in [0,t] of equal length has the same probability of an event occurring during this interval. This conclusion is confirmed, for $s \leq t$,

$$\begin{aligned} P\{T_1 < s|N(t) = 1\} &= \frac{P\{T_1 < s, N(t) = 1\}}{P\{N(t) = 1\}} \\ &= \frac{P\{1 \text{ event in } [0, s), 0 \text{ events in } [s, t]\}}{P\{N(t) = 1\}} \\ &= \frac{P\{1 \text{ event in } [0, s)\}P\{0 \text{ events in } [s, t]\}}{P\{N(t) = 1\}} \\ &= \frac{\lambda s e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\ &= \frac{s}{t} \end{aligned}$$

A graph of arrival times in Poisson process is presented below

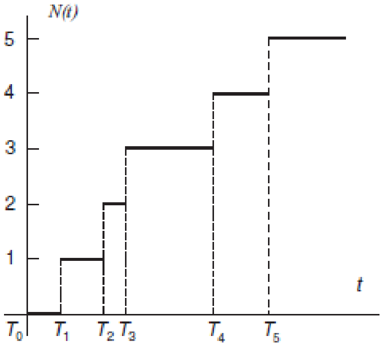

Figure 4.3: The Poisson process and some arrival times T_n

In order to generalize the previous results, it is important to mention a tool of statistics *the order statistics*.

Let $Y_1, Y_2, ..., Y_n$ be n random variables. The terms $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ are known as the order statistics corresponding to $Y_1, Y_2, ..., Y_n$ if $Y_{(k)}$ is the k_{th} smallest value among $Y_1, ..., Y_n$, where $k = 1, 2, ..., n$. For example, if n=4 and $Y_1 = 3, Y_2 = 5, Y_3 = 7$ and $Y_4 = 1$ then $Y_{(1)} = 1, Y_{(2)} = 3, Y_{(3)} = 5, Y_{(4)} = 7$. If the Y_i, i=1, 2,..., n are independent, identically distributed continuous random variables with probability density f, then the joint density of the order statistics $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ is given by [20, p.317]

$$f(y_1, y_2, ..., y_n) = n!\Pi_{i=1}^{n} f(y_i), \quad y_1 < y_2 < ... < y_n$$

As

1. $(Y_{(1)}, Y_{(2)}, ..., Y_{(n)})$ will equal $(y_1, y_2, ...y_n)$ if $(Y_1, Y_2, ..., Y_n)$ is equal to any of the n! permutations of $(y_1, y_2, ..., y_n)$
2. the probability density that $(Y_1, Y_2, ..., Y_n)$ is equal to $y_{i_1}, y_{i_2}, ..., y_{i_n}$ is $\Pi_{j=1}^{n} f(y_j)$ when $i_1, ..., i_n$ is permutation of 1,2,...,n.

Given $Y_i, i = 1, 2, ..., n$ are uniformly distributed over (o,t), the joint density function of the order statistics $Y_{(1)}, Y_{(1)}, ...Y_{(n)}$ is

$$f(y_1, y_2, ..., y_n) = \frac{n!}{t^n}, \quad 0 < y_1 < y_2 < ... < y_n < t$$

Theorem. *Given that $N(t) = n$, the n arrival times $S_1, S_2, ..., S_n$ have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval (0,t).*

Proof. Given that $N(t) = n$, for $0 < S_1 < S_2 < ... < S_n < t$ the event that $S_1 = s_1, S_2 = s_2, ..., S_n = s_n$ is equivalent to the event that the first n+1 interarrival times satisfy $T_1 = s_1, T_2 = s_2 - s_1, ..., T_n = s_n - s_{n-1}, T_{n+1} > t - s_n$. Therefore, the conditional joint density of $S_1, ..., S_n$ is

$$\begin{aligned} f(s_1,...,s_n|n) &= \frac{f(s_1,...,s_n,n)}{P\{N(t)=n\}} \\ &= \frac{\lambda e^{-\lambda s_1}\lambda e^{-\lambda(s_2-s_1)}...\lambda e^{-\lambda(s_n-s_{n-1})}e^{-\lambda(t-s_n)}}{e^{-\lambda t}(\lambda t)^n/n!} \\ &= \frac{n!}{t^n}\ , \quad 0<s_1<...<s_n<t \end{aligned}$$

□

Proposition. *If $N_i(t)$, i=1,...,k represents the number of type i (type I event with probability p or type II event with probability (1-p)) events occurring by time t then $N_i(t)$, i=1,...,k are independent Poisson random variables having*

$$E[N_i(t)] = \lambda \int_0^t P_i(s)ds$$

Proof. In order to compute the joint probability $P\{N_i(t)=n_i, i=1,...,k\}$ there must have been a total of $\sum_{i=1}^k n_i$. Hence, conditioning on N(t) yields

$$\begin{aligned} &P\{N_1(t)=n_1,...,N_k(t)=n_k\} \\ &\qquad = P\{N_1(t)=n_1,...,N_k(t)=n_k|N(t)=\sum_{i=1}^k n_i\} \times P\{N(t)=\sum_{i=1}^k n_i\} \end{aligned}$$

Let an arbitrary event that occurred in the interval [0,t] have probability $P_i(s)$ for being an event of type i. According to a previous theorem, this event will have occurred at some time uniformly distributed on (0,t) and the probability that this event will be type i is given by

$$P_i = \frac{1}{t}\int_0^t P_i(s)ds$$

independently of the other events. Therefore

$$P\{N_i(t)=n_i, i=1,...,k|N(t)=\sum_{i=1}^k n_i\}$$

will equal the multinomial probability of n_i type i outcomes for $i = 1, ..., k$ when each of $\sum_{i=1}^k n_i$ results in outcome i with probability $P_i, i = 1, ..., k$. That is

$$P\{N_1(t) = n_1, ..., N_k(t) = n_k | N(t) = \sum_{i=1}^k n_i\} = \frac{(\sum_{i=1}^k n_i)!}{n_1! \cdots n_k!} P_1^{n_1} \cdots P_k^{n_k}$$

Therefore,

$$\begin{aligned} &P\{N_1(t) = n_1, ...N_k(t) = n_k\} \\ &\quad = \frac{(\sum_i n_i)!}{n_1! \cdots n_k!} P_1^{n_1} \cdots P_k^{n_k} e^{-\lambda t} \frac{(\lambda t)^{\sum_i n_i}}{(\sum_i n_i)!} \\ &\quad = \Pi_{i=1}^k e^{-\lambda t P_i} (\lambda t P_i)^{n_i} / n_i! \end{aligned}$$

□

4.2.5 PASTA: Poisson Arrivals See Time Averages

A great property of the Poisson arrival process is that this process, once the queueuing system enters a steady state, sees the same distribution as a random observer. Supposing P_n represents the probability a balanced queueing system contains n number of customers and a_n represents the probability n customers are already in a queueing system when a new customer is about to enter it, then according to PASTA when the system is at equilibrium $P_n = a_n$.

Proof. Given
N(t) = number of customers in system at time t
$P_n(t) = \mathbb{P}\{\text{System in state n (n customers) at time t}\} = \mathbb{P}\{\text{N(t)=n}\}$
$a_n(t) = \mathbb{P}\{\text{Arrival at time t when the system is in state n}\}$
A(t,t+δt] = an arrival occurs within (t,t+δt]
The probability a customer arrives in the system at time t, given the system already includes n customers is:

$$\begin{aligned}
a_n(t) &= \lim_{\delta t \to 0} \mathbb{P}\{N(t) = n | A(t, t + \delta t]\} \\
&= \lim_{\delta t \to 0} \frac{\mathbb{P}\{N(t) = n \text{ and } A(t, t + \delta t]\}}{\mathbb{P}\{A(t, t + \delta t]\}} \\
&= \lim_{\delta t \to 0} \frac{\mathbb{P}\{A(t, t + \delta t]|N(t) = n\}\mathbb{P}\{N(t) = n\}}{\mathbb{P}\{A(t, t + \delta t]\}} \\
&= \lim_{\delta t \to 0} \frac{\mathbb{P}\{A(t, t + \delta t]\}\mathbb{P}\{N(t) = n\}}{\mathbb{P}\{A(t, t + \delta t]\}} \\
&= \mathbb{P}\{N(t) = n\} = P_n(t)
\end{aligned}$$

□

Since the Poisson arrival process has the memoryless property then $\mathbb{P}\{A(t, t+ \delta t\}$ is independent of the history and the current state of the arrival process N(t), thus:

$$\mathbb{P}\{A(t, t + \delta t]|N(t) = n\} = \mathbb{P}\{A(t, t + \delta t]\}$$

It is important to mention that for M/·/· systems, that special property holds that arriving customers find on average the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time. In general this property is not true. For instance, in a D/D/1 system which is empty at time 0, and with arrivals at 1, 3, 5, . . . and service times 1, every arriving customer finds an empty system, whereas the fraction of time the system is empty is 1/2. [6, p.27]

Chapter 5

Birth & Death Process

The birth and death process constitutes a special case of Markov chains with continuous time parameter. The name "birth and death process" derives from the applications of these processes in the study of biological processes such as the growth of bacteria populations (Bailey, 1964) [3, p.62]. Regarding queueing systems, arrivals to the system are referred to as *births* and departures as *deaths*. Furthermore, variables λ_n and μ_n are now defined as birth rate and death rate, respectively. For all n, it is assumed that λ_n and μ_n are independent of time, in other words they are Markov chains with stationary transition probabilities, but they frequently depend on the occupied state of the system n.

5.1 Definition

A Markov chain $\{x_t,\ \text{t} \in [0,\infty)\}$ with state space the set of non negative integers where [3, p.61]:

$$\begin{aligned}
&\lambda_i = q_{i,i+1} \quad i = 0,1,2,...,\\
&\mu_i = q_{i,i-1} \quad i = 1,2,...,\\
&q_{ij} = 0 \quad \text{for } j \neq i \text{ and } j \neq \pm i+1\ ,\ i = 0,1,2,...\\
&q_i = \lambda_i + \mu_i \quad i = 0,1,... \text{ and } \mu_0 = 0
\end{aligned}$$

is called: *a birth process* if all $\mu_i = 0$, for $i = 1, 2, ...$, *a death process* if all $\lambda_i = 0$, for $i = 0, 1, 2, ...$, *a birth and death process* if at least some of the λ_i and μ_i are positive.

In terms of queueing theory, *birth* is assumed when a new customer arrives in the queueing system and *death* is assumed when a customer exits the system after a successful service. This process describes probabilistically how the number of customers in the system denoted by N(t) ($t \geq 0$) changes during a period of time. The probability distribution of the remaining time until the next arrival, in other words birth, occurs is considered exponential with parameter λ_n, where $n = 0, 1, 2, ...$. The probability distribution of the remaining time until the next completed service (death) occurs is considered exponential with parameter μ_n, where $n = 1, 2, 3, ...$. The results of the analysis of a birth and death process, in this chapter, are deduced based on a steady state condition of the system, in other words λ_n and μ_n are considered independent of time. Some results of a birth and death process when the system is in transient condition are given by S. Karlin and J. McGregor [11].
Given $p_n(t)$ is the probability there are n customers in the system at time t, the probabilities of births and deaths in the system are given below:

$$\begin{aligned}
\mathbb{P}\{\text{One birth in } (t, t+h] | N(t) = n\} &= \lambda_n h + o(h) \\
\mathbb{P}\{\text{One death in } (t, t+h] | N(t) = n\} &= \mu_n h + o(h) \\
\mathbb{P}\{\text{Zero births in } (t, t+h] | N(t) = n\} &= 1 - \lambda_n h + o(h) \\
\mathbb{P}\{\text{Zero deaths in } (t, t+h] | N(t) = n\} &= 1 - \mu_n h + o(h)
\end{aligned}$$

It is evident that a death can occur only after a birth, so the process must move from state i to state i+1. Based on the research paper of Okoro O. Joshua [9] the proof of the probability a birth occurs in (t, $t+h$), where h is considered to be a negligible small interval of time, is :

$$\begin{aligned}
P_{i,i+1}(h) = P(X(t+h) - X(t) = 1 | X(t) = i) &= \frac{(\lambda_i h)^1 e^{-\lambda_i h}}{1!} \frac{(\mu_i h) e^{-\mu_i h}}{0!} + o(h) \\
&= (\lambda_i h) e^{-\lambda_i h} e^{-\mu_i h} + o(h) = (\lambda_i h) e^{-h(\lambda_i + \mu_i)} = (\lambda_i h) \sum_{n=0}^{\infty} \frac{(-h(\lambda_i + \mu_i))^n}{n!} \\
&= (\lambda_i h)(1 - h(\lambda_i + \mu_i) - \frac{1}{2!}(h)^2 (\lambda_i + \mu_i)^2 - ...) + 0(h) \\
&= \lambda_i h + o(h)
\end{aligned}$$

The probability a death occurs is equal to the transition from state i to state i-1 and is given by:

$$P_{i,i-1}(h) = P(X(t+h) - X(t) = -1|x(t) = i) = \mu_i h + o(h)$$

for $\mu_0 = 0$,$\lambda_0 \geq 0$ and μ_i,$\lambda_i \geq 0$ for i =1, 2, 3, ... by following the steps of the previous proof accordingly. The probability of having any other moves other than this two is non-zero and is given by

$$P(X(t+h) - X(t) > 1|X(t) = i) = o(h)$$

for $\mu_0 = 0, \lambda_0 > 0$ and $\mu_i, \lambda_i > 0$ for i=1,2,3,... . This also implies that

$$P(X(t+h) - X(t) = 0|X(t) = 1) = 1 - h(\lambda_i + \mu_i) + o(h)$$

Generally, the probability a birth and death process will occur, from state i to state j and within a small interval of time h, is represented by:

$$P_{i,j}(h) = \begin{cases} \lambda_i h + o(h), & \text{if j=i+1} \\ \mu_i h + o(h), & \text{if j=i-1} \\ 1 - h(\lambda_i + \mu_i) + o(h) & \text{if j=i} \\ o(h), & \text{otherwise} \end{cases}$$

This can also be written as $P_ij(h) = \delta_{i,j} + r_{i,j}(h) + h(0)$, where $\delta_{i,j} = \begin{cases} 1 & \text{j=i} \\ 0 & \text{j}\neq\text{i} \end{cases}$ the Kronecker's delta

$$g_{i,j}(h) = \begin{cases} \lambda_i & \text{if j=i+1} \\ \mu_i & \text{if j=i-1} \\ -(\lambda_i + \mu_i) & \text{if j=i} \\ 0 & otherwise \end{cases}$$

According to the above results, the matrix G is the infinitesimal generator of the process X(t) defined by $g_{i,j}$, where $g_{i,j}$ is called transition rate. [9, p.5]

$$G = \begin{vmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_i & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \end{vmatrix}$$

It is also important to mention that $\delta_{i,j} = P_{i,j}(0)$, since the probability of a process remaining in the same state at zero step is 1 and the probability a transition occurs to another state in zero step is also 1. By using the form of $g_{i,j}$

$$g_{i,j} = \frac{P_{i,j}(h) - \delta_{i,j}}{h} = \frac{P_{i,j}(h) - P_{i,j}(0)}{h} = P'_{i,j(0)}$$

Hence, by differentiating term by term for t=0 and $\sum_j g_{i,j}(0) = 0$ it is concluded that

$$g_{i,j}(t) = \begin{cases} P'_{i,j}(0) \geq 0 & \text{for } i \neq j \\ P'_{i,j}(0) \leq 0 & \text{otherwise} \end{cases}$$

Additionally, $\sum_j P_{i,j}(t) = 1$

The state transition diagram presented below depicts the rate transitions among the states in a birth-death process.

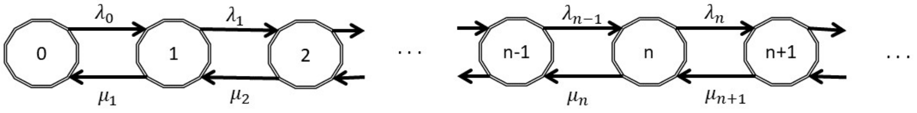

Figure 5.1: Transistion-rate diagram in birth-death process

Probability transition of birth-death process can be described by Kolmogorov backward differential equation on the initial point i [9]:

$$P_{ij}(t+h) = \sum_{k=0}^{\infty} P_{ik}(h)P_{kj}(t)$$

$$= P_{i,i-1}(h)P_{i-1,j}(t) + P_{i,i+1}(h)P_{i+1,j}(t) + P_{i,i}(h)P_{i,j}(t) + \sum_{k}^{\prime} P_{ik}(h)P_{kj}(t)$$

The last summation of the form is written in this way so as to exclude k = i-1,i and i+1.

$$\begin{aligned} P_{ij}(t+h) =& (\mu_i h + o(h))P_{i-1,j}(t) + (\lambda_i h + o(h))P_{i+1,j}(t) + \\ & (1 - h(\lambda_i + \mu_i) + o(h))P_{i,j}(t) + \sum_{k}^{\prime} P_{ik(h)Pkj(t)} \end{aligned}$$

But

$$\begin{aligned}\sum_k{}^{'} P_{ik(h)Pkj(t)} &\leq \sum_k{}^{'} P_{ik(h)} = 1 - (P_{i,i}(h) + P_{i,i+1}(h)) \\ &= 1 - (1 - h(\lambda_i + \mu_i) + o(h) + (\mu_i h + o(h) + (\lambda_i h + o(h)) \\ &= o(h)\end{aligned}$$

So the new form is :

$$\begin{aligned}P_{ij}(t+h) &= \mu_i h P_{i-1,j}(t) + \lambda_i h P_{i+1,j}(t) + (1 - h(\lambda_i + \mu_i))P_{i,j}(t) \\ &\quad + o(h)(P_{i-1,j}(t) + P_{i+1,j}(t) + P_{i,j}(t) + 1) \\ &= \mu_i h P_{i-1,j}(t) + \lambda_i h P_{i+1,j}(t) + (1 - h(\lambda_i + \mu_i)P_{i,j}(t) + o(h) \\ &= \mu_i h P_{i-1,j}(t) + \lambda_i h P_{i+1,j}(t) + P_{i,j}(t) - P_{i,j}(t) h(\lambda_i + \mu_i) + o(h)\end{aligned}$$

Also,

$$\begin{aligned}\frac{P_{i,j}(t+h) - P_{i,j}(t)}{h} &= \frac{\mu_i h P_{i-1,j}(t) + \lambda_i h P_{i+1,j}(t) - P_{i,j}(t) h(\lambda_i + \mu_i) + o(h)}{h} \\ &= \mu_i P_{i-1,j}(t) + \lambda_i P_{i+1,j}(t) - P_{i,j}(t)(\lambda_i + \mu_i) + o(1)\end{aligned}$$

So,

$$P_{i,j}^{'}(t) = \mu_i P_{i-1,j}(t) + \lambda_i P_{i+1,j}(t) - P_{i,j}(t)(\lambda_i + \mu_i)$$

Given that there is no birth without death, in other words $\mu_0 = 0$:

$$\begin{aligned}P_{0j}^{'}(t) &= \mu_0 P_{0-1,j}(t) + \lambda_0 P_{0+1,j}(t) - P_{0,j}(t)(\lambda_0 + \mu_0) \\ &= \lambda_0 P_{1,j}(t) - \lambda_0 P_{0,j}(t)\end{aligned}$$

It is known from Chapman-Kolmogorov equation that:

$$P_{ij}(s+t) = \sum_k P_{ik}(s) P_{kj}(t)$$

By differentiating with respect to s a new form develops:

$$P_{ij}^{'}(s+t) = \sum_k P_{ik}^{'}(s) P_{kj}(t)$$

Setting $s = 0$:

$$P_{ij}^{'}(t) = \sum_k P_{ik}^{'}(0) P_{kj}(t)$$

While the probability distribution of a state in time t can be described by the forward Kolomogorov differential equation, given that the initial point is fixed [9]:

$$\begin{aligned} P_{ij}(t+h) &= \sum_{k=0}^{\infty} P_{ik}(t)P_{kj}(h) \\ &= P_{i,j-1}(t)P_{j-1,j}(h) + P_{i,j+1}(t)P_{j+1,j}(h) + P_{i,j}(t)P_{j,j}(h) + \sum_{k}^{'} P_{ik}(t)P_{kj}(h) \end{aligned}$$

The last summation of the form is written in this way so as to exclude k = i-1,i and i+1.

$$P_{ij}(t+h) = P_{i,j-1}(t)\lambda_{j-1}h + P_{i,j+1}(t)\mu_{j+1}h + P_{i,j}(t)(1-h(\lambda_j+\mu_j)) + o(h)$$

Following the same steps as in Kolmogorov back differential equation :

$$\begin{aligned} &P_{ij}^{'}(t) = P_{i,j-1}(t)\lambda_{j-1} + P_{i,j+1}(t)\mu_{j+1} - P_{i,j}(t)(\lambda_j+\mu_j) \\ &P_{i0}^{'}(t) = P_{i,1}(t)\mu_1 - P_{i,0}(t)\lambda_0 \end{aligned}$$

It is known from Chapman-Kolmogorov equation that:

$$P_{ij}(s+t) = \sum_k P_{ik}(s)P_{kj}(t)$$

By differentiating with respect to t a new form develops:

$$P_{ij}^{'}(s+t) = \sum_k P_{ik}(s)P_{kj}^{'}(t)$$

Setting $t = 0$:

$$P_{ij}^{'}(s) = \sum_k P_{ik}(s)P_{kj}^{'}(0)$$

therefore

$$P'(t) = P(t)G$$

For

$$\begin{aligned} P(0) &= 1 \\ P(t) &= e^{Gt} \end{aligned}$$

where

$$e^{Gt} = \sum_{n=0}^{\infty} \frac{G^n t^n}{n!} = 1 + \sum_{n=1}^{\infty} \frac{G^n T^n}{n!}$$

5.2 The birth process

In the birth process everyone lives forever, there are no deaths. For instance, a colony of bacteria could be modeled as a birth process, as each cell randomly and independently divides into two cells at some future time. This cell dichotomy replicates over time, for each divided cell. It is assumed that births start at time t=0 with n_0 cells. It is also assumed that any cell division in the time interval $(t, t + \delta t)$ is proportional to the time interval δt, where h is considered small. If λ is the birth rate of the process, then the probability a cell divides in the interval is $\lambda\delta t$. For n cells divisions the probability will be $\lambda n\delta t$. Furthermore, the probability that two or more births occur in the time interval δt is $o(\delta t)$ and the probability of no dichotomies at all during this period will be $1 - \lambda n\delta t - o(\delta t)$. This simple birth process described is also known as the Yule[1] process, name after one of its originators.
If N(t) is the random variable then

$$P\{N(t) = n\} = p_n(t)$$

where $p_n(t)$ is the probability that the population size is n at time t. If the initial population size is $n_0 \geq 1$ at time t=0 then

$$p_{n_0}(0) = 1 \text{ and } p_{n_0} = 0 \text{ for } n > n_0$$

If the size population is n-1 at time t then the probability a birth occurs is

$$\lambda(n-1)\delta t + o(\delta t)$$

Accordingly, for $n \geq n_0 + 1$

$$p_n(t + \delta t) = p_{n-1}(t)[\lambda(n-1)\delta t + o(\delta t)] + p_n(t)[1 - \lambda n\delta t + o(\delta t)]$$

$$\frac{p_n(t + \delta t) - p_n(t)}{\delta t} = \lambda(n-1)\delta t p_{n-1}(t) - \lambda n p_n(t) + o(1)$$

For $n = n_0$ then

$$p_{n_0}(t + \delta t) = p_{n_0}(t)[1 - \lambda n_0\delta t + o(\delta t)]$$

$$\frac{p_{n_0}(t + \delta t) - p_{n_0}(t)}{\delta t} = -\lambda n_0 p_{n_0}(t) + o(1)$$

[1] George Undy Yule (1871-1951), Scottish statistician

As $\delta t \to 0$ they become derivatives in the limit so that

$$\frac{dp_{n_0}(t)}{\delta t} = -\lambda n_0 p_{n_0}(t) \tag{5.1}$$

with solution

$$p_{n_0}(t) = e^{-\lambda n_0 t}$$

and by putting $n = n_0 + 1$ in

$$\frac{dp_{n_0}(t)}{\delta t} = \lambda(n-1)p_{n-1}(t) - \lambda n p_n(t) \tag{5.2}$$

its solution is

$$p_{n_0+1}(t) = n_0 e^{-\lambda n_0 t}(1 - e^{-\lambda t})$$

It is importnat to pinpoint that since this process is a simple birth process then $p_n(t) = 0$ for n<n_0.
The probability generating function is

$$G(s,t) = \sum_{n=n_0}^{\infty} p_n(t) s^n$$

A proof of the final form of the probability generating function G(s,t) is presented by Jones, and Smith in their book "Stochastic Processes: An Introduction" [8, p.121-123]. For simple birth process

$$G(s,t) = \frac{1}{\left[1 + \frac{(1-s)}{s} e^{\lambda t}\right]^{n_0}} = \frac{s^{n_0 e^{-\lambda n_0 t}}}{\left[1 - (1 - e^{-\lambda t})s\right]^{n_0}}$$

The mean population size at time t is given by

$$\mu(t) = n_0 e^{\lambda t}$$

5.3 The death process

In the death process no birth occurs and the population numbers decline through deaths. The probability a death occurs in a short time interval δt is $\mu n \delta t$, where μ is the death rate and the size population is n. In case of multiple deaths in time interval δt the probability is negligible. By arguments similar to those for the birth process

$$p_0(t+\delta t) = [\mu\delta t + o(\delta t)]p_1(t) + [1 + o(\delta t)]p_0(t)$$

For $1 \leq n \leq n_0 - 1$

$$p_n(t+\delta t) = [\mu(n+1)\delta t + o(\delta t)]p_{n+1}(t) + [1 - \mu n\delta t - o(\delta t)]p_n(t)$$

If the initial population size is n_0, then for all t, $p_n(t) = 0$ for $n > n_0$ and

$$p_{n_0}(t+\delta t) = [1 - \mu n_0\delta t + o(\delta t)]p_{n_0}(t)$$

Thus

$$\frac{p_0(t+\delta t) - p_0(t)}{\delta t} = \mu p_1(t) + o(1)$$

$$(1 \leq n \leq n_0 - 1), \quad \frac{p_n(t+\delta t) - p_n(t)}{\delta t} = \mu(n+1)p_{n+1}(t) - \mu n p_n(t) + o(1)$$

$$\frac{p_{n_0}(t+\delta t) - p_{n_0}(t)}{\delta t} = -\mu n_0 p_{n_0}(t) + o(1)$$

Let $\delta t \to 0$

$$\begin{cases} \frac{dp_0(t)}{dt} = mp_1(t) \\ \frac{dp_n(t)}{dt} = \mu(n+1)p_{n+1}(t) - \mu n p_n(t) \quad (1 \leq n \leq n_0 - 1) \\ \frac{dp_{n_0}(t)}{dt} = -\mu n_0 p_{n_0}(t) \end{cases}$$

If the initial population size is n_0 at time t=0 then $p_{n_0}(0) = 1$.
The probability generating function G(s,t) is defined as

$$G(s,t) = \sum_{n=0}^{n_0} p_n(t)s^n$$

and its final form is [8, p.126]

$$G(s,t) = \sum_{n=0}^{n_0} \left(\begin{array}{c} n_0 \\ n \end{array} \right) e^{-n\mu t}(1 - e^{-\mu t})^{n_0 - n} s^n$$

5.4 The birth-death process

Given a population of size n can arise at time $t + \delta t$

$$p_0(t + \delta t) = [\mu\delta t + o(\delta t)p_1(t)] + [1 + o(\delta t)]p_0(t)$$

For $n \geq 1$ $$\begin{aligned} p_n(t + \delta t) &= [\lambda(n-1)\delta t + o(\delta t)]p_{n-1}(t) + [1 - (\lambda n + \mu n)\delta t + o(\delta t)p_n(t)] \\ &+ [\mu(n+1)\delta t + o(\delta t)]p_{n+1}(t) \end{aligned}$$

For $\delta t \to 0$, $p_n(t)$ satisfies

$$\begin{cases} \frac{dp_0(t)}{dt} = \mu p_1(t) \\ \frac{dp_n(t)}{dt} = \lambda(n-1)p_{n-1}(t) - (\lambda + \mu)np_n(t) + \mu(n+1)p_{n+1}(t), \quad n \geq 1 \end{cases}$$

The death process equations result if $\lambda = 0$ and the birth process equations result from $\mu = 0$. Given a birth occurs, the probability generating function is

$$G(s,t) = \sum_{n=0}^{\infty} p_n(t)s^n$$

Its final form is

$$G(s,t) = \left[\frac{\mu(1-s) - (\mu - \lambda s)e^{-(\lambda\mu)t}}{\lambda(1-s) - (\mu - \lambda s)e^{-(\lambda\mu)t}} \right]^{n_0}$$

The expected population size at time t, for $\lambda \neq \mu$ is

$$\begin{aligned} \mu(t) &= \sum_{n=1}^{\infty} np_n(t) = G_s(1,t) \\ &= \frac{n_0(-\mu + \lambda e^{-(\lambda-\mu)t})}{-(\mu - \lambda)e^{-(\lambda-\mu)t})} - \frac{n_0(-\lambda + \lambda e^{(\lambda-\mu)t})}{-(\mu - \lambda)e^{-(\lambda-\mu)t})} \end{aligned}$$

In case the birth rate equals the death rate, in other words $\lambda = \mu$, then the probability generating function is given by

$$G(s,t) = \left[\frac{1 + (\lambda t - 1)(1-s)}{1 + \lambda t(1-s)} \right]^{n_0}$$

Furthermore, in case $\lambda = \mu$ the probability of extinction at time t is given by

$$p_0(t) = G(0,t) = \left[\frac{1+(\lambda t - 1)}{1+\lambda t}\right]^{n_0} = \left[\frac{\lambda t}{1+\lambda t}\right]^{n_0}$$

As $t \to \infty$ it is assumed that

$$\lim_{t\to\infty} p_0(t) = \lim_{t\to\infty}\left[\frac{1}{1+\frac{1}{\lambda t}}\right]^{n_0} = 1$$

The previous results indicate that the birth and death rates are in balance, then the ultimate extinction is certain. An example of a birth and death process is given in the article "The numerical solution of a birth-death process arising in multimedia synchronization". [17]

5.5 Balance Equation

According to Hillier in his book 'Introduction to Operations Research' [14, p.850] the balance equation for state n is based on a key principle:

$$\textbf{Rate In} = \textbf{Rate Out}$$

which states that the mean birth rate equals mean death rate for any state of the system n (n = 0,1,2,..).
The rate at which the process enters a state n at time t is given by:

$$\lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t)$$

and the rate at which the process leaves a state n at time t is given by:

$$(\lambda_n + \mu_n)P_n(t)$$

For example, supposing the process wants to enter state 0, the only way to enter it is from state 1. The probability P_1 represents the possible transition from state 1 to state 0 within a period of time. The mean rate of entering state 0 from state 1 is μ_0 and from any other state the transition rate is 0. Consequently, the mean entering rate is:

$$\mu_1 P_1 + 0(1 - P_1) = \mu_1 P_1$$

and the mean leaving rate must be:

$$\lambda_0 P_0$$

Combing the above information, the balance equation for state 0 is:

$$\mu_1 P_1 = \lambda_0 P_0$$

It is important to underline that the transition to state 0 is a unique case as for every other transition between states there are two possibilities into or out of the state.

Balance equations for the birth-death process

State 0:

$$\mu_1 P_1 = \lambda_0 P_0 \Rightarrow P_1 = \frac{\lambda_0}{\mu_1} P_0$$

State 1:

$$\begin{aligned} \lambda_0 P_0 + mu_2 P_2 = (\lambda_1 + \mu_1) P_1 &\Rightarrow P_2 = \frac{\lambda_1}{\mu_2} P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) \\ &\Rightarrow P_2 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0 \end{aligned}$$

State 2:

$$\begin{aligned} \lambda_1 P_1 + mu_3 P_3 = (\lambda_2 + \mu_2) P_2 &\Rightarrow P_3 = \frac{\lambda_2}{\mu_3} P_2 + \frac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) \\ &\Rightarrow P_3 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0 \end{aligned}$$

... ...

State n-1:

$$\lambda_{n-2}P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1})P_{n-1} \Rightarrow P_n = \frac{\lambda_{n-1}}{\mu_n}P_{n-1} + \frac{1}{\mu_n}(\mu_{n-1}P_{n-1} - \lambda_{n-2}P_{n-2})$$
$$\Rightarrow P_n = \frac{\lambda_{n-1}\lambda_{n-2}...\lambda_0}{\mu_n\mu_{n-1}...\mu_1}P_0$$

State n:

$$\lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1} = (\lambda_n + \mu_n)P_n \Rightarrow P_{n+1} = \frac{\lambda_n}{\mu_{n+1}}P_n + \frac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1}P_{n-1})$$
$$\Rightarrow P_{n+1} = \frac{\lambda_n\lambda_{n-1}...\lambda_0}{\mu_{n+1}\mu_n...\mu_1}P_0$$

... ...

Let:

$$C_n = \frac{\lambda_{n-1}\lambda_{n-2}...\lambda_0}{\mu_n\mu_{n-1}...\mu_1} \text{ for n=1,2,..}$$

Then, the general form of steady-state probabilities is:

$$P_n = C_n P_0 \text{ for n=1,2,..}$$

The property $\sum_{n=0}^{\infty} P_n = 1$ is modified to $(\sum_{n=0}^{\infty} C_n)P_0 = 1$ so that:

$$P_0 = (\sum_{n=0}^{\infty} C_n)^{-1}$$

It is also necessary to adjust the key measures of a queuing system such as L, L_q, W and W_q when a queueing model follows a birth-death process. Thus, the new forms of L and L_q are:

$$L = \sum_{n=0}^{\infty} nP_n \quad \text{and} \quad L_q = \sum_{n=s}^{\infty}(n-s)P_n$$

Furthermore, taking into consideration the relationships among the key measures which are:

$$W = \frac{L}{\overline{\lambda}} \quad \text{and} \quad W_q = \frac{L_q}{\overline{\lambda}}$$

where $\overline{\lambda}$ constitutes the average arrival rate and its form is given by:

$$\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$$

Chapter 6

Some queueing models

In 1909, Agner K. Erlang published 'The Theory of Probabilities and Telephone Conversations' in which he proposed models that would describe the Copenhagen telephone exchange [5]. His research formed the base for the evolution of Queueing theory. In queueing systems, it is important that the utilization time of each server be determined. In a multiserver queueing system with unlimited waiting time, the level of service utilization increases in proportion to the number of servers and the arrival rate. The simplest multiserver system is the single server, which is known as the M/M/1 queue and the multiserver queueing systems are denoted by the notation M/M/c, where c stands for the number of channels. Given a birth and death process, there is a great flexibility in modeling systems where Poisson input and exponential arrival times are considered.

6.1 The M/M/1 queueing system

The most simple queueing system is the Single Server one, which is denoted as the M/M/1 system. Suppose that a single server service station follows a Poisson process with arrival rate λ. The mean rate of successive arrivals is $1/\lambda$, where each arrival is considered as an independent exponential random variable. When a customer arrives in the system, he either goes directly into service if the server is idle or he either joins the waiting line. Once

a customer is served he leaves the system (death occurs) and the next one enters the service. The mean rate of successive service times is $1/\mu$ and each service time is considered as an independent exponential random variable.
The previous process depicts a M/M/1 queueing system. The first letter M stands for the Markovian interarrival process since it is a Poisson process and the second M stands for the exponential service distribution and, hence, Markovian. The number 1 shows the number of servers in the system. A representation of how a single server queueing system works is presented below.

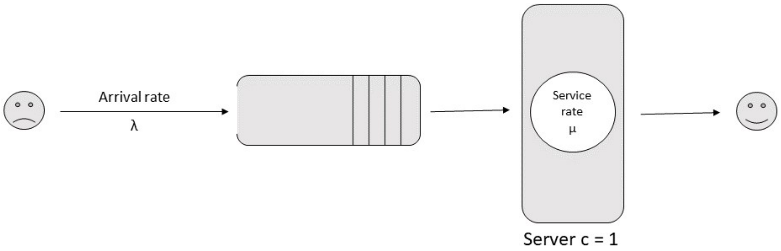

Figure 6.1: The M/M/1 queue

Results for the Single Server Queueing system

The C_n factors of this process reduce to [14, p.879]:

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n \text{ for n=0, 1, 2, ...}$$

Therefore,

$$P_n = \rho^n P_0 \text{ for n=0, 1, 2, ...}$$

where

$$\begin{aligned} P_0 &= (\sum_{n=0}^{\infty} \rho^n)^{-1} \\ &= (\frac{1}{1-\rho})^{-1} \\ &= 1-\rho \end{aligned}$$

Thus,

$$P_n = (1-\rho)\rho^n \text{ for n=0, 1, 2, ...}$$

Furthermore,

$$\begin{aligned}
L &= \sum_{n=0}^{\infty} n(1-\rho)\rho^n \\
&= (1-\rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n) \\
&= (1-\rho)\rho \frac{d}{d\rho}(\sum_{n=0}^{\infty} \rho^n) \\
&= (1-\rho)\rho \frac{d}{d\rho}(\frac{1}{1-\rho}) \\
&= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}
\end{aligned}$$

Accordingly,

$$\begin{aligned}
L_q &= \sum_{n-1}^{\infty}(n-1)P_n \\
&= L - 1(1-P_0) \\
&= \frac{\lambda^2}{\mu(\mu-\lambda)}
\end{aligned}$$

Supposing the mean arrival rate λ is greater or equal to the mean service rate μ ($\lambda \geq \mu$) then the queueing system would grow endlessly. In case the system starts operating without customer presence then the system would operate properly only for a short period of time but eventually would be impossible to avoid breakdown. Even when $\lambda = \mu$ the probabilities of rising number of customers in the system increases significantly over time. On the other hand, when the mean service rate μ is greater than the mean arrival rate λ ($\mu > \lambda$) and the queue discipline is FIFO then a new arriving customer in the system would have to wait through n+1 exponential service times including his own. Let $T_1, T_2, ...$ represent the service time random variables which follow an exponential distribution with mean service rate μ, then the total service time is represented by:

$$S_{n+1} = T_1 + T_2 + \ ... + T_n \text{ for n =0, 1, 2, ...}$$

The probability that a random arriving customer will wait more than time t in the system, taking into consideration that there are already n customers in it, is given by:

$$P(W > t) = \sum_{n=0}^{\infty} P_n P(S_{n+1} > t) = e^{-\mu(1-\rho)t} \text{ for } t \geq 0$$

The waiting time in the system, including waiting and service time, follows an exponential distribution with parameter $\mu(1-\rho)$ and is given by:

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

The waiting time in the queue regarding a random arrival when the queue discipline is FCFS (or FIFO) is denoted as W_q. Supposing a customer arrives in an idle system then he will be served immediately, it is concluded that:

$$P(W_q = 0) = P_0 = 1 - \rho$$

In case there are already n customers in the system then the probability the waiting time in queue line is more than t is given by:

$$\begin{aligned}
P(W_q > t) &= \sum_{n=1}^{\infty} P_n P(S_n > t) \\
&= \sum_{n=1}^{\infty} (1-\rho)\rho^n P(S_n > t) \\
&= \rho \sum_{n=0}^{\infty} P_n P(S_{n+1} > t) \\
&= \rho P(W > t) \\
&= \rho e^{-\mu(1-\rho)t} \text{ , for } t \geq 0
\end{aligned}$$

It is important to underline that W_q does not follow an exponential distribution as $P(W_q = 0) > 0$. The *conditional distribution of* W_q is the one that follows an exponential distribution, because

$$P(W_q > t | W_q > 0) = \frac{P(W_q > t)}{P(W_q > 0)} = e^{-\mu(1-\rho)t} \text{ , for } t \geq 0$$

The waiting time in queue results from the mean of the distribution of W_q or by applying $L_q = \lambda W_q$ or $W_q = W - 1/\mu$ and the final form of W_q is:

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

6.2 The M/M/c queueing system

This queue model constitutes a special birth and death process where the mean arrival rate λ_n and the mean service rate μ_n are considered constant as λ and μ per busy server, regardless of the system state.

Theorem. *The process $\{N(t), t \geq 0\}$ is a birth and death process with birth rate $\lambda_n = \lambda$ for all $n \geq 0$ and with death rate $\mu_n = \mu$ for all $n \geq 1$.[15, p.4]*

Proof. The process $\{N(t), t \geq 0\}$ is a Markov process as its interarrival and its service times follow the exponential distribution. Given $0(\Delta t)$ is the probability of two events occurring in a small time interval $(t, t + \Delta t)$, the following transitional probabilities are applied:

$$\begin{aligned}
\mathbb{P}[N(t+\Delta t) = n+1/N(t) = n] &= \lambda\Delta t + 0(\Delta t) \ , \ n \geq 0 \\
\mathbb{P}[N(t+\Delta t) = n-1/N(t) = n] &= \mu\Delta t + 0(\Delta t) \ , \ n \geq 1 \\
\mathbb{P}[N(t+\Delta t) = n/N(t) = n] &= 1 - (\lambda+\mu)\Delta t + 0(\Delta t) \ , \ n \geq 1 \\
\mathbb{P}[N(t+\Delta t) = n/N(t) = n] &= 1 - \lambda\Delta t + 0(\Delta t) \ , \ n = 0 \\
\mathbb{P}[N(t+\Delta t) = n/N(t) = n] &= 0(\Delta t) \ , \ |k-n| \geq 2
\end{aligned}$$

The term $0(\Delta t)$ is a quantity so that $lim_{\Delta t \to \infty} 0(\Delta t) = 0$. Consequently, it is assumed that the process N(t) is a birth-death process. The first equation gives the probability when the state variable increases by one (a single birth). The parameter $\lambda_i = \lambda$ is considered as the instantaneous birth rate. Likewise, the second equation gives the probability when the state variable is reduced by one (a single death). The parameter $\mu_i = \mu$ denotes the instantaneous death rate. The third equation refers to the case where the state variable does not change. More precisely, the term $[1 - (\lambda_i + \mu_i)]dt$ reflects the probability that neither a single birth nor a single death may occur during this infinitesimal period of time. Furthermore, multiple single births and single deaths as well as simultaneous births and deaths are being calculated by the $0(dt)$ terms of the equations. It is also important to underline the fact that the probability for these events to occur is negligible as $dt \to 0$. □

The M/M/c queueing system is similar to the M/M/1 but in this case there are c identical channels for service. This queue is also known as a queue with parallel servers. The arrival rate of customers follows a Poisson process with $\lambda_n = \lambda$ for all n and the service discipline is FIFO (or FCFS) with service rate μ. In case the number of customers is greater or equal to the number of servers ($n \geq c$) then all the service channels are occupied and the service rate will now be equal to $c\mu$. The *effective* service rate in also called the mean system output rate (MSOR) [22, p.419]. On the other hand if the number of servers is greater than the number of customers ($c \geq n$) in the system then the MSOR will be equal to $n\mu$. Furthermore, a new term needs to be mentioned the *load dependent service center* which is a center where the customers departure rate is described as a function of the number of customers in the system. The load dependent service center is given by $\mu_n = min(n, c)\mu$ where μ =service rate, n=customers in the system and c the number of channels [22, p.419]. A representation of this kind of queueing system is presented below:

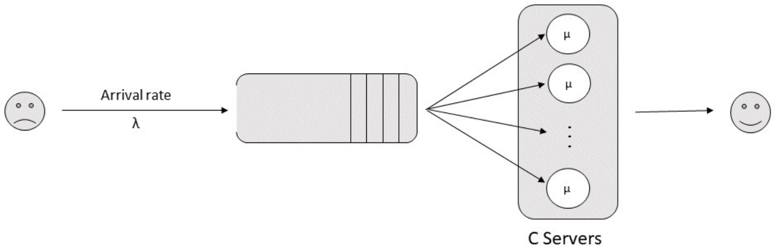

Figure 6.2: The M/M/c queue

The M/M/c queueing system is defined by its birth and the death rates which are :

$$\lambda_n = \lambda \quad \text{for all n,}$$

$$\mu_n = \begin{cases} n\mu & , 1 \leq n \leq c \\ c\mu & , n \geq c \end{cases}$$

Based on the general equations of a birth - death process :

$$P_n = P_0 \prod_{i=1}^{n} \frac{\lambda_i - 1}{\mu_i}$$

The equations for the M/M/c queue become:

$$P_n = P_0 \prod_{i=1}^{n} \frac{\lambda}{i\mu} = P_0\,(\frac{\lambda}{\mu})^n\,\frac{1}{n!} \quad \text{if } 1 \leq n \leq c,$$

and

$$P_n = P_0 \prod_{i=1}^{c} \frac{\lambda}{i\mu} \prod_{i=c+1}^{n} \frac{\lambda}{c\mu} = P_0 \, (\frac{\lambda}{\mu})^n \, \frac{1}{c!} \, (\frac{1}{c})^{n-c} \quad \text{if } n \geq c$$

The state transition diagram of the M/M/c queueing system depicts the behaviour of birth and death rates during a period of time:

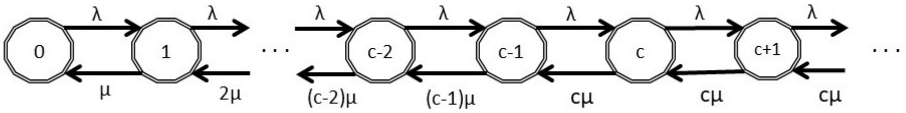

Figure 6.3: The M/M/c state transition diagram

The load factor in this kind of system derives from the relation $\rho = \lambda/c\mu$. Additionally, the condition $\rho < 1$ is required, in order to have a system in equilibrium. By applying this condition, the maximum number of busy service channels is set to $c\rho = \lambda/\mu$, which means that the mean arrival rate must be less than the mean service rate.

The new form of the probabilities is now:

$$P_n = \begin{cases} P_0 \frac{(c\rho)^n}{n!} & \text{for } n \leq c, \\ \\ P_0 \frac{(c\rho)^n}{c^{n-c} c!} = P_0 \frac{\rho^n c^c}{c!} & \text{for } n \geq c \end{cases}$$

In order to calculate P_0, the property $\sum_{n=0}^{\infty} P_n = 1$ is used:

$$\begin{aligned} \sum_{n=0}^{\infty} P_n = 1 = P_0 + \sum_{n=1}^{\infty} P_n \\ = P_0 \Big[1 + \sum_{n=1}^{c-1} \frac{(c\rho)^n}{n!} + \sum_{n=c}^{\infty} \frac{\rho^n c^n}{c!} \Big] \end{aligned}$$

So, the new form of P_0 is:

$$\begin{aligned} P_0 &= \Big[1+\sum_{n=1}^{c-1}\frac{(c\rho)^n}{n!}+\sum_{n=c}^{\infty}\frac{\rho^n c^n}{c!}\Big]^{-1} \\ &= \Big[1+\sum_{n=1}^{c-1}\frac{(c\rho)^n}{n!}+\frac{1}{c!}\sum_{n=c}^{\infty}\rho^n c^n\Big]^{-1} \\ &= \Big[1+\sum_{n=1}^{c-1}\frac{(c\rho)^n}{n!}+\frac{(c\rho)^c}{c!}\sum_{n=c}^{\infty}\rho^{n-c}\Big]^{-1} \\ &= \Big[1+\sum_{n=1}^{c-1}\frac{(c\rho)^n}{n!}+\frac{c\rho^c}{c!}\frac{1}{1-\rho}\Big]^{-1} \end{aligned}$$

Results for the M/M/c Queueing system

The expected queue length L_q has meaning only when $n \geq c$, in other words only when customers wait in the queue and is given by:

$$L_q = \sum_{n=c}^{\infty}(n-c)P_n$$

where

$$P_n = \frac{(\rho c)^n}{c^{n-c}c!}P_0 \quad \text{for } n \geq c$$

So the expected queue length L_q in the M/M/c queue is given by:

$$L_q = \sum_{n=c}^{\infty}\frac{n}{c^{n-c}c!}(\rho c)^n P_0 - \sum_{n=c}^{\infty}\frac{c}{c^{n-c}c!}(\rho c)^n P_0$$

In order to achieve a more elegant form, the right hand terms are going to be analyzed separately according to William J. Stewart's proof [22, p.421]. The first term is:

$$\frac{P_0}{c!}\sum_{n=c}^{\infty}\frac{n(\rho c)^n}{c^{n-c}}$$

where

$$\frac{n\rho^n c^n}{c^{n-c}} = n\rho^n c^c = \frac{n\rho^{n-c-1}\rho^{c+1}c^{c+1}}{c} = \frac{(\rho c)^{c+1}}{c}n\rho^{n-c-1}$$

According to the above and by using derivatives of the geometric series, the first term is modified to :

$$
\begin{aligned}
\frac{P_0}{c!}\sum_{n=c}^{\infty}\frac{n(\rho c)^n}{c^{n-c}} &= \frac{P_0}{c!}\left[\frac{(\rho c)^{c+1}}{c}\right]\sum_{n=c}^{\infty}\left[(n-c)\rho^{n-c-1}+c\rho^{n-c-1}\right] \\
&= \frac{P_0}{c!}\left[\frac{(\rho c)^{c+1}}{c}\right]\left\{\sum_{n=c}^{\infty}(n-c)\rho^{n-c-1}+\sum_{n=c}^{\infty}c\rho^{n-c-1}\right\} \\
&= \frac{P_0}{c!}\left[\frac{(\rho c)^{c+1}}{c}\right]\left\{\frac{1}{(1-\rho)^2}+\frac{c}{\rho}\frac{1}{1-\rho}\right\} \\
&= \frac{P_0}{c!}\left[\frac{(\rho c)^{c+1}}{c}\right]\left\{\frac{1}{(1-\rho)^2}+\frac{c/\rho}{1-\rho}\right\}
\end{aligned}
$$

The second term is modified to:

$$
\begin{aligned}
\frac{P_0}{c!}\sum_{n=c}^{\infty}\frac{c(\rho c)^n}{c^{n-c}} &= \frac{P_0}{c!}\sum_{n=c}^{\infty}c\rho^c\rho^{n-c}c^c = \frac{P_0}{c!}c(\rho c)^c\sum_{n=c}^{\infty}\rho^{n-c} \\
&= \frac{P_0}{c!}\frac{c(\rho c)^c}{1-\rho} = \frac{P_0}{c!}\left[\frac{(\rho c)^{c+1}}{c}\right]\frac{c/\rho}{1-\rho}
\end{aligned}
$$

By combining the modified terms, the mean number of customers waiting in the queue can now be calculated by:

$$L_q = \frac{P_0}{c!}\left[\frac{(\rho c)^{c+1}}{c}\right]\left\{\frac{1}{(1-\rho)^2}+\frac{c\rho}{1-\rho}-\frac{c/\rho}{1-\rho}\right\}$$

Thus, the final form of L_q is :

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

or

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

The mean time a customer spends in the queue waiting to receive service is calculated with the help of Little's Law and the new form is:

$$W_q = \left[\frac{(\lambda/\mu)^c\mu}{(c-1)!(c\mu-\lambda)^2}\right]P_0$$

In order to calculate W, in other words the mean time spent in the system (waiting and service time), the relationship $W = W_q + 1/\mu$ is used. Thus, the new form is:

$$W = \left[\frac{(\lambda/\mu)^c \mu}{(c-1)!(c\mu-\lambda)^2}\right] P_0 + \frac{1}{\mu}$$

Finally, the mean number of customers in the system L is also calculated with the help of Little's Law and the new form is:

$$L = \left[[\frac{(\lambda/\mu)^c \mu}{(c-1)!(c\mu-\lambda)^2}\right] P_0 + \frac{\lambda}{\mu}\right]$$

In case an arriving customer is forced to wait for service due to lack of available servers then the 'Erlang-C formula' will be used in order to calculate this probability, which is:

$$\begin{aligned} \mathbb{P}\{\text{queueing}\} &= \sum_{n=c}^{\infty} P_n = P_0 \sum_{n=c}^{\infty} \frac{c^c}{c!}\left[\frac{\rho^c}{1-\rho}\right] \\ &= \frac{(c\rho^)c}{c!(1-\rho)} P_0 = \frac{(\lambda/\mu)^c \mu}{(c-1)!(c\mu-\lambda)} P_0 \end{aligned}$$

The Erlang-C formula is denoted by $C(c, \lambda/\mu)$ and the performance measures of this special system are:

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

$$W_q = \frac{1}{\lambda}\frac{\rho C(c, \lambda/\mu)}{(1-\rho)} = \frac{C(c, \lambda/\mu)}{c\mu-\lambda}$$

$$W = \frac{1}{\lambda}\frac{\rho C(c, \lambda/\mu)}{(1-\rho)} + \frac{1}{\mu} = \frac{C(c, \lambda/\mu)}{c\mu-\lambda} + \frac{1}{\mu}$$

$$L = \frac{\rho C(c, \lambda/\mu)}{(1-\rho)} + c\rho = \frac{\lambda C(c, \lambda/\mu)}{c\mu-\lambda} + \frac{\lambda}{\mu}$$

6.3 The M/M/∞ queueing system

The M/M/∞ queueing system is considered to be the system where the number of servers is unlimited. This is the reason why the M/M/∞ queue

is also known as the infinite server.
Let:

$$\lambda_n = \lambda \quad \text{for all n}, \quad \mu_n = n\mu \quad \text{for all n}$$

The new form of the probabilities is:

$$P_n = \frac{\lambda^n}{n\mu(n-1)\mu \cdots 2\mu 1\mu} P_0 = \frac{\lambda^n}{n!\mu^n} P_0$$

By taking into consideration the property:

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

The probability a system is empty, P_0, is given by:

$$P_0 = \left[\sum_{n=0}^{\infty} \frac{\lambda^n}{n!\mu^n} \right]^{-1} = e^{-\lambda/\mu}$$

Therefore, the final form of the probability P_n is:

$$P_n = \frac{(\lambda/\mu)^n e^{-\lambda/\mu}}{n!}$$

Inferring from the above, the probability P_n follows the Poisson distribution with parameter λ/μ.
In an infinite server, the waiting time for customers does not have any meaning at all, because no one will ever wait for service. The waiting time in this system will be:

$$L = \sum_{n=1}^{\infty} nP_n = e^{-\lambda/\mu} \sum_{n=1}^{\infty} \frac{(\lambda/\mu)^n}{(n-1)!} = e^{-\lambda/\mu} \frac{\lambda}{\mu} \sum_{n=1}^{\infty} \frac{(\lambda/\mu)^{n-1}}{(n-1)!} = \frac{\lambda}{\mu}$$

In other words the mean time spent in the system will be the mean service time $(1/\mu)$ times the mean number of customers in the system λ. It is obvious that in a M/M/∞ queue the measures L_q and W_q will be equal to 0 as the meaning of waiting time is considered non-existent.

6.4 The $\mathcal{E}_r$/M/1 queue (Erlang-r Arrival model)

In the previous queues, the only probability distribution used to define arrival and service times was the exponential distribution. In these queueing systems, which are referred to as birth and death processes, the transitions from state to state are only allowed between adjacent states. In everyday life though, the need for queueing models with phase-type laws is crucial. For example, there are many real time queueing systems whose services need to be completed in many phases such as inventory and distribution systems. The most common distributions used in this kind of systems are the Erlang and hypergeometric distributions. Queueing systems with phase-type service or arrival mechanisms are known as *quasi-birth-death* QBD processes. Solution on these systems used to be given by ζ-transform but now Neut's method is used.

The E_r/M/1 queue refers to a single server queue whose service time follows an exponential distribution with rate μ and the arrival process follows an Erlang-r distribution. An illustration of how an Erlang-r arrival model works is shown below:

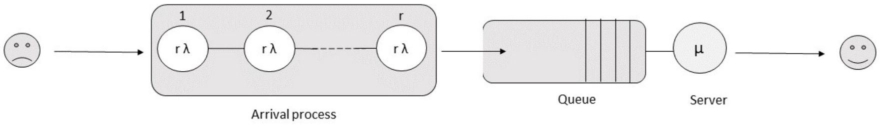

Figure 6.4: The E_r/M/1 queue

The density functions of the arrival and service processes are given by:

$$a(t) = \frac{(r\lambda)^r t^{r-1} e^{-r\lambda t}}{(r-1)!},\ t \geq 0$$
$$b(x) = \mu e^{-\mu x},\ x \geq 0$$

When a customer arrives in an E_r/M/1 queue, there are certain r phases to be passed through, which follow an exponential distribution with parameter λ, until the next customer begins the left-most exponential phase. Thus, only one customer is able to enter the system at any given instant time and only after all the r-phases are completed can a new one begin the process. In order to accomplish a successful description of an Erlang-r arrival model, the k number of customers present in the system, the current phase i an arriving customer is found is important information. A state of the system is depicted by the pair (k,i), where k is the number of customers in the system and i is the current phase of service. The transition state diagram proposed by William J. Stewart [22, p.451] is ideal for depicting the transition rate matrix form of this kind of queues. The diagram shows how the phases of arrival are arranged according to number of customers in the system.

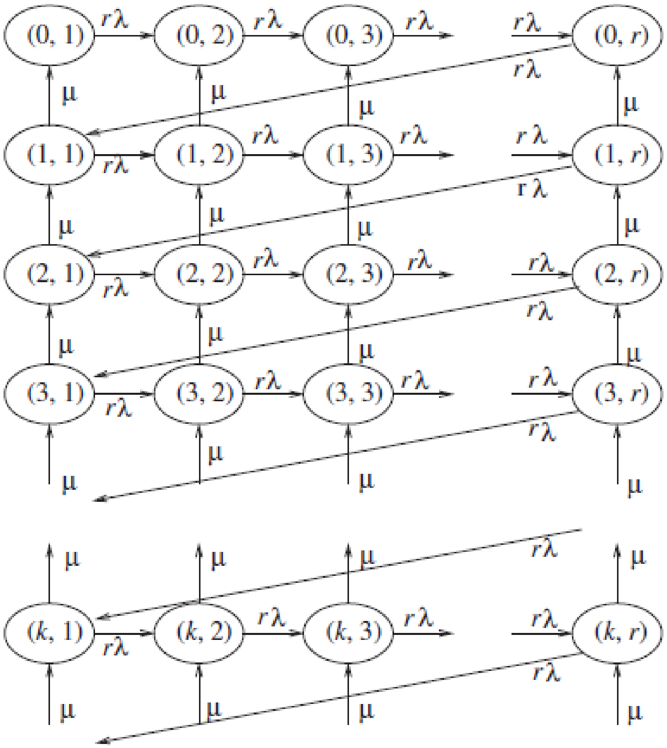

Figure 6.5: State transition diagram - E_r/M/1 queue

Based on the previous framework, the transition rate matrix will have a block-tridiagonal (or quasi birth-death) form, where all the subblocks are square and of order r:

$$Q = \begin{pmatrix} B_{00} & A_2 & 0 & 0 & 0 & 0 & \cdots \\ A_0 & A_1 & A_2 & 0 & 0 & 0 & \cdots \\ 0 & A_0 & A_1 & A_2 & 0 & 0 & \cdots \\ 0 & 0 & A_0 & A_1 & A_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \end{pmatrix}$$

Service completions at rate μ are represented by the matrices A_0 showing the transition from a state at some level k where i-1 arrival phases have already been completed to the state with the same number of completed arrival phases but with one customer less.

$$A_0 = \mu I$$

Arrivals to the system are represented by the matrices A_2. If an arrival completes successfully its last phase at rate $r\lambda$ then the arrival process begins immediately again but this time the number of customers in the system will be increased by 1. Consequently, the matrices A_2 will have one nonzero element $A_2(r, 1)$ with value $r\lambda$, representing the transition from state (k,r) to (k+1,1).

$$A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r\lambda & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

Furthermore, the superdiagonal elements of the matrices A_1 depict the completion of one arrival phase $i < r$ at rate $r\lambda$ and the beginning of the next one. In other words, these matrices represent the transition from state (k,i) to state (k,i+1). The only nonzero elements are the diagonal ones which are equal to the sum of the off-diagonal elements of Q.

$$A_1 = \begin{pmatrix} -\mu - r\lambda & r\lambda & 0 & 0 & \cdots & 0 \\ 0 & -\mu - r\lambda & r\lambda & 0 & \cdots & 0 \\ 0 & 0 & -\mu - r\lambda & r\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \vdots & r\lambda \\ 0 & 0 & 0 & 0 & \cdots & -\mu - r\lambda \end{pmatrix}$$

The diagonal elements of the matrix B_{00} are all equal to $-r\lambda$ and verify that the sum of elements across each row of Q is equal to zero.
Next step is the computation of the stationary probability vector π, with $\pi Q = 0$ using the matrix geometric method. This vector π is written as

$$\pi = (\pi_1, \pi_1, \pi_2, ..., \pi_k, ...)$$

where $k = 0, 1, 2, ..$ is a row vector of length r whose i^{th} component gives the probability of the arrival process having completed exactly i-1 phases, given there are k customers in the system waiting for service. The successive subvectors of π satisfy the relationship

$$\pi_{i+1} = \pi_i R \quad \text{for i =1,2,...}$$

where R, known as Neuts' rate matrix, is given by

$$R_{l+1} = -(V + R_l^2 W) \quad \text{l=0,1,2,...}$$

where $V = A_2 A_1^{-1}$, $W = A_0 A_1^{-1}$and $R_0 = 0$ is set as the initiative term. The sequence R_l is motone increasing and converges to R, according to Neut. The inverse of A_1 is an upper triangular matrix, whose non-zero ij elements are given by

$$(A_1^{-1})_{ij} = (-1)^{j-i} \frac{1}{d} (\frac{a}{d})^{j-i} \quad \text{for } i \leq j \leq r$$

where $d = -(\mu + r\lambda)$ and $a = r\lambda$. Also,

$$V_{ri} = -\left(\frac{r\lambda}{\mu + r\lambda} \right)^i \quad \text{for } 1 \leq i \leq r$$

and $V_{ki} = 0$ for $i \leq k < r$ and $1 \leq i \leq r$.
After that there is the calculation of the subvector π_0 by using:

$$\pi_{i+1} = \pi_i R = \pi_0 R^{i+1} \quad \text{for } i = 0, 1, 2, ..$$

and from $\pi Q = 0$ it is concluded that

$$(\pi_0, \pi_1, \pi_2, ..., \pi_i, ...) \begin{pmatrix} B_{00} & A_2 & 0 & 0 & 0 & 0 & \cdots \\ A_0 & A_1 & A_2 & 0 & 0 & 0 & \cdots \\ 0 & A_0 & A_1 & A_2 & 0 & 0 & \cdots \\ 0 & 0 & A_0 & A_1 & A_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \end{pmatrix} = (0, 0, 0, ..., 0, ...)$$

As a result a new form is created:

$$\pi_0 B_{00} + \pi_1 A_0 = \pi_0 B_{00} + \pi_0 R A_0$$

with constraint

$$1 = \sum_{k=0}^{\infty} \pi_k e = \sum_{k=0}^{\infty} \pi_0 R^k e = \pi_0 (I - R)^{-1} e$$

The system is homogeneous without a unique solution. By taking $\pi_{01} = 1$ a specific solution is achieved, which is then normalized according to the given constraint.
An alternative approach of $\mathcal{E}(r)$/M/1 queue is proposed by M.P.Wiper in his article "Bayesian analysis of Er/M/1 and Er/M/c queues" [28].

Performance Measures The probability that there are k customers in the queuing system is given by the sum of components of the k^{th} subvector

$$P_k = \|\pi_k\|_1 = \left\|\pi_1 R^{k-1}\right\|_1$$

The probability the system is empty is given by $P_0 = \|\pi_0\|_1$ and in case of a busy system by $1 - P_0$, accordingly.

By using Neut's R matrix, the average number of customers in the queuing system is obtained as

$$\begin{aligned}
E[N] &= \sum_{k=1}^{\infty} k\|\pi_k\|_1 = \sum_{k=1}^{\infty} k\left\|\pi_1 R^{k-1}\right\|_1 = \left\|\pi_1 \sum_{k=1}^{\infty} \frac{d}{dR} R^k\right\| \\
&= \left\|\pi_1 \frac{d}{dR}\left(\sum_{k=1}^{\infty} R^k\right)\right\|_1 = \left\|\pi_1 \frac{d}{dR}((I-R)^{-1}-I)\right\|_1 \\
&= \left\|\pi_1 (I-R)^{-2}\right\|_1
\end{aligned}$$

Let $E[N_q]$ the mean number of customers in the waiting line then

$$E[N_q] = E[N] - \lambda/\mu$$

Furthermore, the average response time $E[R]$ is calculated by

$$E[R] = E[N]/\lambda$$

and the average time spent in waiting is given by

$$E[W_q] = E[N_q]/\lambda$$

Chapter 7

Application of queueing theory in small business – A Case study of Bougioukos Bakery

In order for small businesses to survive it is essential that the development of certain factors be applied. These factors concern a good strategy, principles and tactics. Of these three, the strategy and principles have to be carefully taken into consideration since they are really important so as the business can flourish. In case of violating or trying to run a business without taking the fundamental principles into account, this might lead to company failure [26]. Companies which follow the fundamental principles of business are certain to survive both in good and hard times.

Company competitiveness is determined by the strength and steadiness of the possessed competitive advantages [18]. Companies which are successful in cost and quality based competitions, are looking for ways that will provide them further competitive advantage. Time has turned into a strategic resource and as a consequence, its importance has become equivalent to the significance of money, productivity and innovation (Stalk, 1988). Consequently time has turned into a strategic resource.

Waiting in lines constitutes a part of the everyday life. Queues form when the demand for a service exceeds its supply [10]. For many patients or customers, waiting in lines or queuing is annoying [16] or negative experience [21]. The unpleasant experience of waiting in line may adversely affect a customer's experience with a particular firm. The way in which managers address the waiting line issue is critical to the long term success of their firms [4].

This case study comes to ascertain that difficulties such as providing enough capacity for sufficient service and refrain from unnecessary expenses can be controlled by using queuing models. Average arrival rate of customers, average service rate, system utilization factor, cost of service and the probability of a specific number of customers in the system can be calculated and used for achieving a better waiting line performance. The aim is to find balance between minimizing operation costs which derived from optimization of a queuing system and minimizing the cost of waiting of the customers.

7.1 Information about Bougioukos bakery

7.1.1 Operating time

Bougioukos bakery opens from 5:00 am to 21:00 pm, however business hours are from 7:00 am to 21:00 pm. This means that the bakery is open to customers 14 hours a day, all the week. The rest of the hours remaining are for preparation and cleaning. The peak hour of customers arrival is between 11:00 am to 14:00 pm.

7.1.2 Customer serving procedure

Nowadays due to the pandemic, bakery stores are obliged to receive a certain amount of customers inside the store according to the square meters of the store. As a result, customers have to queue so as to enter the bakery, especially during peak hours.

In the present case study, Bougioukos bakery is 60 square meters so it can accept up to 2 customers. The rest of the customers must wait in queue outside the store. The typical procedure of entering a bakery in order to buy bread, relevant baking and pastry goods and coffee is as follows:

- Customer arrives, places the order, receives the traded good and pays for the order at the cash desk.

- In case of buying a cup of coffee, the customer arrives, places and pays for the order at the cash desk. By using the receipt, the customer waits until the product is ready. The receipt is used as a proof to which customer is first, second, third etc. according to the time and number of the receipt each customer has received.

It has to be taken into consideration the fact that as far as buying a cup of coffee is concerned, the orders can be also placed over the phone. Consequently, when these customers arrive to collect their order, this means that they are not obliged to wait in the waiting line. Their order has already been completed, they just pay, receive their receipt and leave. The whole procedure reduces the queue size.

Another issue to be taken into account is that the bakery is situated on a central street, really close to a supermarket. This means that during peak hours, waiting lines can be really long since customers try to combine their shopping in the supermarket with buying pastry goods and bread.

The bakery staff consists of one cashier who is also in charge of the service and three employees who are actually the bakers. In some cases, when there is great demand of service, there is an extra employee helping with the service. Therefore, it is assumed that a single channel system is used and hence a M/M/1 queuing model is the one that describes the way the bakery operates. In other words, the customers visiting the bakery join the queue or in case of idle system they proceed directly to service.

The queuing system at Bougioukos bakery is illustrated as depicted below:

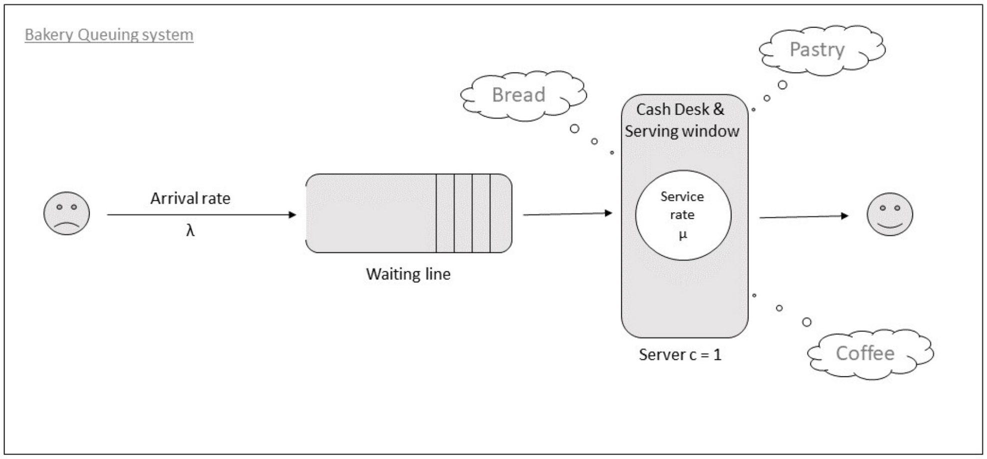

Figure 7.1: The queuing system of Bougioukos bakery

In this case study, customers have been observed coming in pairs or single, but due to Covid-19 restrictions, individual customers are the majority of bakery's daily clients. Arrivals occur randomly and independently and as a result an estimation of arrival occurrence is difficult to be determined. Therefore, the Poisson distribution constitutes the best way of describing an arrival pattern [1].

7.2 Objectives of the Case study

The main goals of this study are :

- To examine the operation of a business through Queuing theory
- To define in numbers the waiting line performance
- To make remarks on the efficiency of the bakery's organization flow according to the results of the study
- To enhance the management of queues in order to avoid customer loss and dissatisfaction

The outcome of the study can be used to increase efficiency and decrease cost in services. After all, the goal of every business is the fulfilment of the requests of each customer and the gaining of new customers through allusions.

Scope of the study and limitations
The data used were collected within a period of four weeks where arrival and departure times were recorded. It is important to mention that the survey was conducted over a short period of time and some aspects may considered to be under investigated.

7.3 Calculations

Weekly Customer Counts
By using the total daily receipts of the bakery a table of weekly customer counts is shown below: According to the survey, during week days the number of customers ranges between 80 to 120 customers per day while during weekends the range is between 200 to 220 customers.

Day Week	M	T	W	T	F	S	S
1^{st} Week	108	103	94	99	112	208	170
2^{nd} Week	102	109	92	89	107	214	183
3^{rd} Week	113	110	100	96	111	217	189
4^{th} Week	98	104	97	104	105	222	185

Table 7.1: Records from purchase receipts

Arriving customers per hour
The arrivals occur, taking into consideration that 75% of daily customers arrive between 11:00 am and 2:00 pm and supposing that they are uniformly distributed. according to the table below. Results from Table 7.1 are used for these calculations.

Day / Arrivals/hour	M	T	W	T	F	S	S	Average	
1st Week	20	19	17	18	21	39	31	23.57	
2nd Week	19	20	17	16	20	40	34	23.71	
3rd Week	21	20	18	18	20	40	35	24.57	
4th Week	18	19	18	19	19	41	34	24	**23.96**

Table 7.2: Records of arrivals per hour

Queue Length as per every twenty minutes
The table presented below shows how queue length is formed during rush hours. The data were received every twenty minutes.

Hour / Day	11:00 am			12:00 am			13:00 pm			14:00 pm			Average	
Minutes	20 Min	40 Min	60 Min	20 Min	40 Min	60 Min	20 Min	40 Min	60 Min	20 Min	40 Min	60 Min		
1st Day	2	3	3	5	6	5	6	4	4	6	5	4	4.4	
2nd Day	4	2	3	5	6	6	6	5	5	4	3	3	4.3	
3rd Day	3	4	2	6	6	4	5	4	5	3	2	2	3.8	**4.16**

Table 7.3: Queue length per twenty minutes during rush hours

According to the data, the bakery is the busiest between 12:00am to 13:00 pm and as a result the queue length is the longest. There is also the necessity to mention that some arrivals may occur as people in groups and consequently a longer queue will be formed.

Service times
By observing the service time completions of six customers in a row the results presented below were obtained:

Customer No. Time (min)	1^{st} Cust.	2^{nd} Cust.	3^{rd} Cust	4^{th} Cust	5^{th} Cust.	6^{th} Cust.	Average
Completed service time	2	1.5	2.5	2	1.5	1	**1.75**

Table 7.4: Records of services per minutes

7.4 Implementation of M/M/1 queuing model

For this study customer arrivals are considered random and independent and hence, a Poisson distribution is assumed. Furthermore, First Come First Served (FIFO) scheduling discipline is assumed, let alone that in real life some customers refuse to follow the queue, due to age or personal temperament. Infinite population is also assumed.
The form of this queuing model is:

$$M/M/C : FIFO(\text{or } FCFS)/\infty/\infty$$

where :
M stands for Markovian or Poisson arrival and exponential service time,
C stands for multi-server,
FIFO for First In First Out,
∞ indicates infinite system limit and
∞ infinite source limit.

Taking into consideration the results of the survey it is concluded that :
Number of servers: c = 1
Arrival rate: $\lambda = 23$ customers per hour
Serving rate: $\mu = 34$ customers per hour

Load factor: $\rho = \lambda/\mu = 67.64\%$
This factor shows that the serving rate is higher than the arrivals one. Consequently, it is assumed that there is a bit of efficiency in this queuing system.
The probability the channel is idle: $P_0 = 1 - \rho = 0.3236$
Average number of customers in the system: $L = \lambda W = 2.07$ per hour
Average number of customers in the queue: $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = 1.41$ per hour
Time spent in the waiting line: $W_q = E(W_q) = \frac{\lambda}{\mu(\mu-\lambda)} = 0.0614$ hours or 3.684 minutes

Total time spent in the queuing system: $W = E(W) = \frac{1}{\mu - \lambda} = 0.09$ hours or 5.4 minutes

According to the table presented below, the probability of customers in the waiting line decreases while the number of customers in the system increases. It is also deducted that the cumulative probability is quickly approaching 1, as for 10 customers $P_{10} = 0.9861$, implying that it is rare to have more than 10 customers in the queue.

N	Probability $P(n)$	Cumulative probability $P(n)$
0	0.3236	0.3236
1	0.2188	0.5424
2	0.1481	0.6905
3	0.1	0.7905
4	0.0677	0.8582
5	0.0458	0.904
6	0.0309	0.9349
7	0.0209	0.9558
8	0.0142	0.97
9	0.0096	0.9796
10	0.0065	0.9861

Table 7.5: Table of Probabilities

7.5 Implementation for $\mathcal{E}_r$/M/1 queuing model

As mentioned before, due to the pandemic, all businesses are obliged to follow certain rules as far as the customers are concerned. A specific number of customers is allowed inside the shop and consequently, a queue of customers

is formed outside the shop. This process could be considered to have arrival phases. For instance, a customer arrives at the bakery at half past eleven. Since it is peak hour, all customers have to wait in queue outside the bakery until the queue inside the shop decreases and is below two customers. This is arrival phase 1. Once one customer enters the bakery, the waiting line has to be followed until the bakery assistant serves the customer. This is arrival phase two.
The arrival process, in this case, follows an Erlang-r distribution and the service time follows an exponential one with parameter μ. Furthermore, First Come First Served (FIFO) scheduling discipline and infinite population are assumed.
The form of this queuing model is:

$$E_r/M/1 : FIFO(\text{or } FCFS)/\infty/\infty$$

where :
E_r stands for r-phase arrival process
M for exponential service time,
C stands for multi-server,
FIFO for First In First Out,
∞ indicates infinite system limit and
∞ infinite source limit.

The parameters of the queuing system according to the case study are $\lambda = 23$, $\mu = 34$ and $r = 2$. The occupation rate is calculated by:

$$\rho = \frac{\lambda}{c\mu} = 0.6764 \quad \text{or } 67.64\%$$

According to section 6.4, the submatrices have the following form:

$$A_0 = \begin{pmatrix} 34 & 0 \\ 0 & 34 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -80 & 46 \\ 0 & -80 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 46 & 0 \end{pmatrix},$$

$$B_{00} = \begin{pmatrix} -34 & 34 \\ 0 & -34 \end{pmatrix}$$

Based on the submatrices, then the form of infinitesimal generator is:

$$Q = \left(\begin{array}{cc|cc|cc|c} -34 & 34 & 0 & 0 & 0 & 0 & \cdots \\ 0 & -34 & 46 & 0 & 0 & 0 & \\ \hline 34 & 0 & -80 & 46 & 0 & 0 & \cdots \\ 0 & 34 & 0 & -80 & 34 & 0 & \\ \hline 0 & 0 & 34 & 0 & -80 & -46 & \cdots \\ 0 & 0 & 0 & 34 & 0 & -80 & \\ \hline \multicolumn{2}{c|}{\vdots} & \multicolumn{2}{c|}{\vdots} & \multicolumn{2}{c|}{\vdots} & \ddots \end{array}\right)$$

Next step is the computation of the stationary probability vector π, with $\pi Q = 0$. This vector π is written as

$$\pi = (\pi_1, \pi_1, \pi_2, ..., \pi_k, ...)$$

where $k = 0, 1, 2, ..$ is a row vector of length r whose i^{th} component gives the probability of the arrival process having completed exactly r-1 phases, given there are k customers in the system waiting for service. The steps presented in chapter 6.4 are followed. The inverse of submatrix A_1 is

$$A_1^{-1} = \begin{pmatrix} -0.0125 & -0.0072 \\ 0 & -0.0125 \end{pmatrix}$$

hence

$$W = A_0 A_1^{-1} = \begin{pmatrix} -0.425 & -0.2443 \\ 0 & -0.425 \end{pmatrix}$$

and

$$V = A_2 A_1^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ -0.575 & -0.3306 & \end{pmatrix}$$

Given $R_0 = 0$, Neuts' rate matrix $R_{l+1} = -(V + R_l^2 W) \quad$ l=0,1,2,...

$$R_1 = \begin{pmatrix} 0 & 0 \\ 0.575 & 0.3306 \end{pmatrix}, \quad R_2 = \begin{pmatrix} 0 & 0 \\ 0.6558 & 0.4235 \end{pmatrix}$$

and with help of python code it is concluded that R_{l+1} converges into R. Following the suitable calculations, it is considered that $R \simeq R_{50}$

$$R_{50} = \begin{pmatrix} 0 & 0 \\ 0.7661 & 0.5868 \end{pmatrix} \simeq R$$

The next step is the calculation of the subvector π_0 by using:

$$\pi_{i+1} = \pi_i R = \pi_0 R^{i+1} \quad \text{for } i = 0, 1, 2, ..$$

and from $\pi Q = 0$ it is concluded that

$$\pi_0 B_{00} + \pi_1 A_0 = \pi_0 B_{00} + \pi_0 R A_0$$

with constraint

$$1 = \sum_{k=0}^{\infty} \pi_k e = \sum_{k=0}^{\infty} \pi_0 R^k e = \pi_0 (I - R)^{-1} e$$

The system is homogeneous without a unique solution so by taking $\pi_{01} = 1$ a specific solution is achieved:

$$\pi_0 (B_{00} + RA_0) = (\pi_{01}, \pi_{02}) \begin{pmatrix} -34 & 1 \\ 26.0464 & 0 \end{pmatrix} = (0, 1)$$

The solution is $\pi_0 = (1, 1.3054)$. According to the restriction

$$(1, 1.3054) \begin{pmatrix} 1 & 0 \\ 1.8543 & 2.4205 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 6.5802$$

By dividing each component of π_0 by 7.0442 the normalized answer is:

$$\pi_0 = (0.1519, 0.1984)$$

The remaining subvectors are calculated by using $\pi_k = \pi_{k-1}R = \pi_0 R^k$

$$\begin{aligned}
\pi_1 &= \pi_0 R = (0.1519, 0.1164)\\
\pi_2 &= \pi_1 R = (0.0891, 0.0683)\\
\pi_3 &= \pi_2 R = (0.0523, 0.04)\\
\pi_4 &= \pi_3 R = (0.0307, 0.0235)\\
&\dots
\end{aligned}$$

The probability of having 0,1,2,... customers in the system is given by the sum of the previous subvectors:

$$p_0 = 0.3503, \quad p_1 = 0.2683, \quad p_2 = 0.1574, \quad p_3 = 0.1628, \quad ...$$

the average number of customers in the queuing system

$$E[N] = \left\|\pi_1 (I - R)^{-2}\right\|_1 = 1.5725$$

Let $E[N_q]$ the mean number of customers in the waiting line then

$$E[N_q] = E[N] - \lambda/\mu = 0.8960$$

Furthermore, the average response time $E[R]$ is calculated by

$$E[R] = E[N]/\lambda = 0.0683$$

and the average time spent in waiting is given by

$$E[W_q] = E[N_q]/\lambda = 0.0389$$

7.6 Conclusions

This case study constitutes a comparative application of two queuing systems, using data collected within a four week period of time. The subject of the study has been a queuing system of a bakery and the goal of the study is to determine the performance of this special system through queuing theory. The system's performance is outlined by calculating the average arrival and service rate of customers, the system utilization factor and the probability a specific number of customers being present in the system during a period of time.

The bakery's system is first described as an $M/M/1$ queuing model since there is one channel of service, the cashier and exponential arrival and service times are assumed. Then another queuing model, the $E_r/M/1$ queuing model is also assumed ideal for modeling this specific case study. Due to the pandemic nowadays, businesses are obliged to adopt arrival phases, in order to avoid customer crowding. As long as bakery's queuing system is considered, the model is constructed assuming there are two arrival phases. The first arrival phase represents the waiting line outside the store until the number of customers inside the store are lower than the customer limit (2 customers) inside and the second arrival phase represents the queue inside the store until the customer is served.

Based on the previous results, a comparative table of the two queuing systems is presented below:

Parameters	ρ	P_0	L	L_q	W	W_q
M/M/1	67.64%	0.3236	2.07	1.41	0.0614	0.09
M/E_r/1	67.64%	0.3503	1.5725	0.8960	0.0389	0.0683

Table 7.6: Comparative table between M/M/1 and $E_r/M/1$ queuing model

It is assumed that the Erlang arrival model is more efficient than the M/M/1 queue. Consequently, it is concluded that the new terms of our every day life have ameliorated the performance of this queuing system.

Chapter 8

Final Conclusions and Further Work

The main goal of this thesis is to represent the most essential parts of Queuing theory and the basic knowledge of Markov chains as used in favour of queuing modeling. In Chapter 1, an introduction to Operations Research (OR) is presented as both the origin and the definition of this special scientific field are illustrated. The definition of the basic steps of a successful implementation of OR is also presented. In Chapter 2, the basic terms and concepts of Markov chains are introduced as a great tool of OR. Markov Chains are ideal for defining how a process will evolve in the future taking into consideration only the present state of the process and by ignoring the past events. Chapters 3 and 4 contain the basic clues of Queuing theory. A great insight of the basics of the structure and construction of a queuing model is illustrated extensively, such as arrivals and service processes and performance measures. Next, in Chapter 5 remarks are found on Birth and Death process as one of the basic terms concerning queuing theory and then on Chapter 6 some basic queuing models are presented. Finally, in Chapter 7, a case study of a bakery's queuing system is illustrated and two different queuing models are applied and then compared for efficiency.

Due to the pandemic, aspects of everyday life have changed dramatically according to the restrictions humanity has to comply with. Businesses of all kinds of forms have to adapt more effectively to different kind of queuing systems than the ones they were used to. Queuing models with arrival phases are about to become more and more popular even in small businesses and as a result, further improvements in this special queuing models should be made. Simulation algorithms such as Discrete Event Simulation (DES) or Metropolis-Hastings algorithm could be adapted and used for improving the performance of various queuing systems, especially the phase type ones. By achieving to predict efficiently the performance of a system, the businesses will have the opportunity to prepare against challenging situations and improve even more their management and operations system. It is an indisputable fact that the enhancement of businesses operation management will lead to an elevating satisfaction rate of customers service which is, after all, the number one goal of every organization.

Appendices

Appendix A

Python

Listing A.1: Python code

```
import numpy
from numpy.linalg import matrix_power
c=1
lmd = 23
mu = 34
r = 2  # r-phases define the matrix dimensions
#Check stability
rho = lmd/mu
if rho >= 0.9999:
    print('ERROR:␣System␣is␣unstable')
else:
    print('System␣is␣stable!')
# submatrices for Er/m/1
A0 = numpy.array([[mu, 0],[0,mu]])
A1 = numpy.array([[-mu - r*lmd, r*lmd],[0,-mu - r*lmd]])
A2 = numpy.array([[0, 0],[r*lmd,0]])
B00 = numpy.array([[-mu, mu],[0,-mu]])
# inverse matrix of A1
A11 = numpy.linalg.inv(A1)
# Neuts' R Matrix
W = numpy.dot(A0, A11)
V = numpy.dot(A2, A11)
```

```
Ra = numpy.zeros((r,r))
R = -(V+numpy.dot(matrix_power(Ra, 2), W))
i=1
for i in range(1,51,1):
        Ra = R
        R = -(V+numpy.dot(matrix_power(Ra, 2), W))
        i = i+1
print('R=',R)
#Boundary equations, subvector pi
N = (B00 + numpy.dot(R,A0))
## first component equal to p01=1
N[0,r-1] = 1
N[1,r-1] = 0
print('N=', N)
## un-normalized pi0,pi1
res = numpy.zeros((r-1,r))
res[0,r-1] = 1
print(res)
unnormpi = numpy.dot(res,numpy.linalg.inv(N))
print('un',unnormpi)
## normalized pi0,pi1
e = numpy.ones((r,1))
lp = numpy.linalg.inv((numpy.identity(r))-R)
sol = numpy.dot(unnormpi,lp)
sol1 = numpy.dot(sol,e) #divide p0 by sol1
#results of pi
p0 = unnormpi/sol1
p1 = numpy.dot(p0,R)
pi = numpy.dot(p0,matrix_power(R,1))
k=1
for k in range(1,5,1): #alter range to find exact pi
    p0 = pi
    pi = numpy.dot(p0,matrix_power(R,1))
    k=k+1
print('p0',p0,'p1',p1,'number␣to␣divide',sol1)
print('pi',pi)
#Measures of effectiveness
z1=numpy.linalg.inv(numpy.identity(r)-R)
```

```
z2=  matrix_power(z1,2)
En = numpy.linalg.norm(numpy.dot(p1,z2),1) #mean number of customers
     #in the system
Eq = En - lmd/mu #mean number of customers in the waiting line
Er = En/lmd #average response time
Ewq = Eq/lmd #average time in waiting line
print('mean␣number␣of␣customers␣in␣the␣system',En,
      'mean␣number␣of␣customers␣in␣the␣waiting␣line',Eq,
      'average␣response␣time',Er,
      'average␣time␣in␣waiting␣line',Ewq)
```

Bibliography

[1] Sidonia Cernea, Mihaela Jaradat, and Mohammad Jaradat. "Characteristics Of Waiting Line Models – The Indicators Of The Customer Flow Management Systems Efficiency". In: *Annales Universitatis Apulensis Series Oeconomica* 2 (Dec. 2010), pp. 13–13. DOI: 10.29302/oeconomica.2010.12.2.13.

[2] J Laurie Snell Charles M. Grinstead. *Introduction to Probability*. 2nd ed. American Mathematical Society, 1997. ISBN: 0-8218-0749-8.

[3] J. W. Cohen. *The Single Server Queue*. 2 Sub. Applied Mathematics and Mechanix. North-Holland, 1982.

[4] Mark Davis, Nicholas Aquilano, and Richard Chase. *Fundamentals of Operations Management*. Jan. 2003.

[5] A.K. Erlang. "The theory of probabilities and telephone conversations". In: *Nyt Tidsskrift for Matematik B* 20 (Jan. 1909), pp. 33–39.

[6] Jacques Resing Ivo Adan. "Queueing Systems". In: 2015.

[7] William S. Jewell. *A simple proof of $L = \lambda W$*. University of California - Berkeley, 1966.

[8] P.W. Jones and P. Smith. *Stochastic Processes: An Introduction, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2017. ISBN: 9781498778121.

[9] Okoro Otonritse Joshua. "On Markovian Queueing Model as Birth-Death Process". In: *Double Blind Peer Reviewed International Research Journal* 13.1 (2013).

[10] Cagin Kandemir-Cavas and Levent Cavas. “An Application of Queueing Theory to the Relationship Between Insulin Level and Number of Insulin Receptors”. In: *Türk Biyokimya Dergisi* 32 (Jan. 2007), pp. 32–38.

[11] Samuel Karlin and James McGregor. “Many server queueing processes with Poisson input and exponential service times.” In: *Pacific J. Math.* 8.1 (1958), pp. 87–118.

[12] Takis Konstantopoulos. *One Hundred Solved Exercises for the subject: Stochastic Processes I.* Accessed: 7-1-2020.

[13] Dieter Baum L. Breuer. *An introduction to queueing theory and matrix-analytic methods.* 1st ed. Springer, 2005. ISBN: 978-1-4020-3631-6.

[14] McGraw-Hill. *Introduction to Operations Research.* The McGraw-Hill Companies, 2001.

[15] Frode B. Nielsen. *Queueing Systems: modelling Analysis and Simulation.Research Report 259.* Department of Informatics, University of Oslo, 1998.

[16] J.K. Obamiro. ““Application of Queuing Model in Determining the Optimum number of Service Facility in Nigerian Hospitals””. M. Sc. Project submitted to Department of Business Administration, University of Ilorin, 2003.

[17] P.R. Parthasarathy, N. Selvaraju, and R.B. Lenin. “The numerical solution of a birth-death process arising in multimedia synchronization”. In: *Mathematical and Computer Modelling* 34.7 (2001), pp. 887–901. ISSN: 0895-7177. DOI: `https://doi.org/10.1016/S0895-7177(01)00107-8`. URL: `http://www.sciencedirect.com/science/article/pii/S0895717701001078`.

[18] Michael Porter. “The Five Competitive Forces That Shape Strategy”. In: *Harvard business review* 86 (Feb. 2008), pp. 78–93, 137.

[19] K. Prasad and B. Usha. “A comparison between M/M/1 and M/D/1 queuing models to vehicular traffic atKanyakumari district”. In: 2015.

[20] Sheldon M. Ross. *Introduction to probability models.* 9th ed. Academic Press, 2007. ISBN: 978-0-12-598062-3.

[21] R. Scotland. “Customer Service: A Waiting Game”. In: *Marketing* 11 (1991), pp. 1–3.

[22] William J. Stewart. *Probability, Markov chains, queues and simulation. The mathematical basis of performance modeling*. Princeton University Press, 2009.

[23] Coletsos I. Stogiannis D. *Introduction to Operations Research*. Kalamara Elli, 2017.

[24] Coletsos I. Stogiannis D. *Operations Research: Theory, Algorithms & Applications*. Symeon, 2021.

[25] H.A. Taha. *Operations Research An Introduction*. Pearson, 2017.

[26] Bonga Wellington Garikai. "An Empirical Analysis of the Queuing Theory and Customer Satisfaction: Application in Small and Medium Enterprises A Case Study of Croc Foods Restaurant". In: *SSRN Electronic Journal* (Jan. 2013). DOI: 10.2139/ssrn.2348304.

[27] Wayne L.(Wayne L. Winston) Winston. *Operations Research: Applications and Algorithms*. 4th ed. Duxbury Press, 2003. ISBN: 978-0534380588.

[28] M.P. Wiper. "Bayesian analysis of Er/M/1 and Er/M/c queues". In: *Journal of Statistical Planning and Inference* 69.1 (1998), pp. 65–79. ISSN: 0378-3758. DOI: https://doi.org/10.1016/S0378-3758(97)00124-9. URL: https://www.sciencedirect.com/science/article/pii/S0378375897001249.