People's Democratic Republic of Algeria

Ministry of Higher Education

And Scientific Research

UNIVERSITY OF ECHAHID HAMMA LAKHDAR

EL OUED

FACULTY OF EXACT SCIENCES

Master's Thesis

Domain:Probability and statistics

Sector:Mathematics

Specialty: Fundamental and applied mathematics

Title:

Queues and their Applications

Presented by:

Safa Naroura and Chaima Hamidani

Discussed by the jury:

Professor:	nedjet doudi	(President)
Professor:	Med el said touati ibrahim	(Supervisor)
Professor:	Dehda el bachir	(Examiner)

Academic year :2024-2025

بسم الله الرحمن الرحيم

شكر وتقدير

الحمد لله الذي بنعمته تتم الصالحات، وبفضله تُنال الغايات، وعلى توفيقه اعتمدت في كل خطوات هذا العمل العلمي.

أتقدّم بأسمى عبارات الشكر والامتنان إلى أستاذي المشرف الفاضل **الدكتور محمد السعيد تواتي إبراهيم**، الذي كان لمتابعته الدقيقة، وتوجيهاته العلمية القيّمة، الأثر البالغ في إنجاز هذه المذكرة. لقد كان مثالًا في التفاني، والدقة، والدعم الأكاديمي والإنساني، فله مني خالص التقدير والاحترام.

كما لا يفوتني أن أتوجّه بجزيل الشكر والعرفان إلى جميع أساتذتي الكرام الذين أسهموا في تكويني العلمي، وإلى كل من قدّم لي يد العون والمساندة طيلة مشواري الجامعي، سواء بكلمة، أو نصيحة، أو دعاء صادق.

أخص بالشكر زملائي وأصدقائي الذين كانوا خير سند، وشاركوني هذه الرحلة بما حملته من تحديات ونجاحات.

وإلى **أسرتي الغالية**، التي كانت دومًا الداعم الأول في حياتي... شكري وامتناني لا يوفيكم حقكم.

إهداء

إلى من غرسا فيّ حب العلم، وزرعا الأمل، وسهرا من أجلي الليالي...
إلى من كانت دعواتهما لي زادًا ونورًا في دربي...
إلى والديَّ العزيزين، أهدي هذه المذكرة عرفانًا وتقديرًا لما قدّمتماه لي من حب وتضحيات.

إلى **أمي الحنون**، التي كانت كلماتها بلسمًا، وصبرها مصدر إلهام.
إلى **والدي القدير**، الذي علّمني معنى المسؤولية والاعتماد على النفس.
إلى **إخوتي وأخواتي**، شكرًا لوجودكم الدائم الذي منحني القوة.
إلى **أصدقائي وزملائي**، رفقاء الدرب وسند اللحظات الحاسمة.

أهدي إليكم جميعًا هذه المذكرة ...**عربون وفاء ومحبة وامتنان.**

صلى الله عليه وسلم

Contents

General Introduction

Queueing theory is one of the most prominent branches of applied mathematics and statistics, due to its crucial role in analysing service systems characterized by the random arrival of units—such as people, requests, or taskmaster limited resources, whether a single server or a pool of servers. With the rapid development of industrial and service sectors, the need has arisen for precise quantitative tools to understand and evaluate such systems, with the aim of reducing waiting times, improving resource utilization, and achieving operational efficiency [17].

The foundations of this theory began to take shape at the beginning of the 20^{th} century, with the Danish engineer **Agner Krarup Erlang** (1878–1929), who is considered the actual founder of this scientific field. In 1909, he conducted studies on congestion problems in telephone lines in Copenhagen, where he developed the first mathematical model to analyze the probability of call rejection due to saturated lines [10]. Through this, he laid the groundwork for queueing theory by introducing concepts of probability and time distributions into service modeling.

Subsequent scientific contributions deepened this field and expanded its range of applications. Notably, **David G. Kendall** (1918–2007) introduced, in the 1950s, the well-known *Kendall's notation* (A/B/s/N) for classifying queueing systems [23]. Additionally, **John D. C. Little** formulated the famous *Little's Law*:

$$L = \lambda W$$

which provides a fundamental relationship between the average number of customers in the system (L), the arrival rate (λ), and the average time spent in the system (W) [29]. Moreover, **Leonard Kleinrock** was one of the first researchers to apply queueing theory to computer networks and played a key role in the development of the internet [25].

This theory is built on a set of fundamental mathematical concepts, such as stochastic processes, Markov chains, Poisson processes, and service time distributions [34]. It also relies on key performance indicators that quantify the effectiveness of service systems, including average waiting time, number of customers in the queue, server utilization rate, and probability of rejection. Due to its flexible modelling nature, queueing theory is applied across many vital fields, including:

- **Healthcare systems**: managing emergency rooms and patient prioritization.
- **Communication and data networks**: data flow control and resource allocation.
- **Call centers**: improving service levels and reducing response times.
- **Manufacturing industries**: production planning and quality control.
- **Cloud computing**: task and resource distribution.

- **Traffic and transportation systems**.

In light of its importance, this thesis aims to explore both the theoretical and practical aspects of queueing models through three main chapters: The structure of this thesis is organized into three main chapters.

- **Chapter 1: Preliminaries**
  Addresses fundamental concepts related to stochastic processes, Markov chains, Poisson processes, and birth-death processes, which form the basis of queueing models.

- **Chapter 2: Classical Queueing Systems**
  Focuses on classical queueing systems, particularly the M/M/1 model, in addition to multi-server models such as M/M/c and M/M/∞, with an in-depth study of performance metrics and the application of Little's Law.

- **Chapter 3: Priority Queueing Models Among Customer Classes**
  Analyzes priority queueing systems involving multiple customer classes, both preemptive and non-preemptive, and presents the derivation of performance measures, supported by numerical analysis and real-world applications across various sectors.

Chapter 1

Preliminaries

1.1 Introduction

In this chapter, we review the concept of stochastic processes, focusing on discrete-time Markov processes defined over a finite or countable state space E. We then explore continuous-time Markov processes, with particular emphasis on Poison processes and birth-death processes, which we introduce and analyse. The concepts presented in this chapter are based on the references [5, 33].

1.2 Stochastic Processes

1.2.1 Definition and Notation

Definition 1.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $T \subset \mathbb{R}^+$ a time (or parameter) space, and $E \subset \mathbb{R}$ a state space. A stochastic process is a measurable mapping:*

$$X : \Omega \times T \to E, \quad (\omega, t) \mapsto X(t, \omega) = X_t(\omega), \tag{1.1}$$

where, for a fixed t_0, the function $\omega \mapsto X_{t_0}(\omega)$ is a random variable, and for a fixed ω_0, the function $t \mapsto X(t, \omega_0)$ is a measurable mapping called a trajectory or sample path.
*Stochastic processes are widely used in probability theory and various applications such as queueing systems, finance, and physics [**?**, **?**, 34].*

1.3 Stationary Processes

1.3.1 Strongly Stationary Process

Definition 1.2. *A stochastic process $(X_t)_{t\in T}$ is said to be strongly stationary (strictly stationary) if its finite-dimensional distributions are invariant under translation. That is, for all $h > 0$ and for all $t_1, t_2, \ldots, t_n \in T$, with $n \in \mathbb{N}^*$, the vector*

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$$

has the same distribution as the vector

$$(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}).$$

In particular, X_t and X_{t+h} share the same distribution for all $t, h \in T$.
This concept is fundamental in the study of stochastic processes and time series analysis [6, 9, 19].

1.3.2 Weakly Stationary Process

Definition 1.3 (Weak Stationarity [6,8,19]). *A stochastic process $(X_t)_{t\in T}$ is said to be* weakly stationary *(or stationary in the wide sense) if the following conditions hold for all $t, s \in T$:*

1. *The expected value is constant:*
$$\mathbb{E}(X_t) := m_t = m < +\infty.$$

2. *The second moment is finite and constant:*
$$\mathbb{E}(X_t^2) := \delta_t^2 = \delta^2 < +\infty.$$

3. *The covariance function depends only on the time difference:*
$$Cov(X_t, X_s) := \mathbb{E}[(X_t - m)(X_s - m)] = \Gamma(t, s) = \Gamma(t - s),$$
where Γ is the covariance function of the process.

1.3.3 Processes with Independent and Stationary Increments (PISI)

Independent Increment Processes (IIP)

Definition 1.4. *A stochastic process $(X_t)_{t\in T}$ is said to have* ***independent increments*** *(IIP) if, for all $t_0, t_1, \dots, t_n \in T$ and any integer n, the random variables*
$$X_{t_n} - X_{t_{n-1}},\ X_{t_{n-1}} - X_{t_{n-2}},\ \dots,\ X_{t_2} - X_{t_1},\ X_{t_1} - X_{t_0}$$
are independent.
This property is fundamental in the study of stochastic processes such as the Wiener process and the Poisson process [?].

Stationary Increment Processes (SIP)

Definition 1.5. *A stochastic process $(X_t)_{t\in T}$ is said to have* ***stationary increments*** *(SIP) if, for all $t, s \in T$, the law of the increment $X_t - X_s$ depends only on the difference $t - s$. That is, the distribution of $X_t - X_s$ is the same as that of X_{t-s}*
Processes with this property are time-homogeneous in terms of their increments, which is a common characteristic in models of physical and financial systems [8].

Independent and Stationary Increment Processes (PISI)

Definition 1.6. *A stochastic process $(X_t)_{t\in T}$ is said to have* ***independent and stationary increments*** *(PISI) if it satisfies both the independent increment property (IIP) and the stationary increment property (SIP).*
Examples of such processes include Lévy processes, which form a rich class of stochastic models [35].

Example 1.1. 1. Poisson Process

The Poisson process $(N_t)_{t\geq 0}$ is a counting process where:

- $N_0 = 0$.

- *It has independent increments: the number of events in disjoint time intervals are independent.*
- *It has stationary increments: the number of events in an interval of length h is Poisson distributed with parameter λh, independent of the starting time.*

Therefore, the Poisson process is a classic example of a process with independent and stationary increments (PISI) [7].

2. Wiener Process (Brownian Motion)

The Wiener process $(B_t)_{t\geq 0}$ satisfies:

- $B_0 = 0$ *almost surely.*
- $B_t - B_s \sim \mathcal{N}(0, t-s)$ *for* $0 \leq s < t$.
- *It has independent increments: non-overlapping increments are independent.*
- *It has stationary increments: the distribution of increments depends only on the time difference.*

Hence, the Wiener process is another standard example of a PISI process [20].

1.4 Markov Chains

The Markov process provides a simple modelling tool for a particular class of systems with a discrete state space. A key characteristic of a Markov process is the "memoryless" property, which states that the future evolution of the process depends only on the present state and not on the sequence of events that preceded it.
Markov chains are widely used in modelling and analysing systems that evolve randomly over time, especially in the context of discrete events. The analysis of Markov processes forms a necessary prerequisite for the study of queuing systems, where arrivals, services, and departures can be described probabilistically using transition probabilities and state dynamics

A state j is transient if and only if:

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

In this case, we also have:

$$\sum_{n=0}^{+\infty} P_{ij}^{(n)} < +\infty, \quad \forall i \in E.$$

The states in the same communication class are either all recurrent or all transient [?].

1.4.1 Continuous-Time Markov Chains

Definition 1.7. *A stochastic process $(X_t)_{t\geq 0}$ is called a continuous-time Markov chain (CTMC) if, for all $0 \leq t_0 < t_1 < \cdots < t_n < t$, and for all $j, i_0, i_1, \ldots, i_n \in E$ with $n \in \mathbb{N}$,*

$$P(X_t = j \mid X_{t_0} = i_0, \ldots, X_{t_n} = i_n) = P(X_t = j \mid X_{t_n} = i_n).$$

Furthermore, the chain is said to be homogeneous if, for all $t \in \mathbb{R}^+$ and $i, j \in E$,

$$P(X_{t+s} = j \mid X_s = i) = P(X_t = j \mid X_0 = i) = p_{ij}(t).$$

Transition Matrix

Definition 1.8. *The transition matrix $P(t)$ of the process, whose general term is given by:*

$$p_{ij}(t) = P(X_t = j \mid X_0 = i),$$

is a stochastic matrix that satisfies:

$$\forall i, j \in E, \quad p_{ij}(t) \geq 0, \quad \text{and} \quad \sum_{j \in E} p_{ij}(t) = 1, \quad \forall i \in E.$$

The transition probabilities of a continuous-time Markov chain satisfy the Chapman-Kolmogorov equations. For all $i, j \in E$ and for all non-negative real numbers s, t, we have:

$$p_{ij}(t+s) = \sum_{k \in E} p_{ik}(t) p_{kj}(s).$$

Continuous-Time Markov Chains (CTMCs)

Proposition 1.1. ***Infinitesimal Generator Matrix*** *We associate with a continuous-time Markov chain (CTMC) the matrix Q, called the infinitesimal generator, defined as follows:*

$$\forall i \neq j, \quad q_{ij} = \mu_{ij} = \mu_i P_{ij}, \tag{1.2}$$

$$q_{ii} = -\sum_{j \neq i} \mu_{ij} = -\sum_{j \neq i} \mu_i P_{ij} = -\mu_i. \tag{1.3}$$

Thus, the generator matrix Q takes the form:

$$Q = \begin{bmatrix} -\sum_{j\neq 1} \mu_{1j} & \mu_{12} & \mu_{13} & \cdots & \mu_{1n} \\ \mu_{21} & -\sum_{j \neq 2} \mu_{2j} & \mu_{23} & \cdots & \mu_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mu_{n1} & \mu_{n2} & \mu_{n3} & \cdots & -\sum_{j \neq n} \mu_{nj} \end{bmatrix}.$$

It satisfies the property:

$$\sum_j q_{ij} = 0. \tag{1.4}$$

Proposition 1.2. ***Sojourn Time Distribution*** *For any state $i \in E$, the time spent in i before leaving follows an exponential distribution with parameter μ_i [32].*

Corollary 1.1. ***Transition Probability Matrix***
The transition matrix $P(t)$ of a CTMC can be expressed as:

$$P(t) = e^{tQ} = \sum_{n=0}^{+\infty} \frac{t^n}{n!} Q^n, \tag{1.5}$$

*where Q is the infinitesimal generator associated with the CTMC [**?**].*

1.4.2 Poisson Process

The Poisson process is a significant stochastic process in continuous time, frequently used to model the random occurrence of events of the same type.

Definition 1.9. ***Counting Process*** *A stochastic process $(N(t))_{t \in \mathbb{T}}$ is called a counting process if it satisfies the following:*

1. $N(t) \in \mathbb{N}, \quad \forall t \in \mathbb{R}^+$,
2. $\forall t > s \in \mathbb{R}^+, \quad N(t) \geq N(s)$,
3. $N(t) - N(s)$ *represents the number of events in the interval* $[s, t]$.

Definition 1.10. ***Poisson Process*** *A counting process* $(N(t))_{t\geq 0}$ *is called a Poisson process if it satisfies the following three conditions:*

C1 Time Homogeneity: *The probability of observing* k *events in an interval of given length depends only on* t*, not on the location of the interval along the time axis:*

$$P(N(t) = k) = P(N(s+t) - N(s) = k), \quad \forall s > 0, t > 0.$$

[?]

C2: Independent Increments: *The process* $N(t)$ *has* ***independent increments****, which means that for all* $s > 0$ *and* $t > 0$*, the number of events in disjoint time intervals are independent. Formally,*

$$P(N(s+t) - N(s) = k,\, N(s) = j) = P(N(s+t) - N(s) = k) \cdot P(N(s) = j).$$

This property implies that the number of arrivals in the interval $(s, s+t]$ *is independent of the history up to time* s *[27, 34].*

C3: Increment Probability: *Let* $p_k(\Delta t)$ *denote the probability that exactly* k *events occur during the small time interval* Δt*. Assuming the rate* $\lambda = \mathbb{E}[N(t+1) - N(t)]$*, then:*

$$p_k(\Delta t) = \begin{cases} 1 - (\lambda \Delta t + o(\Delta t)), & \text{if } k = 0, \\ \lambda \Delta t + o(\Delta t), & \text{if } k = 1, \\ o(\Delta t), & \text{if } k \geq 2, \end{cases}$$

where $o(\Delta t)$ *is a function such that* $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$*. This characterization is fundamental to the construction of the Poisson process [27, 34].*

Remark and Proposition on the Poisson Process

Remark 1.1. *The parameter* λ *is called the intensity of the Poisson process* $\{N(t), t \geq 0\}$*. It represents the expected number of events occurring in a unit time interval [34].*

Proposition 1.3. *Let* $\{N(t)\}_{t\geq 0}$ *be a Poisson process. Then there exists a constant* $\lambda > 0$ *such that, for all* $0 \leq s < t$*, the random variable* $N(t) - N(s)$ *follows a Poisson distribution with parameter* $\lambda(t - s)$*, i.e.,*

$$\mathbb{P}(N(t) - N(s) = k) = e^{-\lambda(t-s)} \frac{[\lambda(t-s)]^k}{k!}, \quad k \in \mathbb{N}. \tag{1.6}$$

This property is known as the stationary independent increments of the process [?].

Poisson Process and the Exponential Law

We characterize the Poisson process through its arrival times. Let A_n *denote the time of the* n*-th arrival:*

$$A_n = \inf\{t \geq 0 \,:\, N(t) = n\}. \tag{1.7}$$

Define T_n *as the* n *the interarrival time for* $n \in \mathbb{N}^*$*:*

$$T_n = A_n - A_{n-1}, \quad \textit{where } A_0 = 0. \tag{1.8}$$

Thus, the arrival time of the n *th event can be expressed as the sum of interarrival times:*

$$A_n = \sum_{i=1}^{n} T_i, \tag{1.9}$$

and the number of events up to time t *is given by:*

$$N(t) = \max\{n \geq 0 \,:\, A_n \leq t\}. \tag{1.10}$$

These properties follow from the standard construction of the Poisson process as a counting process with exponentially distributed interarrival times [13].

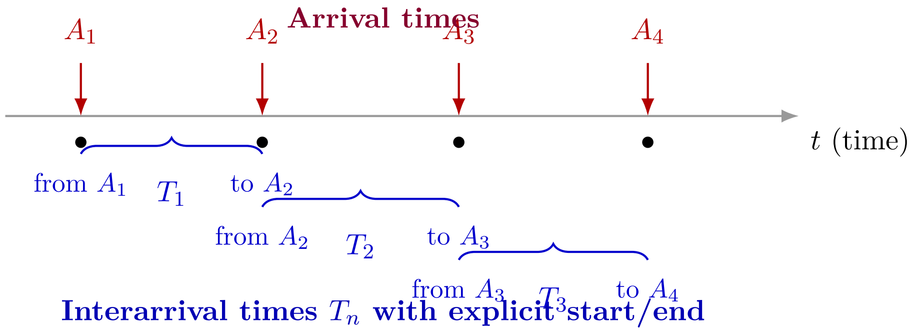


Figure 1.1: Arrival times A_n and interarrival times T_n in a Poisson process, with start/end of each T_n labeled

Characterization via Interarrival Times

A process $\{N_t\}_{t \in \mathbb{R}_+}$ *is a Poisson process with rate* λ *if and only if the interarrival times* T_n *are independent and identically distributed (i.i.d.) according to the exponential distribution* $\varepsilon(\lambda)$*, with the probability density function:*

$$f_{T_n}(t) = \lambda e^{-\lambda t} \mathbf{1}_{]0,+\infty[}(t), \tag{1.11}$$

*as explained in [34] and [**?**].*

Distribution of T_1

We show that T_1 *follows an exponential distribution:*

$$\mathbb{P}(T_1 > t) = \mathbb{P}(N_t = 0) = e^{-\lambda t} = 1 - F_1(t), \tag{1.12}$$

where F_1 *is the cumulative distribution function (CDF) of* T_1*. Hence,* $T_1 \sim \varepsilon(\lambda)$ *[36].*

Distribution of T_2

Using the independence of increments:

$$\begin{aligned} \mathbb{P}(T_2 > t \mid T_1 = t_1) &= \mathbb{P}(N_{t_1+t} - N_{t_1} = 0 \mid N_{t_1} = 1, N_s = 0 \textit{ for } s < t_1) \\ &= \mathbb{P}(N_{t_1+t} - N_{t_1} = 0) \\ &= \mathbb{P}(N_t = 0) = e^{-\lambda t}. \end{aligned} \tag{1.13}$$

Thus, T_2 *is independent of* T_1 *and follows an exponential distribution* $\varepsilon(\lambda)$ *[5].*

General Case by Induction

By induction, we show that for $k \geq 2$:

$$P([T_k > t] \mid [T_1 = t_1] \cap \ldots \cap [T_{k-1} = t_{k-1}]) = P(N_{t_{k-1}+1} - N_{t_{k-1}} = 0) \\ = P(N_t = 0) = e^{-\lambda t}. \quad (1.14)$$

Therefore, T_k is independent of $T_1, \ldots, T_{k-1}$ and follows an exponential distribution $\varepsilon(\lambda)$.

Transition Graph

In the case of continuous-time processes, particularly the Poisson process, a transition graph can be used to visualize the process's behavior through state transitions characterized by exponential holding times and memoryless properties [34].

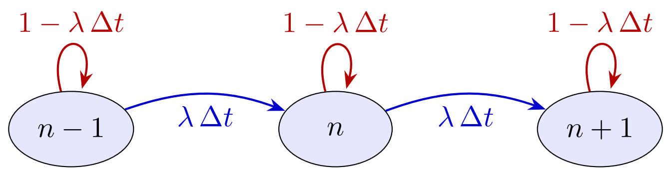


Figure 1.2: Colored transition graph of the Poisson process

Poisson Process and Transition Probabilities

In continuous-time Markov chains, the time unit is often replaced by a small interval Δt. Each arc is labeled with the corresponding transition probability, as shown in Figure ?? (typically a state transition diagram). To avoid overloading the diagram, the $+o(\Delta t)$ term, which appears in every transition probability, is usually omitted for simplicity [32].
*The Poisons process is a fundamental stochastic model used in queueing theory and reliability analysis. It is characterized by independent and exponentially distributed interarrival times, with a constant rate λ [34]. The memoryless property of the exponential distribution makes the Poisons process particularly useful in modeling arrivals in service systems and failures in reliability theory [**?**].*

1.4.3 Birth and Death Processes

Generalities

Birth and death processes are continuous-time stochastic processes with a discrete state space $n = 0, 1, 2, \ldots$. These processes are memoryless, meaning that the future evolution depends only on the current state, not on the past history [5]. From a given state n, transitions are possible only to the neighbouring states $n+1$ and $n-1$ (with $n \geq 1$). These transitions are commonly referred to as "births" and "deaths," respectively.
Such processes are widely used to model queueing systems, such as the number of customers in a service facility, and population dynamics, where individuals are born and die at certain rates [36].

Definition 1.11. *Let $\{X(t); t \geq 0\}$ be a time-homogeneous discrete-state stochastic process, meaning:*

$$P(X(t+s) = j \mid X(s) = i) = p_{ij}(t),$$

*which does not depend on s. The process $\{X(t); t \geq 0\}$ is a **birth and death process** if it satisfies the following conditions:*

$$\begin{cases} p_{i,i+1}(\Delta t) = \lambda_i \Delta t + o(\Delta t), & i \geq 0, \\ p_{i,i-1}(\Delta t) = \mu_i \Delta t + o(\Delta t), & i \geq 1, \\ p_{i,i}(\Delta t) = 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t), & i \geq 0, \\ p_{i,j}(\Delta t) = o(\Delta t), & |i-j| \geq 2, \end{cases}$$

where the positive coefficients λ_i and μ_i are called the transition rates, specifically the birth (or growth) rate for λ_i and the death (or decay) rate for μ_i [?, 34].

Transient Regime

*The **transient regime** of a Markov chain refers to the period during which the state probabilities are still evolving over time and have not yet reached the stable values of the steady-state regime. Specifically, it is the time interval t during which the probabilities $p_j(t)$ of being in state j at time t still depend on the initial distribution and vary with t.*
*The system of equations known as the **Chapman-Kolmogorov equations** associated with the process $(X_t)_{t \geq 0}$ is defined by the following system:*

$$(S) \quad \begin{cases} p_0'(t) = \mu_1 p_1(t) - \lambda_0 p_0(t), & n = 0, \\ p_n'(t) = \mu_{n+1} p_{n+1}(t) - (\lambda_n + \mu_n) p_n(t) + \lambda_{n-1} p_{n-1}(t), & n \geq 1, \\ \sum_{n \geq 0} p_n(t) = 1, \\ p_n(0) = P[X_0 = n] = \begin{cases} 1, & n = 0, \\ 0, & n > 0. \end{cases} \end{cases}$$

Remark 1.2. *To derive the system (S), the total probability theorem (see Appendix) is used. Since it is a system of differential and recurrence equations, its analytical solution involves orthogonal Bessel functions in the transient (non-stationary) case [2, 11].*

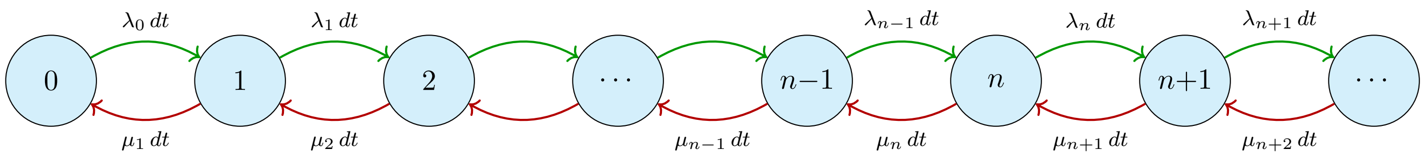


Figure 1.3: Transition graph with all states (including symbolic ones) inside circles

Steady-State Regime of Markov Chains

The steady-state regime of a Markov chain is reached when the transition probabilities between states stabilize at fixed values, independent of the initial distribution of states. In other words, as $t \to +\infty$, the limits:

$$p_n = \lim_{t \to \infty} p_n(t)$$

exist and are independent of the initial state of the process [34]. At steady state, we also have:

$$\lim_{t \to \infty} p_n'(t) = 0,$$

where p_n is a stationary distribution if the steady-state regime exists. This leads to the so-called balance equations [31].
The previous system of equations (S) becomes:

$$(S') \quad \begin{cases} \mu_1 p_1 - \lambda_0 p_0 = 0, & \text{for } n = 0, \\ \mu_n p_n - \lambda_{n-1} p_{n-1} = \mu_{n-1} p_{n-1} - \lambda_{n-2} p_{n-2}, & \text{for } n \geq 1, \\ \sum_{n \geq 0} p_n = 1. \end{cases}$$

Solving the system (S') [27]:
For $n = 0$:

$$p_1 = \frac{\lambda_0}{\mu_1} p_0,$$

For $n = 1$:

$$\mu_2 p_2 - \lambda_1 p_1 = \mu_1 p_1 - \lambda_0 p_0,$$

For $n = 2$:

$$\mu_3 p_3 - \lambda_2 p_2 = \mu_2 p_2 - \lambda_1 p_1.$$

By recurrence, we obtain:

$$p_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_1\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_2\mu_1} p_0.$$

Thus, we can write:

$$p_n = \left(\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right) p_0 = a_n p_0, \quad \forall n \geq 1,$$

where the sequence $(p_n)_{n\geq 0}$ represents a probability distribution. Since:

$$\sum_{n\geq 0} p_n = 1,$$

we have:

$$p_0 = \frac{1}{1 + \sum_{n\geq 1} a_n}.$$

Conclusion. *If $\sum_{n\geq 1} a_n < +\infty$, then:*

$$p_n = \frac{a_n}{1 + \sum_{n\geq 1} a_n}, \quad n \geq 1, \quad \text{and} \quad p_0 = \frac{1}{1 + \sum_{n\geq 1} a_n},$$

where:

$$a_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}.$$

This approach is fundamental in queuing theory and stochastic processes, particularly for birth-death processes [13].

Chapter 2

Queueing Systems

2.1 Problem Statement

Queueing systems are widespread in daily life, appearing in areas such as customer service, traffic, and telecommunications. Various stochastic models have been developed to analyse these systems, with the M/M/1 Markovitz model being the most common due to its memoryless nature [15]. Key questions in performance optimization include customer wait time, queue length, server idle time, and the service capacity required to avoid congestion [?]. Markov chains and stochastic processes offer reliable answers based on the model used. Thus, queueing theory serves as a vital tool for analysing and enhancing service systems, with model selection playing a crucial role in addressing real-world challenges [?].

2.2 The Formalism of Queueing Systems

Queueing theory aims to model and analyse various seemingly different situations. In these models, customers arrive at random intervals to a system with multiple servers where they request service. The service time at each server is also random. After being served, customers leave the system.

Definition 2.1. *A queueing system is the mathematical abstraction of a system that can be described by the following elements:*

1. *The arrival process of customers;*
2. *The source of customers;*
3. *The behavior of customers;*
4. *The service time distribution;*
5. *The service discipline;*
6. *The number of servers;*
7. *The queue capacity.*

- ***The Arrival Process of Customers:*** *It is often assumed that inter-arrival times are independently and identically distributed (i.i.d.). In many cases, the arrival process is modeled as Poissonian, meaning that the inter-arrival times follow an exponential distribution [15]. However, this assumption is not always realistic. Customers may arrive individually or in batches.*

- ***The Source of Customers:*** *In most models, the customer source is assumed to be infinite. However, in reliability models, the source is finite. For example, in a factory setting where a repairman services machines, the number of machines needing repair is limited [?].*
- ***Customer Behavior:*** *Some customers are patient and willing to wait for extended periods. Others become impatient and leave the system after some time. This is the case in telephone call centers where customers hang up if they wait too long before a line becomes available, only to try calling again later [?].*
- ***Service Time Distribution:*** *Service times are generally assumed to be independently and identically distributed (i.i.d.) and independent of the inter-arrival times. However, this is not always true. For instance, in a production system, machine processing time may increase significantly when the number of tasks becomes too large [15].*
- ***Service Discipline:*** *Customers can be served individually or in batches. Several service order disciplines are commonly considered:*
  - ***FIFO (First In, First Out):*** *Entities leave the system in the order they entered. This is the most commonly used discipline.*
  - ***LIFO (Last In, First Out):*** *The last entity to enter is the first to be served. This is typical in computer stacks.*
  - ***Random:*** *All entities have the same probability of being served first.*
- ***Number of Servers:*** *The number of servers may vary depending on the nature of the service. It can be a single server or multiple servers working simultaneously [?].*
- ***Queue Capacity:*** *In many cases, the queue is assumed to be infinite. However, in some situations, it is finite. Such cases arise in practical systems where space or resource limitations prevent indefinite queueing [?].*

2.3 Classification of Queueing Systems

To classify queueing systems, we use the symbolic notation introduced by Kendall in the early 1950s [23].

2.3.1 Kendall Notation

This notation consists of four symbols arranged in the order $A/B/s/N$, where:

- A*: The distribution of time intervals separating two consecutive customer arrivals in the system (arrival process).*
- B*: The distribution of the random time it takes for a customer to receive service (service time).*
- s*: The number of service stations available simultaneously.*
- N*: The total capacity of the system (maximum number of customers that can be present in the system at the same time).*

The last symbol is omitted if the system has infinite capacity (no customer will ever be refused). To specify the distributions A and B, the following symbols are used:

- M*: Markovian process (exponential distribution).*
- E_k*: Erlang distribution of order k.*

- H_k: *Hyperexponential distribution.*
- G: *General distribution.*
- D: *Deterministic case.*

We also introduce:

- λ: *The arrival rate (average number of arrivals per unit time).*
- $\frac{1}{\lambda}$: *The average time interval separating two consecutive arrivals.*
- μ: *The service rate (average number of customers served per unit time).*
- $\frac{1}{\mu}$: *The average service time duration.*

In the following, we focus on $M/M/1/(\infty)$ *systems, where:*

- $A = M$: *The distribution of intervals between two arrivals is exponential.*
- $B = M$: *The distribution of service times is exponential.*
- $s = 1$: *A single service station is available at a time.*
- $N = \infty$: *The system capacity is infinite (the queue can grow indefinitely).*

2.4 Queueing System M/M/1

2.4.1 Model Description

The M/M/1 queueing system is a fundamental model in queueing theory. It describes a system where customers arrive according to a Poisons process with rate λ *(average number of customers arriving per unit time). This implies that the time intervals between consecutive arrivals follow an exponential distribution with parameter* $\lambda > 0$. *Service is provided by a single server. Upon arrival, if the server is free, the customer is immediately served. Otherwise, the customer joins the queue. The queue capacity is unlimited (infinite number of positions and no other restrictions). The queue discipline is First-In-First-Out (FIFO). The service times follow an exponential distribution with parameter* $\mu > 0$. *Consequently, the service rate is* μ *(average number of customers served per unit time), and the average service time for a customer is* $\frac{1}{\mu}$. *The random variables representing the intervals between consecutive arrivals and the service times are mutually independent. We are interested in the evolution of the number of customers* $N(t)$ *present in the system at time* t $(t > 0)$. $N(t)$ *is a stochastic process, and its behavior depends on the system parameters [?, ?, ?].*

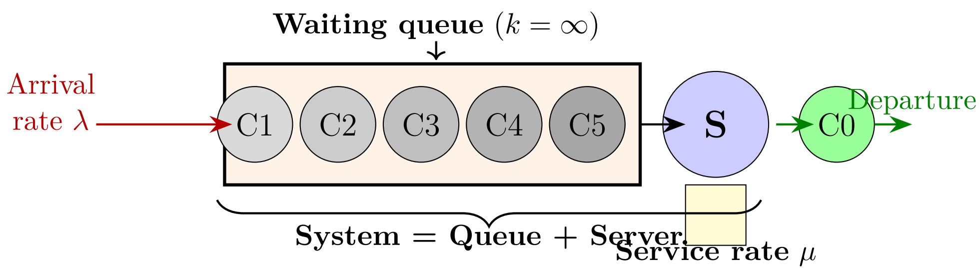


Figure 2.1: M/M/1 Queue with arriving, waiting, served, and departed customers.

2.5 General Approach

The mathematical study of a queueing system is most often conducted by introducing a stochastic process defined in an appropriate manner. In general, we are interested in the number $N(t)$ of customers in the system at time t ($t \geq 0$). Based on the quantities that define the system structure, we aim to calculate:

- *The state probabilities $p_n(t) = P(N(t) = n)$, which define the transient regime of the process. These probabilities $p_n(t)$ must obviously depend on the initial state or the initial distribution of the process. In the case of M/M/1 systems, the calculations involve Bessel functions, which are complex. Therefore, we will not calculate the transient regime.*
- *The stationary regime of the stochastic process is defined by:*

$$p_n = \lim_{t \to +\infty} p_n(t) = P(N(+\infty) = n), \quad n = 0, 1, 2, \ldots$$

From the stationary distribution of the process $\{N(t)\}_{t \geq 0}$, it is possible to obtain other operational characteristics of the system, such as:

- *The average number L of customers in the system.*
- *The average number L_q of customers in the queue.*
- *The average waiting time W_q of a customer.*
- *The average sojourn time W in the system (waiting + service).*
- *The utilization rate of the service stations.*
- *The percentage of customers who could not be served.*
- *The duration of an activity period, i.e., the time interval during which there is always at least one customer in the system.*

2.6 Transient Regime

Although the transient regime is difficult to calculate, thanks to the properties of the Poisson process and the exponential distribution, we obtain results necessary for calculating the stationary regime [?, ?].

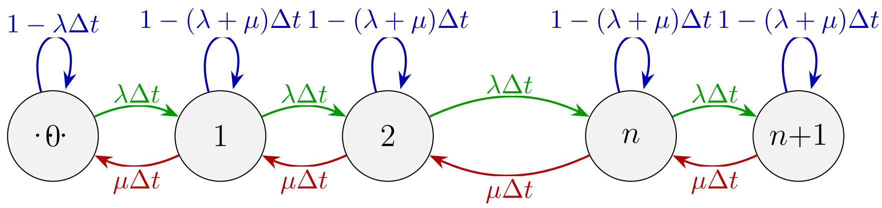


Figure 2.2: Transition Graph of the M/M/1 Queue (colored: arrivals in green, services in red, self-loops in blue)

Stationary Regime Focus

Due to the complexities involved in the transient regime, particularly the appearance of Bessel functions, we focus on the steady-state behavior of the M/M/1 queue [?, 38].
For a small time interval Δt, the following probabilities are given [?]:

- $\Pr(\textit{exactly one arrival during } \Delta t) = \lambda\Delta t + o(\Delta t)$
- $\Pr(\textit{no arrivals during } \Delta t) = 1 - \lambda\Delta t + o(\Delta t)$
- $\Pr(\textit{two or more arrivals during } \Delta t) = o(\Delta t)$
- $\Pr(\textit{exactly one departure during } \Delta t \mid N(t) > 0) = \mu\Delta t + o(\Delta t)$
- $\Pr(\textit{no departures during } \Delta t \mid N(t) > 0) = 1 - \mu\Delta t + o(\Delta t)$
- $\Pr(\textit{two or more departures during } \Delta t) = o(\Delta t)$

Let $p_{ij}(\Delta t) = \Pr(N(t + \Delta t) = j \mid N(t) = i)$, where $i = 0, 1, 2, \ldots$. These probabilities depend only on Δt, not on t. Because arrivals and departures are independent, $N(t)$ is a homogeneous Markov process [?].
We compute the transition probabilities as follows:

$$\begin{aligned}
p_{n,n+1}(\Delta t) &= (\lambda\Delta t + o(\Delta t))(1 - \mu\Delta t + o(\Delta t)) \\
&= \lambda\Delta t + o(\Delta t) && (2.1) \\
p_{n,n}(\Delta t) &= \lambda\Delta t \cdot \mu\Delta t + (1 - \lambda\Delta t)(1 - \mu\Delta t) + o(\Delta t) \\
&= 1 - (\lambda + \mu)\Delta t + o(\Delta t), \quad n \geq 1 && (2.2) \\
p_{0,0}(\Delta t) &= 1 - \lambda\Delta t + o(\Delta t) && (2.3) \\
p_{n+1,n}(\Delta t) &= (1 - \lambda\Delta t)\mu\Delta t + o(\Delta t) \\
&= \mu\Delta t + o(\Delta t), \quad n \geq 0 && (2.4) \\
p_{n,m}(\Delta t) &= o(\Delta t), \quad |m - n| \geq 2 && (2.5)
\end{aligned}$$

From the law of total probability:

$$p_0(t + \Delta t) = \mu\Delta t\, p_1(t) + (1 - \lambda\Delta t)\, p_0(t) + o(\Delta t) \tag{2.6}$$
$$p_n(t + \Delta t) = \mu\Delta t\, p_{n+1}(t) + \lambda\Delta t\, p_{n-1}(t) + (1 - (\lambda + \mu)\Delta t)p_n(t) + o(\Delta t), \quad n \geq 1 \tag{2.7}$$

2.6.1 Differential Equations

Taking the limit as $\Delta t \to 0$, we obtain the Kolmogorov forward differential equations:

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) + \mu p_1(t) \tag{2.8}$$
$$\frac{dp_n(t)}{dt} = -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), \quad n \geq 1 \tag{2.9}$$

Solving this system analytically involves Bessel functions [38], which are beyond the scope of this discussion.

2.6.2 Steady-State Probabilities

In the long run (as $t \to \infty$), the system reaches a steady state. That is:

$$\lim_{t\to\infty} p_n(t) = p_n, \quad n \geq 0 \tag{2.10}$$
$$\lim_{t\to\infty} \frac{dp_n(t)}{dt} = 0 \tag{2.11}$$

These results allow us to analyze the equilibrium behavior of the queue, independent of initial conditions [?].

Balance Equations

Equations (2.1) and (2.2) yield the following system of balance equations: [?]

$$\mu p_1 = \lambda p_0 \tag{2.12}$$
$$\lambda p_{n-1} + \mu p_{n+1} = (\lambda + \mu) p_n, \quad n \geq 1 \tag{2.13}$$

The solution of this system proceeds as follows: [?]

- *For* $n = 1$: $\lambda p_0 + \mu p_2 = (\lambda + \mu) p_1 \Rightarrow p_2 = \left(\frac{\lambda}{\mu}\right)^2 p_0$.
- *For* $n > 1$: $p_n = \left(\frac{\lambda}{\mu}\right)^n p_0$.

To find the value of p_0*, we apply the normalization condition:*

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow p_0 \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = 1 \tag{2.14}$$

This is a geometric series with ratio $\rho = \frac{\lambda}{\mu}$*. It converges if* $\rho < 1$*:*

$$p_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho \tag{2.15}$$

Thus, the stationary distribution is: [15]

$$p_n = (1 - \rho)\rho^n, \quad n \geq 0 \tag{2.16}$$

- *An M/M/1 queue is stable if* $\lambda < \mu$*.*
- *The utilization factor is* $\rho = \frac{\lambda}{\mu} = 1 - p_0$*.*
- *If* $\rho > 1$*, the system becomes unstable and the queue length diverges.*

2.7 Performance Measures

Once the stationary regime is established, we define: [1]

- $L = \mathbb{E}[N]$*: average number of customers in the system.*
- L_q*: average number of customers in the queue.*
- W*: average time a customer spends in the system.*
- W_q*: average waiting time in the queue.*
- W_q^**: average waiting time for customers who must wait.*

According to Little's law: [29]

$$L = \lambda W \tag{2.17}$$
$$L_q = \lambda W_q \tag{2.18}$$
$$W = W_q + \frac{1}{\mu} \tag{2.19}$$
$$L = L_q + \frac{\lambda}{\mu} \tag{2.20}$$

For the M/M/1 system ($\lambda_e = \lambda$), we compute:

$$L = \sum_{n=0}^{\infty} np_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \tag{2.21}$$

$$L_q = L - \rho = \frac{\lambda^2}{\mu(\mu-\lambda)} \tag{2.22}$$

$$W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{2.23}$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda} \tag{2.24}$$

2.8 Waiting Time Distribution

When a customer arrives and finds n customers ahead, the total waiting time T is the sum of n exponential random variables of rate μ plus his own service time. This follows a Gamma distribution $\Gamma(n+1,\mu)$: [?]

$$P(T \le t \mid n) = \int_0^t \frac{\mu^{n+1}\tau^n}{n!} e^{-\mu\tau} d\tau \tag{2.25}$$

In steady-state, the cumulative distribution function (CDF) is:

$$P(T \le t) = \sum_{n=0}^{\infty} P(T \le t \mid n) P(X=n) \tag{2.26}$$

$$= \sum_{n=0}^{\infty} \left[\int_0^t \frac{\mu^{n+1}\tau^n}{n!} e^{-\mu\tau} d\tau \cdot (1-\rho)\rho^n \right] \tag{2.27}$$

$$= (1-\rho)\mu \int_0^t e^{-\mu\tau} \sum_{n=0}^{\infty} \frac{(\rho\mu\tau)^n}{n!} d\tau \tag{2.28}$$

$$= (1-\rho)\mu \int_0^t e^{-\mu\tau} e^{\rho\mu\tau} d\tau \tag{2.29}$$

$$= (1-\rho)\mu \int_0^t e^{-\tau(\mu-\rho\mu)} d\tau = (1-\rho)\mu \int_0^t e^{-\tau(\mu-\lambda)} d\tau \tag{2.30}$$

$$= 1 - e^{-t(\mu-\lambda)} \tag{2.31}$$

Hence, $T \sim Exp(\mu-\lambda)$ with mean:

$$\mathbb{E}[T] = \frac{1}{\mu-\lambda} \tag{2.32}$$

This confirms Little's law:

$$L = \lambda W \tag{2.33}$$

[29]

Example 2.1. *Derivation of Waiting Time Distribution The derivation involves calculating the probability that the waiting time T is less than or equal to t, given that there are n customers ahead in the queue. This conditional probability is based on the assumption that the service times are independent and identically distributed.*
Given n customers ahead, the total service time before the tagged customer is the sum of n

independent service times. If the service time of each customer follows an exponential distribution with rate μ, then the sum of n such exponential random variables follows a Gamma distribution with parameters n and μ. Therefore, the conditional probability that the waiting time $T \leq t$, given n customers ahead, is given by the cumulative distribution function of the Gamma distribution:

$$P(T \leq t \mid n) = \int_0^t \frac{\mu^n x^{n-1} e^{-\mu x}}{(n-1)!} dx$$

Now, suppose the number of customers ahead follows a Poisson distribution with parameter λ, then the total (unconditional) probability is obtained by summing over all possible values of n:

$$P(T \leq t) = \sum_{n=0}^{\infty} P(T \leq t \mid n) \cdot \frac{\lambda^n e^{-\lambda}}{n!}$$

This method combines the Poisons distribution for the number of customers in the system and the Gamma distribution for the sum of exponential service times, which is a standard approach in the analysis of queueing systems such as the M/M/1 model [?, ?, 38].

2.9 Description of Other Queueing Systems

2.9.1 The Queue M/M/C

This model is an extension of the M/M/1 queue, with C identical and independent servers. Customer arrivals follow a Poisons process with rate λ, and service times are exponentially distributed with rate μ per server [17, 25].
The state space is $E = \{0, 1, 2, \dots\}$. This is a birth-death process with transition rates:

$$\lambda_n = \lambda, \quad \mu_n = \begin{cases} n\mu & \text{if } 0 < n < C \\ C\mu & \text{if } n \geq C \end{cases}$$

For $n < C$, all customers are being served. The probability of a transition from state n to $n-1$ in time Δt is:

$$p_{n,n-1}(\Delta t) = \left(\sum_{j=1}^{n} (\mu \Delta t + o(\Delta t))(1 - \mu \Delta t + o(\Delta t))^{n-1} \right) (1 - \lambda \Delta t + o(\Delta t))$$

Using a first-order Taylor expansion:

$$p_{n,n-1}(\Delta t) = n\mu \Delta t + o(\Delta t)$$

For $n \geq C$, only C servers are active, giving a transition rate of $C\mu$. The arrival rate remains λ.The ***stability condition*** *is:*

$$\lambda < C\mu$$

The steady-state probabilities satisfy:

$$\begin{cases} P_n = \rho P_{n-1} & \text{for } 1 \leq n < C \\ P_n = \dfrac{\rho}{C} P_{n-1} & \text{for } n \geq C \end{cases} \qquad \text{where } \rho = \frac{\lambda}{\mu}$$

Performance Metrics

The average waiting time is given by:

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

To compute L_q (the average number in queue), we use:

$$L_q = \sum_{n=C}^{\infty} (n-C)P_n = \frac{\rho^C}{C!} P_0 \sum_{n=1}^{\infty} n\rho^{n-1} = \frac{\rho^C}{C!} P_0 \cdot \frac{\rho}{(1-\rho)^2}$$

Hence,

$$W = \frac{\rho^C}{C!(C-\rho)^2} \cdot \frac{P_0}{\mu} + \frac{1}{\mu}$$

$$L = \lambda W = \frac{\rho^{C+1}}{(C-1)!(C-\rho)^2} P_0 + \rho$$

2.9.2 The Queue M/M/∞

This system has an infinite number of servers. Each arriving customer is served immediately. Arrivals follow a Poisons process with rate λ, and services are exponentially distributed with rate μ [3, 37].
This yields:

$$\lambda_n = \lambda, \quad \mu_n = n\mu$$

The equilibrium condition:

$$P_n = \frac{\rho^n}{n!} P_0, \quad \textit{where } \rho = \frac{\lambda}{\mu}$$

With:

$$P_0 = \left(\sum_{n=0}^{\infty} \frac{\rho^n}{n!} \right)^{-1} = e^{-\rho} \quad \Rightarrow \quad P_n = \frac{\rho^n}{n!} e^{-\rho}$$

Performance Metrics

Throughput:

$$d = \sum_{n=1}^{\infty} n\mu P_n = \mu \cdot \sum_{n=1}^{\infty} n \cdot \frac{\rho^n}{n!} e^{-\rho} = \lambda$$

Average Number of Customers:

$$L = \sum_{n=1}^{\infty} nP_n = \rho$$

Average Waiting Time:

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

Chapter 3

Priority Queueing Models among Customer Classes

3.1 Mathematical Models

In this chapter, we develop mathematical models for priority queueing systems with multiple customer classes. These models will be derived under the assumption of an M/M/1 system, where arrivals follow a Poisson process and service times are exponentially distributed. We will consider both preemptive and non-preemptive priority disciplines.

3.1.1 M/M/1 Priority Queue

The classical M/M/1 queue is the simplest form of a queueing system with a single server. The system is characterized by the following:

- *Arrivals follow a Poisson process with rate* λ*.*
- *Service times are exponentially distributed with rate* μ*.*
- *There is a single server in the system.*

The steady-state distribution for the number of customers in the system is given by the following probability mass function:

$$P_n = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots$$

where $\rho = \frac{\lambda}{\mu}$ *is the utilization factor of the server.*

The average number of customers in the system is:

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

3.1.2 Priority Queue with Two Classes

Now, consider an M/M/1 queue with two classes of customers: high-priority (class 1) and low-priority (class 2). Each class has its own arrival rate λ_1 *and* λ_2*, respectively. The service rate for both classes is* μ*, and the server serves class 1 customers first (non-preemptive priority).*

Non-preemptive Priority

In the non-preemptive case, if a high-priority customer arrives while a low-priority customer is being served, the low-priority customer continues receiving service until completion. The system's behavior can be modeled as follows:
The probability that the system is serving a high-priority customer is:

$$P_{high} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

The probability of serving a low-priority customer is:

$$P_{low} = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

The average waiting time for high-priority customers is:

$$W_1 = \frac{1}{\mu - \lambda_1}$$

For low-priority customers, the average waiting time is:

$$W_2 = \frac{1}{\mu - \lambda_2} + \frac{1}{\mu - \lambda_1}$$

Preemptive Priority

In the preemptive priority system, the service of a low-priority customer can be interrupted when a high-priority customer arrives. The interrupted customer will resume service after the high-priority customer is served.

The system behavior under preemptive priority can be derived by modifying the arrival and service rates. The key difference is that class 1 customers effectively "block" the server from serving class 2 customers. The effective service rate for class 1 customers is μ, while the service rate for class 2 customers is effectively reduced during times when class 1 customers are being served.

The average waiting time for a high-priority customer is still:

$$W_1 = \frac{1}{\mu - \lambda_1}$$

However, for a low-priority customer, the waiting time is affected by interruptions and is given by:

$$W_2 = \frac{1}{\mu - \lambda_2} + \frac{1}{\mu - \lambda_1}$$

3.1.3 Generalized Priority Models

In more complex scenarios with multiple classes and varying priority schemes, the analysis involves generalizing the M/M/c queue model. For instance, in a system with c servers and multiple classes of customers, each class may have a different service rate and priority, and the system can be described by a system of differential equations or using matrix-analytic methods.

3.2 Mathematical Derivations and Diagrams

3.2.1 Derivation of Waiting Times for Two-Class M/M/1 Queue (Non-Preemptive)

Assume the following:

- *Arrival rate for Class 1 (high-priority):* λ_1
- *Arrival rate for Class 2 (low-priority):* λ_2
- *Service rate (shared):* μ
- $\lambda = \lambda_1 + \lambda_2 < \mu$

Class 1: High Priority Customers

They are unaffected by Class 2 customers. Thus, their waiting time is simply:

$$W_1 = \frac{1}{\mu - \lambda_1}$$

Class 2: Low Priority Customers

These customers must wait for all Class 1 customers ahead of them, so their waiting time includes:

$$W_2 = \frac{1}{\mu - \lambda_2} + \frac{1}{\mu - \lambda_1}$$

Derivation Justification

For Class 1: *The Class 1 customers effectively experience a standard M/M/1 queue because:*

- *Their arrivals follow a Poisson process with rate* λ_1
- *They are never delayed by Class 2 customers*
- *The service rate is* μ

Thus, by standard M/M/1 theory, the expected waiting time in the system is:

$$W_1 = \frac{1}{\mu - \lambda_1}$$

For Class 2: *Class 2 customers must wait:*

1. *For any Class 1 customers already in the queue or arriving before them*
2. *For other Class 2 customers already ahead of them*

Their delay consists of two parts:

- ***Own Queue:*** *If Class 1 didn't exist, they'd behave like M/M/1 with rate* λ_2*:*

$$\textit{Own waiting: } \frac{1}{\mu - \lambda_2}$$

- ***Blocking by Class 1:*** *On average, they are blocked by Class 1 traffic:*

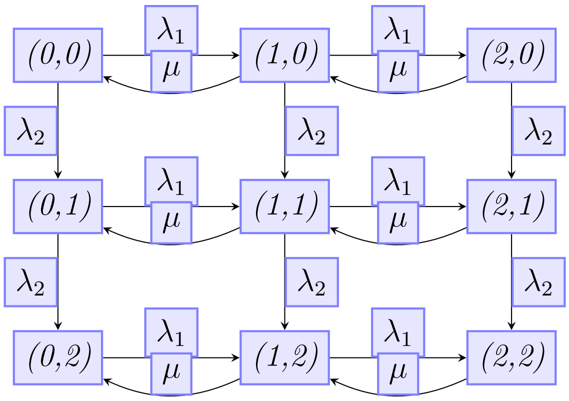
$$\textit{Class 1 blocking time: } \frac{1}{\mu - \lambda_1}$$

Therefore, the total expected waiting time for Class 2 is:

$$W_2 = \frac{1}{\mu - \lambda_2} + \frac{1}{\mu - \lambda_1}$$

Note: *This derivation assumes the non-preemptive rule, i.e., once a service begins for any customer, it completes even if a higher priority customer arrives.*

3.2.2 State Transition Diagram for M/M/1 with Non-Preemptive Priority



3.3 Performance Analysis

In this chapter, we analyze the performance of priority queueing systems using mathematical tools derived in the previous chapter. We focus on key performance indicators such as average waiting time, queue length, and system utilization, with a special emphasis on the effects of customer prioritization.

3.3.1 Average Waiting Time

The average waiting time, denoted by W_i, is one of the most important performance measures in a queueing system. It represents the expected time a customer spends in the queue before receiving service.

For an M/M/1 queue with a single class of customers, the average waiting time is given by the well-known formula:

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

where λ is the arrival rate and μ is the service rate.

For a system with two classes of customers, the waiting times differ between high-priority and low-priority customers. For non-preemptive priority, the average waiting time for a high-priority customer is:

$$W_1 = \frac{1}{\mu - \lambda_1}$$

For low-priority customers, the waiting time is affected by the high-priority customers:

$$W_2 = \frac{1}{\mu - \lambda_2} + \frac{1}{\mu - \lambda_1}$$

3.3.2 Average Number of Customers in the System

The average number of customers in the system, L, is another key performance measure. It represents the expected number of customers present in the system, including those being served.

For an M/M/1 queue with a single class of customers, the average number of customers in the system is given by:

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

For systems with multiple classes, the average number of customers in the system can be found by considering the contributions of both high-priority and low-priority customers. For the M/M/1 queue with two classes, the average number of customers in the system is:

$$L_1 = \frac{\lambda_1}{\mu - \lambda_1}, \quad L_2 = \frac{\lambda_2}{\mu - \lambda_2}$$

Thus, the total number of customers in the system is:

$$L_{total} = L_1 + L_2$$

3.3.3 System Utilization

System utilization is an important performance indicator, particularly in multi-server systems. It is defined as the fraction of time the server is busy. For a system with a single server, the utilization ρ is given by:

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

In a multi-class priority queueing system, the utilization for each class depends on the arrival rate and the service rate of the system. For example, in a non-preemptive priority system, the utilization for high-priority customers is:

$$\rho_1 = \frac{\lambda_1}{\mu}$$

Similarly, the utilization for low-priority customers can be expressed as:

$$\rho_2 = \frac{\lambda_2}{\mu}$$

3.3.4 Probability of Delay

The probability that a customer will have to wait before being served is an important performance measure, especially in systems with strict quality of service requirements. This can be expressed as the probability that the system is not empty, i.e., that there is at least one customer in the system. For an M/M/1 queue, this probability is:

$$P_{delay} = \rho$$

For a multi-class priority system, the probability of delay for high-priority customers is:

$$P_{delay,\ high} = \frac{\lambda_1}{\mu}$$

For low-priority customers, the probability of delay is:

$$P_{delay,\ low} = \frac{\lambda_2}{\mu}$$

3.3.5 Effect of Prioritization on Performance

In general, introducing priority mechanisms into a queueing system can have significant effects on performance metrics. For example, higher-priority customers often experience lower waiting times, but this can result in longer waiting times for low-priority customers, as they are only served when no high-priority customers are in the system.

Graphically, the effect of priority on waiting times can be illustrated by plotting the average waiting time as a function of the arrival rate for both classes. The results typically show that as the arrival rate increases, the waiting time for low-priority customers increases significantly more than for high-priority customers.

In the next chapter, we will provide numerical examples to illustrate these effects and compare the performance of different priority schemes.

3.4 Numerical Examples and Illustrations

In this chapter, we present numerical examples and graphical illustrations to demonstrate the impact of priority queueing models on system performance. We consider both preemptive and non-preemptive disciplines and analyze key performance indicators such as average waiting time, queue length, and system utilization for each customer class.

3.4.1 Assumptions and Parameter Values

We consider a queueing system with the following parameters:

- *Service rate:* $\mu = 10$ *customers per unit time*
- *Arrival rate of high-priority customers:* $\lambda_1 = 4$
- *Arrival rate of low-priority customers:* $\lambda_2 = 3$

The total arrival rate is $\lambda = \lambda_1 + \lambda_2 = 7$*, which is less than* μ*, ensuring system stability.*

3.4.2 Non-Preemptive Priority Analysis

Using the analytical formulas from the previous chapters:

- *Average waiting time for high-priority customers:*

$$W_1 = \frac{1}{\mu - \lambda_1} = \frac{1}{10 - 4} = \frac{1}{6} \approx 0.1667$$

- *Average waiting time for low-priority customers:*

$$W_2 = \frac{1}{\mu - \lambda_2} + \frac{1}{\mu - \lambda_1} = \frac{1}{10 - 3} + \frac{1}{10 - 4} = \frac{1}{7} + \frac{1}{6} \approx 0.3095$$

3.4.3 Preemptive Priority Analysis

In the preemptive case, high-priority customers always preempt low-priority customers, which can reduce their waiting time:

$$W_1 = \frac{1}{\mu - \lambda_1} = \frac{1}{6} \approx 0.1667$$

$$W_2 = \text{(increased due to preemptions, analysis involves Laplace transforms or simulation)}$$

To simplify, we consider the effective service rate for low-priority customers as:

$$\mu_{\mathit{eff}} = \mu - \lambda_1 = 6$$

$$W_2 \approx \frac{1}{\mu_{\mathit{eff}} - \lambda_2} = \frac{1}{6-3} = \frac{1}{3} \approx 0.3333$$

3.4.4 Graphical Illustration

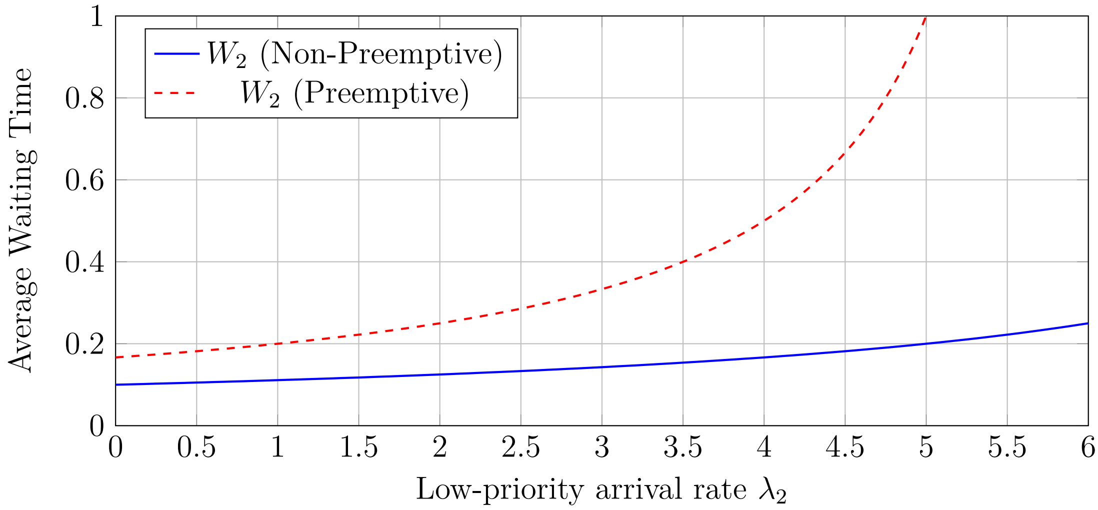


Figure 3.1: Effect of increasing λ_2 on W_2 under both disciplines (with fixed $\lambda_1 = 4$ and $\mu = 10$)

3.4.5 Observations

- *The waiting time for high-priority customers remains stable regardless of the arrival rate of low-priority customers.*
- *The waiting time for low-priority customers increases more rapidly under the preemptive discipline.*
- *Preemptive policies are more favorable for critical services where delay for high-priority customers is unacceptable.*

In the next chapter, we will explore real-world applications of these priority queueing models in various industries.

3.5 Applications in Real-World Systems

Priority queueing models are widely applicable in real-world systems where it is necessary to distinguish between different types of customers, tasks, or data based on urgency, importance, or required quality of service. This chapter highlights key areas where such models are practically implemented.

3.5.1 Healthcare Systems

In hospitals and emergency rooms, patients are typically triaged based on the severity of their conditions. For example:

- *Critical patients (e.g., heart attacks, trauma) are served immediately.*
- *Less severe cases (e.g., mild injuries or checkups) may wait longer.*

This setup can be modeled as a multi-class priority queue, where high-priority patients have either preemptive or non-preemptive service over lower-priority ones. ***Example:*** *In an emergency room with one doctor (server), high-priority patients are served immediately upon arrival (preemptive), while others wait.*

3.5.2 Telecommunication Networks

Priority queueing is essential in communication systems to ensure Quality of Service (QoS). Different data packets are assigned different priorities:

- *Voice or video data (requiring low latency) gets higher priority.*
- *Email or file transfers (delay-tolerant) are low priority.*

Routers use scheduling algorithms like Weighted Fair Queueing (WFQ) or Priority Queuing (PQ) to manage such priorities efficiently.

3.5.3 Manufacturing and Production Lines

In manufacturing, certain orders (e.g., urgent or high-value customers) are given processing priority over others:

- *Rush orders are treated as high priority.*
- *Standard orders follow a regular queue.*

Queueing models with priority help optimize production schedules, minimize delay for key clients, and reduce penalties.

3.5.4 Cloud Computing and Data Centers

Cloud service providers often use priority-based resource allocation for:

- *Premium users (with Service Level Agreements) who require faster processing.*
- *Free-tier users who are served when resources are available.*

This is modeled as a priority queue with either preemptive or probabilistic service discipline.

3.5.5 Customer Service and Call Centers

Call centers frequently employ priority queuing:

- *VIP customers or subscribers with premium packages are queued ahead.*
- *Regular customers may be on hold longer.*

Simulation and optimization of such queueing systems improve overall customer satisfaction and resource efficiency.

3.5.6 Traffic and Transportation Systems

Traffic control systems use priority queueing, for example:

- *Emergency vehicles get priority at intersections.*
- *Public transport lanes provide faster service to buses.*

Queueing models support the design of traffic signals and lane prioritization strategies to improve flow and safety.

3.5.7 Observations

Across these applications:

- *Priority queueing enhances system responsiveness and efficiency.*
- *The choice of preemptive vs. non-preemptive policy depends on context.*
- *Proper analysis and simulation are crucial to avoid unfair service or resource starvation.*

In the next chapter, we summarize the key findings and propose future directions for research.

Conclusion and Future Work

In this study, we have conducted an in-depth exploration of priority queueing models among different customer classes. Through theoretical formulation, mathematical modeling, and numerical analysis, we have demonstrated how prioritization impacts the performance and efficiency of service systems. The main contributions and findings can be summarized as follows:

- *We introduced the foundational concepts of queueing theory and extended them to multi-class priority systems, distinguishing between preemptive and non-preemptive strategies.*
- *We formulated and derived key performance measures such as average waiting time, queue length, and utilization for each priority class.*
- *We presented numerical examples to illustrate the influence of priority rules on customer experience and service fairness.*
- *We examined various real-world applications including healthcare, telecommunication, manufacturing, cloud computing, and transportation systems.*

Overall, priority queueing models are powerful tools for managing heterogeneous service demands, ensuring responsiveness for high-priority entities, and optimizing resource allocation. ***Future Work*** *Despite the progress made in this research, several avenues remain open for future study and enhancement:*

- ***Stochastic Modeling:*** *Incorporating non-Markovian arrival and service processes (e.g., M/G/1, G/M/1) for greater realism.*
- ***Simulation-Based Optimization:*** *Applying discrete-event simulation techniques for more complex systems with dynamic priorities or time-varying rates.*
- ***Machine Learning Integration:*** *Using predictive models to assign dynamic priorities based on customer behavior or system load.*
- ***Fairness and Ethics:*** *Balancing service efficiency with fairness, especially in sensitive contexts like healthcare and public services.*
- ***Energy-Aware Queueing:*** *Considering energy consumption and sustainability in priority-based scheduling for data centers and mobile networks.*

We hope this work serves as a foundation for further research in advanced queueing systems and contributes to both academic understanding and practical implementations.

Bibliography

[1] Allen, A. O. Probability, Statistics, and Queueing Theory with Computer Science Applications. Academic Press, 1990.

[2] Bailey, N. T. J. (1964). The Elements of Stochastic Processes with Applications to the Natural Sciences. John Wiley & Sons.

[3] Bertsekas, D., & Gallager, R. (1992). Data Networks (2nd ed.). Prentice Hall.

[4] Billingsley, P. (1995). Probability and Measure (3rd ed.). New York: John Wiley & Sons.

[5] Brémaud, P. (1999). Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues. Springer.

[6] Brockwell, P. J., & Davis, R. A. (2002). Introduction to Time Series and Forecasting. Springer.

[7] Çinlar, E. (2011). Probability and Stochastics. Springer.

[8] Doob, J. L. (1953). Stochastic Processes. Wiley.

[9] Durrett, R. (2019). Probability: Theory and Examples (5th ed.). Cambridge University Press.

[10] Erlang, A. K. (1909). The theory of probabilities and telephone conversations. Nyt Tidsskrift for Matematik B, 20, 33–39.

[11] Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Vol. 2. 2nd ed., Wiley.

[12] Gardner, W. A. Introduction to Random Processes with Applications to Signals and Systems. McGraw-Hill, 1990.

[13] Grimmett, G. R., & Stirzaker, D. R. (2001). Probability and Random Processes (3rd ed.). Oxford University Press.

[14] Grinstead, C. M., & Snell, J. L. (1997). Introduction to Probability. American Mathematical Society.

[15] Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2018). Fundamentals of Queueing Theory (5th ed.). Wiley.

[16] Gross, D., and Harris, C. M. (1985). Fundamentals of Queueing Theory (2nd ed.). John Wiley & Sons.

[17] Gross, D., & Harris, C. M. (1998). Fundamentals of Queueing Theory (3rd ed.). Wiley.

[18] Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2008). Fundamentals of Queueing Theory (4th ed.). Wiley.

[19] Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press.

[20] Karatzas, I., & Shreve, S. E. (1991). Brownian Motion and Stochastic Calculus. Springer.

[21] *Karlin, S., & McGregor, J. (1957). The classification of birth and death processes. Transactions of the American Mathematical Society, 86(2), 366–400.*

[22] *Karlin, S., & Taylor, H. (1998). An Introduction to Stochastic Modeling. Academic Press, 3rd edition.*

[23] *Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Annals of Mathematical Statistics, 24(3), 338–354.*

[24] *Kingman, J. F. C. (1993). Poisson Processes. Oxford University Press.*

[25] *Kleinrock, L. (1975). Queueing Systems, Volume 1: Theory. Wiley-Interscience.*

[26] *Krengel, U. Ergodic Theorems. Walter de Gruyter, 1985.*

[27] *Kulkarni, V. G. (2016). Modeling and Analysis of Stochastic Systems (2nd ed.). Chapman and Hall/CRC.*

[28] *Levin, D. A., Peres, Y., & Wilmer, E. L. (2009). Markov Chains and Mixing Times. American Mathematical Society.*

[29] *Little, J. D. C. (1961). A Proof for the Queuing Formula:* $L = \lambda W. Operations Research, 9(3), 383–387.$

[30] *Medhi, J. (2003). Stochastic Models in Queueing Theory (2nd ed.). Academic Press.*

[31] *Medhi, J. (2002). Stochastic Processes (3rd ed.). New Age International Publishers.*

[32] *Norris, J. R. (1997). Markov Chains. Cambridge University Press.*

[33] *Ross, S. M. (1996). Stochastic Processes (2nd ed.). Wiley.*

[34] *Ross, S. M. (2014). Introduction to Probability Models (11th ed.). Academic Press.*

[35] *Sato, K. (1999). Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press.*

[36] *Serfozo, R. (2009). Basics of Applied Stochastic Processes. Springer.*

[37] *Tijms, H. C. (2003). A First Course in Stochastic Models. Wiley.*

[38] *Wolff, R. W. (1989). Stochastic Modeling and the Theory of Queues. Prentice Hall.*

ملخص الدراسة

تهدف هذه الدراسة إلى تحليل أنظمة الطوابير باستخدام النماذج الاحتمالية، مع التركيز على نموذج ش\ش\١ ونماذج الطوابير ذات الأولوية بين فئات الزبائن. تبدأ المذكرة بأساسيات العمليات التصادفية وسلاسل ماركوف، ثم تنتقل إلى دراسة تفصيلية لأنظمة الطوابير المختلفة من حيث بنيتها وأدائها. يتم اشتقاق معادلات رياضية وتحليل مقاييس الأداء مثل زمن الانتظار وعدد الزبائن في النظام. كما تشمل الدراسة تطبيقات حقيقية في مجالات متعددة مثل الصحة، الاتصالات، والإنتاج. تختتم المذكرة بخلاصة النتائج وتوصيات للبحث المستقبلي.

الكلمات المفتاحية:

نماذج الطوابير، العمليات التصادفية، سلاسل ماركوف، أولويات الخدمة، مقاييس الأداء.

٥

Résumé de l'étude

Cette étude analyse les systèmes de files d'attente à l'aide de modèles probabilistes, en mettant l'accent sur le modèle M/M/1 et les files d'attente à priorité entre différentes classes de clients. Elle commence par les notions fondamentales des processus stochastiques et des chaînes de Markov, avant de développer une analyse détaillée des systèmes de files d'attente selon leur structure et leurs performances. Des dérivations mathématiques permettent d'évaluer des mesures de performance telles que le temps d'attente ou le nombre moyen de clients. L'étude aborde également des applications pratiques dans les domaines de la santé, des télécommunications et de la production industrielle. Elle se termine par une synthèse des résultats et des pistes pour les travaux futurs.

Mots-Clés :

Modèles de files d'attente, processus stochastiques, chaînes de Markov, files à priorité, mesures de performance.

Study Summary

This study analyzes queueing systems using probabilistic models, with a focus on the M/M/1 model and priority queues among customer classes. It begins with the fundamentals of stochastic processes and Markov chains, then presents a detailed examination of queueing systems in terms of structure and performance. Mathematical derivations are used to analyze key performance measures such as waiting time and the number of customers in the system. The study also includes real-world applications in healthcare, telecommunications, and manufacturing. It concludes with a summary of results and suggestions for future research.

Key words:

Queueing models, stochastic processes, Markov chains, priority queues, performance measures.