one to take retrial policy for SUs into account. Sun et al. [122] modeled the CRN as a multi-channel queueing model of the $MMAP/M_2/N/N$ type, where $MMAP$ stands for marked Markovian arrival process. Therein, a retrial orbit is employed for the SUs and with the aim to enhance their throughput, access of the SUs is restricted via threshold mechanism. The optimal value of the threshold is also obtained then. Gao [123] investigated a single channel retrial queueing model with general retrial times. He showed the application of the proposed model in CRNs, with PUs having preemptive priority over SUs. Further, Dudin et al. [71] generalized the results of retrial queue obtained by Sun et al. [122]. They assumed service times of priority users have much more general, phase type distribution in contrast to the exponential distribution. Additionally, they took different bandwidth requirements into consideration by providing a whole channel to the PU and a sub-channel to the SU for service. Recently, Zhao and Yue [124] proposed a spectrum sharing strategy with a returning threshold and a returning probability to dynamically control the retransmissions of interrupted SUs. In this strategy, when the number of SUs already in the SU buffer reaches a predefined returning threshold, then an interrupted SU will be admitted to the buffer for later retransmission with an adjustable returning probability. To make the proposed strategy more adaptive, this probability is assumed to be inversely proportional to the total number of users in the system.

1.6.3.3 Queues with Working Breakdowns and Vacations

Queueing systems with server breakdown and vacations have also been investigated in different frameworks in recent years. The server can take vacations after being busy for a certain period of time or after serving a (fixed or random) number of customers or packets as discussed by Doshi [125]. An example of a vacation model in telecommunication applications involves the ever-increasing use of wireless cellular networks, which triggers a huge consumption of energy. In order to develop energy efficient wireless cellular networks, researchers proposed hibernation (or sleeping) of a BS in the absence of an active user in the network. A sleeping BS is similar to a server on a vacation. There are a few vacation based BS sleeping models available in the literature. For instance, Marsan

et al. [126] studied the energy-aware management of cellular networks with simple analytical models to optimize the energy saving. When other cells are switched off, they assume that the cells that remain active take care of radio coverage and service provisioning, to ensure that service is available over the whole area. Moreover, according to a deterministic traffic variation pattern over time, they gave a static BS sleep pattern. While based on the traffic variation with respect to certain blocking probability constraint, Gong et al. [127] proposed an energy saving algorithm that dynamically adjusts BSs' working modes (active or sleeping). Later on, Wu et al. [128] surveyed BS sleep mode techniques and their applications to mobile networks under different assumptions on system and power model. They have shown that the advantages of sleep mode strategies are greatly affected by the assumptions. Recently, Feng et al. [129] identified the challenges of designing BS ON-OFF switching (also known as BS sleep control) strategies in 5G wireless networks. They also presented an overview of recent advances on different switching mechanisms.

The assumption that the service environment to be one hundred percent reliable can be impractical, especially in communication systems. The frequency channel can fail due to signal fading, channel interference, weak transmission power, path loss, etc. leading to the study of stochastic models with server subject to random breakdowns and repairs. For instance, Cao and Singhal [130] proposed a fault-tolerant channel allocation algorithm, which is responsible for collecting information from other cells to find the available channels while guaranteeing that the channel assignment does not interfere with other cells. In the algorithm, a borrower need not receive a response from every interference neighbor rather from a small portion of them and hence the algorithm can tolerate network congestion and communication link failures. The performance of the algorithm was evaluated under both environments i.e., with and without failures of communication links or mobile hosts. Later, a fault-tolerant dynamic channel allocation scheme for cellular networks was developed by Boukerche et al. [131], which can handle mobile host failures, BS failures as well as communication link failures under both uniform and non-uniform call arrival distributions. Sattiraju and Schotten [132] introduced a new framework for

modeling and analyzing reliability of the wireless link under the effects of channel fading and retransmissions by considering the channel as a non-repairable system. Recently, Nemouchi and Sztrik [133] presented a finite-source retrial queueing system for CRNs consisting of two interconnected, not independent sub-systems. In the first part, the PU requests with preemptive priority are sent to a single server unit or Primary Channel Service (PCS) and the second sub-system is for SU requests at Secondary Channel Service (SCS). Therein, both servers are subject to random breakdowns and repairs. The failure and repair times are assumed to be generally distributed, in particular hypo-exponentially and hyper-exponentially distributed for simulation results. Further, Zaghouani et al. [134] extended the analysis of Nemouchi and Sztrik [133] by allowing Gamma distributed failure and repair times in addition to the Hypo-Exponential and Hyper-Exponential ones with the same mean and variance. This enables them to capture the effect of the distributions contrary to that of only the first two moments.

Additionally, in order to avoid excessive delays, many authors have considered the provision of a backup server to serve at a reduced rate during the absences (caused by vacations and breakdowns) of the main server, referred as working vacation and working breakdown. The detailed study of these models can be found in the works of Chandrasekaran et al. [135] and Deepa and Kalidass [136]. Servi and Finn [137] were the first to introduce an $M/M/1$ queueing system with working vacations (WV). Wu and Takagi [138] extended the $M/M/1/WV$ queue to an $M/G/1/WV$ queue. Authors like Zhang and Hou [139], Arivudainambi et al. [140], Gao et al. [141], Zhang and Liu [142], Rajadurai et al. [143] and Rajadurai [144, 145] analyzed queueing systems with working vacations. Kalidass and Kasturi [146] were the first to introduce the concept of working breakdowns. For unreliable queueing systems, working breakdown service is considered as a more reasonable repair policy as it can decrease complaints from the users who should wait for the main server to be repaired and reduces the cost of waiting users. Kim and Lee [147] studied an $M/G/1$ queueing system with working breakdown services. Recently, Rajadurai et al. [148] investigated an $M/G/1$ retrial queue with negative customers (disasters) under working vacations and working breakdowns.

1.6.3.4 Queues with Varying Arrival and Service Processes

To develop a detailed analytical model for mobile and wireless cellular systems, the call arrival process and the call holding times (CHT) are the two main characteristics. Different models are derived based on different assumptions for the distribution of inter-arrival times and call durations. For instance, a working vacation model wherein there are two types of priority customers, a possibility of interrupting the vacations, and with CHT following a phase type (PH−) distribution was studied by Goswami and Selvaraju [149]. Later, Goswami and Selvarju [150] analyzed a $PH/M/c$ queue, wherein servers can go for multiple working vacations that is, the (main) server keeps on taking working vacations until the server finds no customers after returning from a vacation.

Empirical studies have shown that the interarrival times of a communication system demonstrate significant correlation as mentioned by Klemm et al. [151]. From this point of view, there are some queueing models that consider MAP arrivals to capture the correlation effect. For instance, Banerjee et al. [152] studied a finite buffer single server queue with MAP arrivals where service times, which depend on the batch size, are generally distributed. The queue length distribution for $MAP/G/1$ working-vacation-interruption queue was derived by Zhang and Hou [153]. The steady-state analysis of multiclass $MAP/PH/c$ retrial queueing system with acyclic PH−retrials was studied by Dayar and Orhan [154], where acyclic PH−distribution is a subclass of PH−distribution. Vadivu [155] investigated a finite buffer $MAP/G/1$ queue with an additional second phase of optional service under vacation policy. Based on queueing system models considered in the survey by Vishnevskii and Dudin [156], recently it was shown that the presence of a positive correlation significantly deteriorates the characteristics of a queueing system. Further, Ye and Liu [157] studied $MAP/M/1$-type queue with working breakdowns and repairs. In addition, the theory of queueing systems with correlated arrival processes developed in the works of M.F. Neuts, D.M. Lucantoni, V. Ramaswami, S.R. Chakravarthy, A.N. Dudin, V.I. Klimenok and others, has had numerous applications in the telecommunication networks, in particular for cellular networks, see e.g., the works

of Lee et al. [158], and Zhou and Zhu [159].

1.7 Research Gaps and Contributions

Based on the above discussions, several research gaps are revealed which are addressed with our proposed solutions throughout this thesis. This section summarizes those research contributions made by the proposed solutions.

There are a number of CAC schemes proposed by various researchers based on different aspects of service management in cellular networks. However, the literature survey identifies that not much focus has been stressed upon the quality of signal. Since deterioration in the quality of received signals may result in call drop therefore, in CAC it is of paramount importance to control the quality of signal for better performance. In view of this, we have designed two CAC schemes to quantify the impact of signal quality on cellular networks that are elaborated in Chapter 3. To support the mobility of users and further enhance the system performance, FGC policy with queueing of handoff requests is taken into consideration in the proposed scheme. Moreover, an algorithm is developed to determine the optimal value of new call acceptance probability associated with the FGC policy.

As mentioned earlier, spectrum sensing plays a significant role in enabling the utilization of spectrum holes by (unlicensed) SUs in CRNs. From the literature survey, it is revealed that most of the works concerning spectrum sensing has focused on sensing carried out by (only) incoming SUs aiming at locating spectrum opportunities. However, in order to appropriately protect licensed PUs, SUs should continuously perform spectrum sensing during their ongoing transmissions as well. False alarm rate (FAR) is an important issue associated with continuous sensing, which is defined as the average number of false alarms per unit of time and can severely degrade the QoS. Motivated by these issues, a queueing analysis of opportunistic access in CRNs is performed in Chapter 4 to examine the effect of sensing errors including the FARs on the performance of CRNs. To gain further insight, system-centric performance metrics such as capacity, blocking probability and forced termination probability are explored by employing queueing and

handoff schemes.

The gap between research work and real implementation of cellular networks exists due to various assumptions made during system modeling for the purpose of analysis convenience. For instance, the homogeneity of traffic that is assumed in many analytical models is not observed in practice since CR users are diverse in traffic types with distinct QoS requirements. With this motivation, Chapters 4 to 6 analyze a CRN with hybrid SU traffic considering both elastic/non-real time and real-time SU services.

In addition, the assumption of Poisson arrival process and/or exponential service times is widely adopted in the literature due to its simplicity and analytical tractability. However, a Poisson arrival process or an independent renewal process cannot capture the correlation of packets in a network. For high speed networks, data, video or voice traffic is seldom uniform and characterized by periods of burstiness i.e. packets are bursty and significantly correlated (see e.g. Jian and Dovrolis [160]). Inspired by this observation, we have studied a queueing model with MAP and phase type distributed service times. In Chapter 7, we model a working-vacation-breakdown-repair queue which incorporates these characteristics.

The popularity of stationary analysis comes from its simplicity and thus the available results of the transient (time-dependent) regime in the literature are usually restricted due to its complexity, as discussed by Parthasarathy and Dharamraja [161] and Legros [162]. However, considering the fact that the duration of a network connection is finite and steady-state might never be attained, it is required to investigate the transient behavior due to the real-time nature of the mobile traffic so as to know how the system will operate up to some specified time. Thereby, the transient analysis of CRNs is carried out in Chapters 5 and 6.

In real-life networks, a wireless link encounters a certain probability of channel failure due to various reasons. The impact of these failures and recoveries on the performance of CRNs is an overlooked area in existing research work due to the complexity of analysis. Indeed, the ignorance of the effect of link failures and recoveries in mathematical modeling generally overestimates the performance measures. Besides, the dependability

aspect of CRNs is also an overlooked topic even though a tremendous amount of research efforts have been made in the area of CRNs. When analyzing CRN performance, motivated by these observations, Chapters 5 and 6 attempt to model a more realistic network scenario from the perspectives of dependability theory by considering channel failures and repairs with the aim of achieving Ultra-Reliable Communication (URC) in 5G and beyond networks.

1.8 Organization of the Thesis

In general terms, the study of this thesis is devoted to develop an analytical framework for analyzing the performance of future cellular networks including CRNs. More specifically, it contributes towards a better understanding of cellular mobile networks and provides a new insight into their operation. The thesis consists of eight chapters. The first chapter provides an introduction, a summary of the research area and the performed work, together with the importance of performance analysis using queueing theory.

Chapter 2 provides an overview of a number of concepts in Markov modeling and analysis, useful for understanding the subsequent chapters. It also summarizes the performance metrics and the available solution methods for the proposed mathematical models.

Chapter 3 proposes two CAC schemes taking handoff prioritization and signal quality into account. A non-linear optimization problem is also formulated to minimize the new call blocking probability under a hard constraint on handoff dropping probability.

Chapter 4 presents a CTMC-based model for analyzing the steady-state performance of CRNs by addressing the issue of sensing errors. The study focuses on the joint analysis of spectrum access and continuous spectrum sensing by employing call buffering and handoff strategies.

Chapters 5 and 6 present transient analysis on reliability and availability performance of CRNs from dependability theory's perspective. The proposed analytical model in Chapter 5 incorporates channel reservation, handoff capability, retrial phenomenon and heterogeneous SU traffic support in a single model. Chapter 6 then presents a CTMC model using different channel modes and distinct failure rates. The proposed model

therein explores the multi-channel reservation scheme for heterogeneous SU traffic.

Chapter 7 discusses a queueing model with server breakdowns, repairs, vacations, and backup server under the steady-state. Using Neuts' versatile point process for the arrivals and modeling the service times with phase type distributions, the proposed model generalizes some of the previously published ones in the literature. Additionally, the decomposition results for the rate matrix and the mean number in the system are proved under some special cases.

Finally, Chapter 8 provides a conclusive summary, highlights the specific contributions of the work and points out a few potential extensions of the proposed schemes and models.

The structure of this thesis is also illustrated in Figure 1.5 where the connections among the research topics are highlighted.

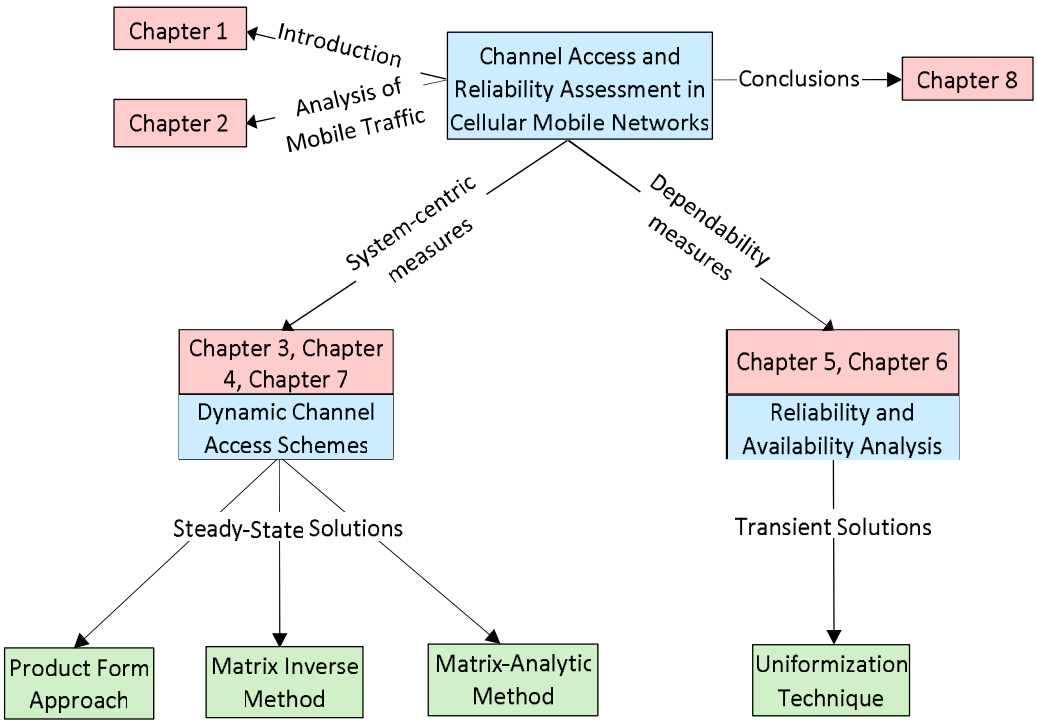

Figure 1.5: The structure of the thesis.

Chapter 2

Analysis and Modeling of Mobile Traffic

"Intellectual growth should commence at birth and cease only at death."
— Albert Einstein

2.1 An Overview of Markov Modeling

Markov analysis has been and continues to be the method of choice for modeling in many diverse fields such as biology, physics, electrical engineering, finance, manufacturing, agriculture, and so on. It is often possible to analyze the behavior of a physical system by describing all the possible states the system may occupy and indicating the evolution of the system by transitions from one state to another in time. The scope of applications of Markovian models to the field of wireless communications and Internet traffic modeling has been on the increase. Because of the memoryless property of Markovian models, these models are amenable to analysis. Markov models attempt to model the activities of a traffic source on a network, by a finite number of states. If the underlying traffic models do not efficiently capture the characteristics of the actual traffic, the result may be the under-estimation or over-estimation of the performance of the network. This would

totally impair the design of the network. Accurate modeling of network events is thus essential to the understanding of system dynamics and to analyze the performance of networks as discussed by Konrad et al. [163].

This chapter gives an overview of Markov processes that are used to model the cellular systems involved in this thesis. The techniques and methods that are utilized to derive the probability vector and the performance measures of interest are provided in the latter part of this chapter. More in-depth treatments of the topics covered here can be found in classical queueing theory books, including by Kleinrock [164], Karlin and Taylor [165], Medhi [166] and Shortle et al. [167].

2.1.1 Markov Chains

A family of random variables $\{X(t), t \in T\}$, defined on a given probability space and indexed by the parameter $t \in T$ is called a *stochastic process* (or random process, or random function). The *parameter space* or index set $T \subset \mathbb{R}$ is sometimes also called the time range. If the parameter space T is countable, for e.g., if we let $t = 0, 1, 2, ...$, then we have a discrete-(time) parameter stochastic process while on the other hand, if T is an interval or an algebraic combination of intervals, we call the process as continuous-(time) parameter stochastic process.

The values assumed by the random variable $X(t)$ denotes the state of the process (observation) at time t and the set of possible states of the process constitutes its *state space*, S. Again, the (real valued) state space can also be either discrete if it is countable (finite or countably infinite); otherwise, it is continuous. Accordingly, a stochastic process can be classified based on the continuous or discrete nature of its parameter space and state space.

So far we have discussed about one-dimensional processes, but the values assumed by the random variable $X(t)$ may be multi-dimensional, and hence one can similarly have multi-dimensional processes. In many queueing systems, multi-state, multi-dimensional Markov chains are required to accurately model the queueing dynamics. An example of a two-dimensional stochastic process in continuous time having continuous state space

is $X(t) = (X_1(t), X_2(t))$ where $X_1(t)$ represents the minimum and $X_2(t)$ represents the maximum temperature at a particular place in an interval of time $(0, t)$.

Any stochastic process satisfying the so-called "Markov property" is said to be a *Markov process*. Specifically, a continuous-time stochastic process is a Markov process if for all integers $k \geq 1$, $t_1 < t_2 < \cdots < t_k$ in the index set and for any real numbers $x_1, x_2, ..., x_k$, we have

$$Pr\{X(t_k) \leq x_k | X(t_1) = x_1, ..., X(t_{k-1}) = x_{k-1}\} = Pr\{X(t_k) \leq x_k | X(t_{k-1}) = x_{k-1}\} \tag{2.1}$$

The above equation (2.1) is called as Markov property which says that given the current state of the process, the next state is independent of any past state taken by the process, and the process thus is said to exhibit memoryless property.

In case the state space of a Markov process is discrete, the Markov process is referred to as a *Markov chain*. Again, depending on the nature of the time range, the Markov chain is classified as a *discrete-time Markov chain* (DTMC) or *continuous-time Markov chain* (CTMC), and the latter is utilized for the analysis of cellular systems in this thesis.

Furthermore, for a Markov chain to satisfy the Markov property, the time spent in any of its states (generally referred to as the sojourn time) must possess the memoryless property. That is, at any time t, the remaining time until the event occurs must be independent of the time already spent in that state. It follows that for CTMCs, the sojourn time must be exponentially distributed as exponential distribution is the only continuous distribution to possess memoryless property.

2.1.2 Relevant Probability Distributions

In this section, probability distributions that are used in most stochastic models are defined and their associated basic properties are presented.

2.1.2.1 Exponential Distribution

A continuous random variable X has an exponential distribution with parameter $\lambda > 0$, denoted by $Exp(\lambda)$ if its probability density function (pdf) is given as

$$f(t) = \begin{cases} \lambda e^{(-\lambda t)}; & t \geq 0 \\ 0; & t < 0. \end{cases} \tag{2.2}$$

The mean and variance are then given by

$$E(X) = \mu_X = \frac{1}{\lambda}, \;\; Var(X) = \sigma_X^2 = \frac{1}{\lambda^2}, \tag{2.3}$$

and hence the coefficient of variation, $c_X = \frac{\sigma_X}{\mu_X} = 1$.

An important property for an exponential random variable X is the *memoryless property*. Mathematically, this property states that for $t \geq 0$ and $s \geq 0$,

$$P\{X > t + s | X > s\} = P\{X > t\}. \tag{2.4}$$

Suppose that X denotes the time. Intuitively, the memoryless property says that given $X > s$ the residual time $X - s$ is independent of the time s that has elapsed. For instance, He [168] discussed that a used (but not too old) light bulb may be as good as a new one, as long as it still works. This implies that the lifetime of a light bulb may possess the memoryless property. Consequently, the lifetime distribution of a light bulb can be approximated by the exponential distribution. Other examples of the exponential distribution include the lifetime of electronic components (e.g., resistors) and the interarrival times of customers at a service facility.

Note that exponential random variables are the only continuous random variables that possess the memoryless property. Further, the following result also plays a key role for stochastic systems in which exponential distribution is involved.

If $\{X_j, j = 1, ..., n\}$ are independent exponential random variables with parameters $\{\lambda_j, j = 1, ..., n\}$, respectively, then $\min\{X_1, ..., X_n\}$ is exponentially distributed with parameter $\lambda_1 + \cdots + \lambda_n$.

2.1.2.2 Erlang-k Distribution

A continuous random variable X is referred to as a k-th order Erlang (or Erlang-k) random variable with parameters (λ, k), denoted by Erl-$k(\lambda)$ for $\lambda > 0$ and $k \in \{1, 2, ...\}$, if its pdf is given by

$$f(t) = \begin{cases} \frac{\lambda^k t^{k-1}}{(k-1)!} e^{-\lambda t}, & t \geq 0; \\ 0, & t < 0. \end{cases} \tag{2.5}$$

For this distribution, we have

$$\mu_X = \frac{k}{\lambda}, \quad \sigma^2(X) = \frac{k}{\lambda^2}, \tag{2.6}$$

and the coefficient of variation c_X is always less than or equal to 1.

The Erlang random variable X is the sum of k independent random variables $X_1, ..., X_k$ having a common exponential distribution with mean $1/\lambda$. For instance, in VoIP the service process consists mainly in transferring the information of caller to another end user. This service process of VoIP proceeds through three phases ($k = 3$): (i) connection establishment, (ii) transferring information, and (iii) accessing differentiated services thus, can be modeled using Erlang-3 distribution. Note that when $k = 1$, the Erlang and exponential distribution coincides.

2.1.2.3 Hyperexponential-k Distribution

A continuous random variable X follows a k-phase hyperexponential distribution with parameters (α_i, λ_i), denoted by Hyp-$k(\alpha_i, \lambda_i)$, for $i = 1, 2, ..., k$ and $\sum_{i=1}^{k} \alpha_i = 1$, if it has the pdf as

$$f(t) = \begin{cases} \sum_{i=1}^{k} \alpha_i \lambda_i e^{(-\lambda_i t)}; & t \geq 0, \alpha_i > 0, \lambda_i > 0 \\ 0; & t < 0. \end{cases} \tag{2.7}$$

The mean and variance are given by

$$\mu_X = \sum_{i=1}^{k} \frac{\alpha_i}{\lambda_i}, \quad \sigma_X^2 = \sum_{i=1}^{k} \frac{2\alpha_i}{\lambda_i^2} - \left(\sum_{i=1}^{k} \frac{\alpha_i}{\lambda_i} \right)^2, \tag{2.8}$$

and the coefficient of variation c_X of this distribution is always greater than or equal to 1.

If a process consists of parallel phases, instead of sequential phases and the system randomly selects one of the phases to process each time according to specified probabilities, each of which has exponential distribution, then the resulting distribution is hyperexponential. For instance, in the context of telephony, where, if someone has a modem and a phone, their phone line usage could be modeled as a hyperexponential distribution where there is probability α_1 of them talking on the phone with rate λ_1 and probability $(1 - \alpha_1)$ of them using their Internet connection with rate λ_2.

2.1.2.4 Phase Type Distribution

Phase type (PH−) distributions were introduced by Neuts [169] as a generalization of the exponential distribution. The above three discussed probability distributions are special cases of PH−distributions that are useful in queuing theory. The primarily theoretical utility of PH−distribution is that any distribution on non-negative real numbers can be approximated by a PH−distribution and the resulting queueing models can be analyzed without losing computational tractability.

For PH−distribution, one can think of a queueing model with finite capacity. Here, customers which upon arrival find no waiting space in the system, are blocked and are called lost customers. We define the state when the arriving customer is blocked as an absorbing state. Now, from the system administrator's viewpoint, in order to keep track of the blocking of a customer, it is required to know the distribution of the time till the first loss, i.e., the time until the system enters the absorbing state. This problem can be addressed by the concept of PH−distribution.

Consider a CTMC $\{X(t),\ t \geq 0\}$ with finite state space of m transient states and one absorbing state $\{1, 2, ..., m, m + 1\}$ and infinitesimal generator

$$Q - \begin{pmatrix} T & \boldsymbol{T}^0 \\ 0 & 0 \end{pmatrix}, \tag{2.9}$$

where T is a square matrix of order m and satisfies $(T)_{ii} < 0$, for $1 \leq i \leq m$ and $(T)_{ij} \geq 0$, for $1 \leq i \neq j \leq m$. Since the row sums of an infinitesimal generator are zero,

the column vector $\boldsymbol{T}^0$ is such that $\boldsymbol{T}^0 = -T\boldsymbol{e}$, where the vector $\boldsymbol{e}$ is a column of 1's of appropriate dimension. The assumption that the first m states are transient implies that the matrix T is non-singular, as mentioned in the study of Latouche and Ramaswami [170]. Note that starting from any initial state, the absorption into the state $m+1$ is certain if and only if the matrix T is non-singular (see e.g., Asmussen [171]).

Let the initial distribution for the CTMC is $(\boldsymbol{\alpha}, \alpha_{m+1})$ with $\boldsymbol{\alpha}$ being a row vector of dimension m and $\alpha_{m+1} = 1 - \boldsymbol{\alpha}\boldsymbol{e}$. Define

$$Z = min\{t : X(t) = m+1,\ t \geq 0\}, \tag{2.10}$$

which denotes the absorption time of state $m+1$ of the CTMC $\{X(t),\ t \geq 0\}$. Then Z is a continuous random variable taking nonnegative values with probability distribution function $F(t)$ given by

$$F(t) = P\{Z \leq t\} - P\{X(t) = m+1\} = 1 - \boldsymbol{\alpha}e^{Tt}\boldsymbol{e},\ t \geq 0. \tag{2.11}$$

Such a variable Z denoting time until absorption is said to follow a PH−distribution with representation $(\boldsymbol{\alpha}, T)$.

The dimension m here is said to be *order* of the distribution $PH(\boldsymbol{\alpha}, T)$. The states $\{1, 2, ..., m\}$ are called *phases* and the CTMC $\{X(t),\ t \geq 0\}$ is usually called the *underlying Markov chain.* The density function of Z is given by

$$f(t) = \boldsymbol{\alpha}e^{Tt}\boldsymbol{T}^0,\ t \geq 0 \tag{2.12}$$

where the matrix exponential of matrix T is given by

$$e^{Tt} = \sum_{n=0}^{\infty} \frac{t^n}{n!} T^n. \tag{2.13}$$

The k^{th} (noncentral) moment of $F(t)$ can be obtained by

$$E(Z^k) = \mu_k = k!\boldsymbol{\alpha}(-T)^{-k}\boldsymbol{e},\ k \geq 0. \tag{2.14}$$

Moreover, the distribution $F(t)$ has a jump of height α_{m+1} at $t = 0$, but in most cases,

it is assumed that the probability of process starting in the absorbing state is zero i.e., $F(0) = 0 = \alpha_{m+1}$.

Next, we discuss some **particular cases** of PH−distribution and their corresponding PH−representations.

When $\boldsymbol{\alpha} = 1, T = -\lambda$ and $m = 1$, the underlying PH−distribution becomes exponential. The Erlangian distribution where there are k phases in series is a PH−distribution with representation ($\boldsymbol{\alpha}$,T) where

$$\boldsymbol{\alpha} = (1, 0, ...0) \text{ and } T = \begin{pmatrix} -\lambda & \lambda & & & \\ & -\lambda & \lambda & & \\ & & \ddots & \ddots & \\ & & & -\lambda & \lambda \\ & & & & -\lambda \end{pmatrix}_{k \times k}. \tag{2.15}$$

A hyperexponential of order k is a PH−distribution with representation ($\boldsymbol{\alpha}$,T) with

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k) \text{ and } T = \begin{pmatrix} -\lambda_1 & & & \\ & -\lambda_2 & & \\ & & \ddots & \\ & & & -\lambda_k \end{pmatrix}. \tag{2.16}$$

Note: The missing elements in matrix T are all zeros. Throughout this thesis, if an element or a block of elements is missing in a matrix, then it is zero or a block of zeros.

Kronecker compositions

The Kronecker composition of matrices has been researched since the nineteenth century. A lot of interesting properties about its trace, determinant and other decompositions have been discovered, many of them are stated and proven in the basic literature about matrix analysis (see e.g., Horn and Johnson [172, Chapter 4]). Here we provide some details on Kronecker composition of matrices, mainly on Kronecker product and Kronecker sum,

which will then be employed to solve matrix equations in Chapter 7.

If $A = (a_{ij})$ is an $m \times n$ matrix and $B = (b_{ij})$ is a $p \times q$ matrix, then the Kronecker product of A and B, denoted by $A \otimes B$, is the $mp \times nq$ block matrix given by

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{pmatrix}. \tag{2.17}$$

If A and B are square matrices i.e., $m = n$ and $p = q$, then the Kronecker sum of A and B, denoted by $A \oplus B$, is defined as $A \oplus B = (A \otimes I_{p\times p}) + (I_{m\times m} \otimes B)$, where the first identity matrix of dimension p and the second of dimension m ensures that both terms are of dimension mp and can thus be added. For matrices A, B, C and D, Kronecker composition possesses a few useful properties listed below as derived by Graham [173].

1. $A \otimes B \neq B \otimes A$, for $A_{m\times n}, B_{p\times q}$;

2. $(A \otimes B) \otimes C = A \otimes (B \otimes C) = A \otimes B \otimes C$, for $A_{m\times n}, B_{p\times q}, C_{r\times s}$;

3. $(A \otimes B)(C \otimes D) = AC \otimes BD$, for $A_{m\times n}, B_{p\times q}, C_{n\times r}, D_{q\times s}$;

4. $A \otimes (B \pm C) = (A \otimes B) \pm (A \otimes C)$, for $A_{m\times n}, B_{p\times q}, C_{p\times q}$;

5. $(A \pm B) \otimes C = (A \otimes C) \pm (B \otimes C)$, for $A_{m\times n}, B_{m\times n}, C_{p\times q}$;

6. $(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha(A \otimes B)$, for scalar α, and $A_{m\times n}, B_{p\times q}$;

7. trace$(A \otimes B) =$ trace$(B \otimes A) =$ trace(A)trace(B), for $A_{m\times m}, B_{p\times p}$;

8. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, if $A_{m\times m}$ and $B_{p\times p}$ are invertible;

9. $(A \otimes B)^T - A^T \otimes B^T$, for $A_{m\times n}, B_{p\times q}$;

10. $e^{(A\oplus B)} = e^A \otimes e^B$ for $A_{m\times n}, B_{p\times q}$.

2.1.3 Birth-Death Process

A birth-death (BD) process is an important sub-class of CTMCs, often used to model a broad class of simple queueing systems and is very well studied in the probability theory subject (see e.g., Feller [174]). It consists of state space, S as the set (or a subset) of non-negative integers, typically denoting the number of customers in the system. 'Birth' corresponds to customer arrival which increases the state variable by one and 'death' corresponds to customer departure which decreases the state by one. More specifically, after the process finishes its stay in state $i \geq 0$, the process transits from state i to $i+1$ upon occurrence of an arrival or transits to $i-1$ at a departure. The process is specified by birth rates λ_i and death rates μ_i. This leads to the following rate transition matrix of the BD process:

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & \cdots \end{array} \\ \begin{pmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & \\ & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{pmatrix} \end{array}, \tag{2.18}$$

where the unspecified elements are zero. The sojourn time in state i is the minimum of the time until the next arrival or the next departure. Since the time until the next arrival and, independently, the time until the next departure is exponentially distributed with rate λ_i and μ_i respectively thus, the sojourn time is exponentially distributed with parameter $\lambda_i + \mu_i$. The rate transition diagram of the BD process is given in Figure 2.1.

Denote by $P_i(t)$ the probability that the process is in state i at time $t \geq 0$. The probabilities $P_i(t), i \geq 0$ satisfy the following system of linear differential equations:

$$\begin{aligned} P_0'(t) &= -\lambda_0 P_0(t) + \mu_1 P_1(t), \\ P_i'(t) &= -(\lambda_i + \mu_i) P_i(t) + \lambda_{i-1} P_{i-1}(t) + \mu_{i+1} P_{i+1}(t), \quad i \geq 1. \end{aligned} \tag{2.19}$$

The BD process with constant birth and death rates as $\lambda_i = \lambda$ and $\mu_i = \mu$ for all $i \geq 0$, is referred to as homogeneous and inhomogeneous otherwise. In particular, the

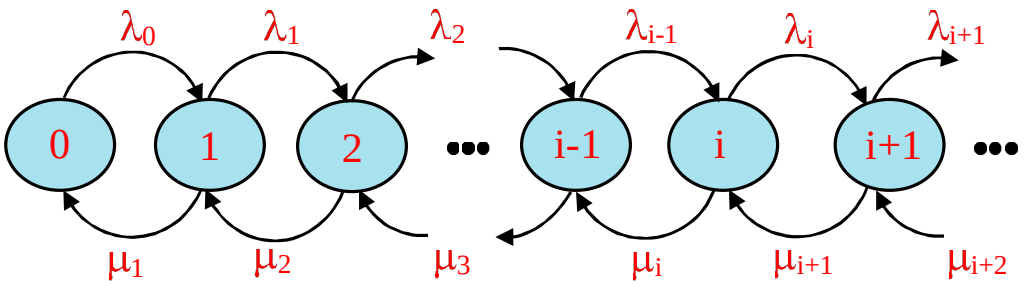

Figure 2.1: Rate transition diagram for a birth-death process.

pure birth process are such that $\mu_i = 0$ and pure death process are such that $\lambda_i = 0$ for all $i \geq 0$. The (homogeneous) Poisson process can be seen as a simplest example of a pure birth process where $\lambda_i = \lambda$ and $\mu_i = 0$ for all $i \geq 0$. Note that in any state, because of the birth-death characteristic, transitions can be made only to the adjacent states however, for more general Markovian models this is not necessarily true.

In order to analyze a stochastic queueing system, we require a useful tool that can provide a good approximation of real-time arrival process and lead to analytically tractable models. The Poisson process and Markovian arrival process (MAP) have proven to be such indispensable stochastic modeling tools. In the following sections, prior to the discussion on MAP, Poisson process and its properties are briefly explained.

2.1.4 Poisson Process

The Poisson process is the simplest one and is most widely used in the modeling of cellular networks. This process is used to count the occurrences of a certain event (arrival) in some time interval. A *counting process* refers to a non-negative, integer-valued, increasing stochastic process.

Consider an arrival counting process $\{X(t), t \geq 0\}$, where $X(t)$ denotes the number of arrivals occurring in $[0, t]$, with $X(0) = 0$. The process $\{X(t), t \geq 0\}$ is a Poisson process with rate $\lambda > 0$ if it satisfies the following three properties:

1. The probability of having one arrival between time t and time $t + \Delta t$ that is, in an interval of length $\Delta t = P\{X(\Delta t) = 1\} = \lambda \Delta t + o(\Delta t)$, where Δt is an

incremental element and

$$lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

2. The probability that more than one arrival occurs between time t and time $t + \Delta t$ — $P\{X(\Delta t) \geq 2\} = o(\Delta t)$.

3. The number of arrivals in non-overlapping intervals are statistically independent; that is, the process possesses the independent increment property. In particular, let $X(s,t) = X(t) - X(s)$ for $t > s$. The counting process $\{X(t), t \geq 0\}$ is said to have *independent increments* if the two increments $X(s,t)$ and $X(u,v)$ are independent for all non-overlapping intervals (s,t) and (u,v).

2.1.4.1 Properties

Assume that $\{X(t), t \geq 0\}$ is a Poisson process with rate λ. Then the process exhibits the following properties.

- $P\{X(t+s) - X(s) = n\} = e^{-\lambda t}(\lambda t)^n / n!, \ n = 0, 1, 2, ...,$ i.e., the random variable denoting the number of arrivals in an interval of duartion t has the *Poisson distribution* with a mean of λt arrivals.

- The process has *stationary increments*; that is, the number of arrivals in intervals of equal width are identically distributed.

- The time between successive arrivals called as *interarrival times* of the process are independent and identically distributed (i.i.d) exponential random variables with parameter λ. Let T be the random variable associated with interarrival times, then T has the exponential distribution with mean $1/\lambda$, where λ is the mean arrival rate (arrivals per unit time).

2.1.5 Markovian Arrival Process

The idea of the MAP is to generalize the Poisson processes in ways that allow the inclusion of non-exponential interarrival time distributions and dependent interarrival times

but still keep the tractability for modeling purposes. As a first step to go beyond the Poisson process, it seems natural to replace the exponential distribution of interarrival times by a general probability distribution and this leads immediately to the class of *renewal processes.* A renewal process is a counting process with i.i.d inter-renewal times, which describe an ordered set of points like, service completion epochs, arrival instants and equipment failure instant on $[0, \infty]$. Their simplifying feature is the independence and equidistribution of successive inter-renewal intervals. A renewal process with phase type distributed inter-renewal intervals is called a **PH renewal process**. Note that the Poisson process is a special PH renewal process where inter-renewal (inter-arrival) times are exponentially distributed.

While in many practical applications, notably in modern communication systems like Internet or other computer networks, arrivals do not usually form a renewal process as there may be a strong correlation between successive interarrival times. In particular, according to Breuer and Baum [175], the arrivals that tend to occur in bursts cannot be modeled with the class of renewal processes. Thus, to introduce a dependence between successive renewal intervals, the MAP as a generalization of the PH renewal process is defined where the correlation aspect is not ignored.

Mathematically, recall from Section 2.1.2.4 and let in a PH renewal process, the inter-renewal times follow a PH(α, T) distribution. Here, instantaneously after the occurrence of an arrival (i.e., a renewal event), the process gets restarted by selecting a new initial state (phase) which is independent of the past phase, using the same probability vector α each time. We then get the infinitesimal generator of PH renewal process as

$$Q = \begin{pmatrix} T & B & & \\ & T & B & \\ & & \ddots & \ddots \\ & & & \ddots \end{pmatrix}, \text{ with } B = \boldsymbol{T}^0\alpha = \begin{pmatrix} T_1^0\alpha \\ T_2^0\alpha \\ \vdots \\ T_m^0\alpha \end{pmatrix}. \tag{2.20}$$

Relaxing this restriction, where choosing a new phase after an arrival depends on the one

immediately before that arrival, we get the generator matrix as

$$Q = \begin{pmatrix} T & A & & \\ & T & A & \\ & & \ddots & \ddots \\ & & & \ddots \end{pmatrix}, \text{ with } A = \begin{pmatrix} T_1^0\alpha_1 \\ T_2^0\alpha_2 \\ \vdots \\ T_m^0\alpha_m \end{pmatrix} \tag{2.21}$$

such that A is non-negative and $\alpha_i \boldsymbol{e} = 1, i = 1, 2, ..., m$, so that still we have $\boldsymbol{T}^0 = -T\boldsymbol{e}$. A Markov chain with such a generator Q gives us a MAP.

MAP is a convenient tool to model both renewal and non-renewal arrivals. However, different from that for a PH−distribution, an underlying Markov chain for a MAP has no absorption state (phase). Here, the arrivals can also occur during the stay in each state of the underlying Markov chain. To be more precise, for a MAP, the transition of states with arrival, the transition of states without arrival and arrivals without a transition of state, are all referred to as events. While MAP is defined for both discrete and continuous times, here we discuss only the continuous-time case which is utilized in this thesis.

2.1.5.1 Construction of Continuous MAP

Let, the MAP is a bivariate Markov process $\{N(t), I(t); t \geq 0\}$ with the state space $S = \mathbb{N} \times \{1, 2, ..., m\}$, where $N(t)$ records the number of arrivals up to time t, while $I(t)$ keeps the track of the state (phase) of the underlying Markov chain. Suppose that $D = (d_{ij})$ denotes the generator of the underlying Markov chain, which is assumed to be irreducible.

At the end of a sojourn time in state $(n, i) \in S$, which is exponentially distributed with parameter $\lambda_i \geq -d_{ii}$, there occurs a transition to another or (possibly) the same phase state. The transition may not correspond to an arrival epoch; that is, the only possible transitions are to the states $\{(n+1, j) : 1 \leq j \leq m\} \in S$ and to the states $\{(n, j) : 1 \leq i \neq j \leq m\} \in S$. The transition rates from state (n, i) are independent of n. More specifically, one of the two events could occur: with probability $p_{ij}(1)$ the transition corresponds to an arrival and the underlying Markov chain is in state j with

$1 \leq i, j \leq m$; with probability $p_{ij}(0)$ the transition corresponds to no arrival and the state of the underlying Markov chain is $j, j \neq i$. Note that $I(t)$ can go from state i to state i only through an arrival. Also, we have

$$\sum_{j=1}^{m} p_{ij}(1) + \sum_{\substack{j=1 \\ j \neq i}}^{m} p_{ij}(0) = 1, \ 1 \leq i \leq m. \tag{2.22}$$

For $k = 0, 1$ and $1 \leq i, j \leq m$, define the matrices $D_k = (d_{ij}(k))$ where the entries $d_{ii}(0) = -\lambda_i$; $d_{ij}(0) = \lambda_i p_{ij}(0)$, for $j \neq i$, and $d_{ij}(1) = \lambda_i p_{ij}(1)$. By assuming D_0 to be a nonsingular matrix, the interarrival times will be finite with probability one and hence the arrival process does not terminate. Thus, D_0 is a stable matrix. It is clear that the generator D is then given by $D = D_0 + D_1$.

The preceding construction shows that the bivariate process $\{N(t), I(t); t \geq 0\}$ is a CTMC with infinitesimal generator

$$Q = \begin{pmatrix} D_0 & D_1 & & \\ & D_0 & D_1 & \\ & & \ddots & \ddots \\ & & & \ddots \end{pmatrix}. \tag{2.23}$$

Note that denoting $D_0 = T$ and $D_1 = A$, the generator of MAP in (2.21) coincides with the one in (2.23). The matrices D_0 and D_1 are called as rate matrices, where D_0 consists of the transition rates without arrivals (except for the diagonal elements) and D_1 consists of the transition rates with arrivals. The sequence of matrices $\{D_k; k = 0, 1\}$ contains all information for Q and thus is usually called the characteristic sequence of the MAP. It follows that the 2-tuple (D_0, D_1) represents the MAP of order m.

If the process $\{I(t) : t \geq 0\}$ is irreducible, then it has a unique stationary distribution say, π. It follows that π is the unique (positive) probability vector satisfying

$$\pi D = \mathbf{0}, \pi \boldsymbol{e} = 1, \tag{2.24}$$

and the fundamental rate of arrival is then defined by $\lambda = \pi D_1 \boldsymbol{e}$, which gives the expected

number of arrivals per unit of time in the stationary version of the MAP.

The construction of MAP is further extended to BMAP (Batch Markovian Arrival Process), which allow arrivals in batches. In BMAP, we have a sequence D_k of matrices where entries of D_0 again corresponds to transition without an arrival and those of D_k corresponds to transitions coupled with a batch arrival of size k $(= 1, 2, ...)$.

MAP is a rich class of point processes, which includes many well-known processes such as Poisson process, Phase type renewal process and Markov-modulated Poisson process (MMPP). Originally, this point process was introduced by Neuts [176] as a versatile Markovian point process (VMPP) with complex notation, which was later simplified by Lucantoni et al. [177]. For more details on MAP and BMAP, one may refer to Artalejo & Gómez-Corral [178]; Chakravarthy et al. [179]; Chakravarthy [180]; Neuts [181]; Lucantoni [182]. Some special cases of MAP are presented below.

2.1.5.2 Special Cases

The following are some well-known processes obtained as particular cases of the MAP.

1. *Poisson process.* The Poisson process with parameter $\lambda > 0$ corresponds to the simplest case where $D_0 = (-\lambda)$, $D_1 = (\lambda)$, $D = (0)$ and $m = 1$.

2. *PH renewal process.* Suppose that the interarrival times in a renewal process are independent and have a common PH–distribution $(\boldsymbol{\alpha}, T)$ with $\boldsymbol{\alpha e} = 1$. Then that renewal process is a MAP with $D_0 - T$ and $D_1 - -T\boldsymbol{e\alpha}$.
   Moreover, this class contains the familiar *Erlang* (E_k) and *Hyperexponential* (H_k) arrival processes as well as finite mixtures of these.

2.1.6 Quasi-Birth-Death Process

The quasi-birth-death (QBD) process, a term that was coined by V. Wallace [183], can be viewed as matrix generalization of (simple) BD process seen earlier. Consider a CTMC $\{X(t), J(t),\ t \geq 0\}$ with two-dimensional state space given by

$$\{(0,1),(0,2),...,(0,m_0)\} \cup \{\{1,2,...\} \times \{1,2,...,m\}\}, \tag{2.25}$$

where m_0 and m are positive integers. Here, the variable $X(t)$ is called the *level variable* and $J(t)$ the *phase variable* i.e., the first coordinate of a state (n, j) denotes the level and the second coordinate j denotes the phase. The number of phases m in each level may be either finite or infinite. The state space can also be partitioned on the basis of levels as $S = \cup_{n\geq 0} l(n)$ where the vector $l(n) = \{(n, j), j = 1, 2, ..., m\}$ for $n \geq 0$ is called the *level n*.

The CTMC is then called a continuous time QBD process if the variable $X(t)$ increases or decreases its value by at most one at each transition. In other words, transition from (n, j) to (n', i) is not possible if $|n-n'| \geq 2$. Besides, if the transition rates are level independent i.e., the transition from (n, j) to (n', i) may depend on j, i, and $n-n'$ but not on specific values of n and n', then the resulting QBD process is called *level independent* QBD process (LIQBD) (see e.g., Neuts [184]). Otherwise, the QBD process is called as a level-dependent QBD process (see e.g., Bright and Taylor [185]). The infinitesimal generator matrix of a LIQBD process has the following structure

$$Q = \begin{pmatrix} B_{00} & B_{01} & & & \\ B_{10} & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{pmatrix}, \tag{2.26}$$

where B_{00} is a square matrix of dimension m_0; B_{01} and B_{10} are rectangular matrices of dimension $m_0 \times m$ and $m \times m_0$ respectively and A_0, A_1 and A_2 are square matrices of order m. Also, we have

$$B_{00}\boldsymbol{e} + B_{01}\boldsymbol{e} - B_{10}\boldsymbol{e} + A_1\boldsymbol{e} + A_0\boldsymbol{e} - A_2\boldsymbol{e} + A_1\boldsymbol{e} + A_0\boldsymbol{e} - 0. \tag{2.27}$$

It is clear that the above generator matrix exhibits a block tridiagonal structure where the block structures correspond to levels and the intra-block structure correspond to phases.

Note that for $n - m - m_0 = 1$, a QBD process reduces to a (simple) BD process discussed in Section 2.1.3. For more details on the QBD process, one may refer to the

book by Bini et al. [186].

2.2 Methodology

The performance evaluation of cellular mobile systems leads to the development of multi-dimensional queueing models. Once the Markov model is developed, the probability vector can be obtained by using different methods. In the case of steady-state solutions, the analysis often involves solving simultaneous linear-algebraic equations while for many transient situations, we are often faced with solving a system of linear first-order differential equations. The following sections illustrate the techniques utilized in this thesis for obtaining such solutions.

2.2.1 Steady-State Solutions

For queues in which a Markov analysis is possible, the steady-state solution can be obtained by solving the stationary equations for a continuous-parameter (time) process. That is, $\boldsymbol{\pi}$ satisfying,

$$\boldsymbol{\pi}Q - 0, \;\; \boldsymbol{\pi}\boldsymbol{e} = 1, \tag{2.28}$$

where $\boldsymbol{\pi}$ is the steady-state probability vector, Q the infinitesimal generator of the CTMC, and $\boldsymbol{e}$ is a column vector of ones. Letting $\boldsymbol{\pi} = (\pi_0, \pi_1, ...)$, π_j gives the (long-term) fraction of time the process will be in state j at an arbitrary time. The $\{\pi_j\}$ here are referred as limiting or steady-state probabilities of the Markov chain.

2.2.1.1 Product Form Approach

For the birth-death process, using the matrix Q given in (2.18), the vector-matrix equation $\boldsymbol{\pi}Q - 0$ can be written in component form as

$$\begin{gathered} -(\lambda_j + \mu_j)\pi_j + \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1} = 0, \quad j \geq 1, \\ -\lambda_0\pi_0 + \mu_1\pi_1 = 0. \end{gathered} \tag{2.29}$$

Then under the stability condition (which is both necessary and sufficient) given by

$$\sum_{j=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} < \infty, \tag{2.30}$$

it follows that the steady-state probabilities of the process

$$\pi_j = \lim_{t \to \infty} P_j(t), \quad j \geq 0, \tag{2.31}$$

are calculated as

$$\pi_j = \pi_0 \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \tag{2.32}$$

with

$$\pi_0 - \left[1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \right]^{1}, \tag{2.33}$$

where $\pi_j,\ j \geq 0$ (typically) denotes the probability of being j number of customers in the system at an arbitrary time.

2.2.1.2 Generating Function Method

Another method that is well suited for queueing models is probability generating function (PGF) technique. It is a useful tool for dealing with discrete random variables. If the sequence $\{\pi_j\}$ is the probability mass function of a random variable X on the nonnegative integers (i.e. $P(X = j) = \pi_j,\ j = 0, 1, 2, ...$), then the PGF of $\{\pi_j\}$ is given by

$$P_X(z) = E(z^X) = \sum_{j=0}^{\infty} \pi_j z^j. \tag{2.34}$$

The convergence of the PGF is guaranteed for all complex numbers z with $|z| \leq 1$. Here, the probabilities $\{\pi_j\}$ can be extracted by using repeated differentiation, as

$$\pi_j - \frac{P_X^{(j)}(0)}{j!},\ j \geq 0. \tag{2.35}$$

Moreover, the PGF is useful for obtaining decomposition results since for two independent random variables, X and Y, we have, $P_{X+Y}(z) = P_X(z) P_Y(z)$.

2.2.1.3 Matrix Inverse Method

Standard inversion techniques are quite adequate for moderately sized systems. Consider the system of equations in (2.28), here one equation of the set $\boldsymbol{\pi}Q = 0$ is always redundant. Incorporating the normalization condition, $\boldsymbol{\pi}\boldsymbol{e} = 1$, the last column of the Q matrix can be replaced by a column of ones and the last 0 element of the vector 0 by a one. It is then required to solve a system of the form

$$\boldsymbol{\pi}A = \boldsymbol{b}, \tag{2.36}$$

where A is the modified Q matrix and $\boldsymbol{b}$ is the modified zero vector i.e., of the form $(0, 0, ...0, 1)$. It suffices to find A^{-1} since the solution vector is then obtained as

$$\boldsymbol{\pi} = \boldsymbol{b}A^{-1}. \tag{2.37}$$

In fact, the last row of the A^{-1} matrix contains the steady-state probabilities $\{\pi_j\}$.

2.2.1.4 Matrix-Analytic Method (MAM)

Since the introduction of the MAMs by M.F. Neuts [187] in the 1970s, they have become an important tool to study the stochastic models, notably the queueing models and reliability models.

Consider a QBD process, using the matrix Q given in (2.26), let $\boldsymbol{\pi}$ be the steady-state probability vector of $A - A_0 + A_1 + A_2$, which is assumed to be irreducible. That is, $\boldsymbol{\pi}$ satisfies,

$$\boldsymbol{\pi}A - 0, \;\; \boldsymbol{\pi}\boldsymbol{e} = 1. \tag{2.38}$$

Suppose that $\boldsymbol{x}$, partitioned as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, ...)$ is the steady-state probability vector of the matrix Q. That is,

$$\boldsymbol{x}Q = 0, \;\; \boldsymbol{x}\boldsymbol{e} = 1. \tag{2.39}$$

Then under the stability condition, $\boldsymbol{\pi}A_0\boldsymbol{e} < \boldsymbol{\pi}A_2\boldsymbol{e}$, the solution probability vector is of the form

$$\boldsymbol{x}_i = \boldsymbol{x}_1 R^{i-1}, i \geq 1, \tag{2.40}$$

where the matrix R is the minimal nonnegative solution to the matrix-quadratic equation:

$$R^2 A_2 + R A_1 + A_0 = 0, \tag{2.41}$$

and $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ can be obtained by solving the first two equations of (2.39):

$$\begin{aligned} \boldsymbol{x}_0 B_{0\mathrm{C}} + \boldsymbol{x}_1 B_{10} &= 0, \\ \boldsymbol{x}_0 B_{01} + \boldsymbol{x}_1 (A_1 + R A_2) &= 0, \end{aligned} \tag{2.42}$$

subject to the (normalizing) condition:

$$\boldsymbol{x}_0 \boldsymbol{e} + \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{e} = 1. \tag{2.43}$$

2.2.2 Transient Solutions

For Markovian queues, if $\pi_j(t)$ denotes the probability that the Markov chain with generator Q is in state j at time t, and $\boldsymbol{\pi}(t)$ is the vector of all such probabilities, then from Chapman-Kolmogorov differential equations

$$\boldsymbol{\pi}'(t) = \boldsymbol{\pi}(t) Q, \tag{2.44}$$

it can be obtained as

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0) e^{Qt}, \tag{2.45}$$

with $\boldsymbol{\pi}(0)$ being the initial probability vector. Here, e^{Qt} is the matrix exponential which is defined by

$$e^{Qt} = \sum_{k=0}^{\infty} \frac{(Qt)^k}{k!}, \; t \geq 0. \tag{2.46}$$

As the complexity of the queueing system increases, one cannot obtain closed form expressions for the transient solution. Hence, the numerical techniques are used to solve the resulting system of equations and to gain an insight into the behavior of system characteristics.

2.2.2.1 Uniformization Technique

Uniformization, also known as the method of randomization, was first described by Jensen [188] in 1953. It is a powerful probabilistic method in the sense that it avoids the use of differential equations as well as matrix exponentials to evaluate the transient solution (see e.g., Neuts [189]). This numerical technique may be applied to the Markov process on a countable state space S as long as the diagonal elements of the Q matrix are bounded, i.e., there exists a Δ such that

$$|q_{ii}| \leq \Lambda < \infty, \; i \in S. \tag{2.47}$$

Letting $\Delta = max_i|q_{ii}|$ and considering the (discretized) one-step transition probability matrix as

$$R = I + Q/\Delta, \tag{2.48}$$

gives

$$e^{Qt} - e^{-\Delta t}e^{(\Delta t)R}. \tag{2.49}$$

Substituting this into (2.45) yields

$$\boldsymbol{\pi}(t) = \sum_{k=0}^{\infty} \boldsymbol{\pi}(0)R^k e^{-\Delta t}\frac{(\Delta t)^k}{k!}, \; t \geq 0, \tag{2.50}$$

which is called the *uniformization equation*. Moreover, while implementing this method, it is required to truncate the above series, i.e., for a given $\epsilon > 0$, choose k^* sufficiently large such that

$$\sum_{k=0}^{k*} e^{-\Delta t}\frac{(\Delta t)^k}{k!} > 1 - \epsilon, \tag{2.51}$$

so that for any consistent vector norm $||\cdot||$, we have

$$\left|\left|\boldsymbol{\pi}(t) - \sum_{k=0}^{k*} \boldsymbol{\pi}(0)R^k e^{-\Delta t}\frac{(\Delta t)^k}{k!}\right|\right|_{\infty} \leq 1 - \sum_{k=0}^{k*} e^{-\Delta t}\frac{(\Delta t)^k}{k!} < \epsilon. \tag{2.52}$$

After obtaining the probability distribution, regardless of the solution technique being used, the next step is to derive the measures of interest.

2.3 Performance and Dependability Measures

When a cellular mobile system is being designed and implemented, it becomes essential to answer "what-if" questions and carry out trade-off studies to choose between a set of contending design alternatives. The broad classes of measures that need to be evaluated are:

- *Dependability*: Dependability evaluation typically accounts for failure and repair characteristics of the system. Dependability analysis endeavors to answer the question "Does the system work, and for how long ?". Metrics that measure the availability and reliability aspects of a network from the dependability theory perspective include the following.

    1. *Channel Availability*: The channel availability represents the probability with which the network will allocate a channel to a new user arrival without blocking its request. Accordingly, it is considered as the main QoS measure for newly arriving users.
    2. *Retainability*: The probability of successful completion of an already commenced connection is referred to as service retainability. This metric is considered more important for ongoing users.
    3. *Unserviceable Probability*: The network unserviceable probability is defined as the probability that a (newly arriving or ongoing) service cannot be completed successfully. This metric is important to be determined to evaluate the overall satisfaction of the network.

- *Performance*: Performance analysis involves the computation of system-centric metrics. The issue addressed by performance measures is "Given that the system works, how well does it work ?". Different metrics may be relevant in the context

of different systems. Metrics that measure system-centric performance include the following.

1. *Blocking Probability*: The blocking probability is defined as the probability that an incoming user will be denied service due to lack of idle channels.
2. *Dropping Probability*: The probability of dropping or forced termination represents the probability that a service in progress is forced to terminate before its communication is finished.
3. *Capacity*: The capacity of a service in a network can be defined as the average number of service completions per unit time.
4. *Spectrum Utilization*: The ratio of the average number of utilized channels to the total number of channels available is referred to as spectrum utilization.
5. *Mean Number in System*: The average number of calls waiting in the queue including the calls being served at a particular time represents the average number of calls in the system.
6. *Probability of Idle System*: The probability of a system being idle is defined as the probability of the system having no user requests.

In this thesis, both system-centric metrics and dependability metrics are evaluated considering different spectrum management schemes. However, there may exist some other metrics which do not lie in any of the above two classes. Regarding system-oriented metrics, more details are provided in Chapters 3, 4 and 7 while reliability-oriented metrics are covered in Chapters 5 and 6. Moreover, we develop MATLAB programs to obtain such required metrics and the numerical results of the traffic models proposed in this doctoral work. Note that the expressions for analyzing the performance measures for each model are derived as general expressions and therefore they are applicable to any scenario within the proposed paradigms.

Chapter 3

Performance Evaluation of Admission Control Based on Signal Quality

"*Quality in a product or service is not what you put into it. It is what the client or customer gets out of it.*"

— *Peter F. Drucker*

3.1 Introduction

The evolving 5G and beyond cellular networks are envisioned to provide new outstanding attributes such as higher mobility, higher support to traffic demands and higher levels of the quality of service (QoS), as discussed by Ojijo and Falowo [190]. Accordingly, to handle the tremendous growth in traffic with limited radio spectrum available, it is imperative to develop efficient call admission control (CAC) schemes that can provide guaranteed QoS to the end users. As we discussed in Chapter 1, in order to sustain the provided QoS to the users and to support users' mobility, the handoff (handover) mechanism is a key element of wireless cellular networks. The mobility of users is primarily responsible for

---

This work (partially) has appeared in **R. Kulshrestha, M. Jain and Shruti, Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, Springer, 90 (2020) 739-747.** (https://doi.org/10.1007/s40010-019-00635-2) and optimization part has been accepted in **R. Kulshrestha and Shruti, International Journal of Mathematics in Operational Research (2020).**

initiating a handoff, which makes a CAC much more challenging in cellular networks.

When the mobile terminal (MT) moves out of the coverage area of a particular base station (BS), the received signal strength becomes weak and the present cell site requests a handoff (see e.g., Lee [191]). Thus, for the successful implementation of the handoff process, the system designers specify an optimum signal level, at which handoff is initiated. For the handoff decision mechanism, some measurements such as relative signal strength (RSS), RSS with threshold, RSS with hysteresis and threshold are considered. Several types and phases of the handoff procedure in cellular systems have been described by Kumar and Purohit [75]. Recently, Al-Rubaye et al. [192] developed a CAC function to adjust thresholds during handoff request signaling.

During mobility of an MT, it is possible that the BS system which is providing service may no longer be capable to give desired signal quality for the service as compared to another neighboring BS system. Therefore, the BS system may decide to handoff the call to a neighboring BS system with good signal quality, instead of dropping the call. In the mobile assisted handoff (MAHO) scheme, decision related to handoff is made based on the received signal strength indicator (RSSI) and bit error rate (BER) of transmissions from neighboring BS systems. The combined MAHO and guard channel (GC) scheme presented by Madan et al. [193] ensures that the handoff call meets the acceptable signal quality standard and availability of channels in the neighboring BS system. It is not always possible to have adequate signal quality when an ongoing call is being handoff in a cellular mobile system. As the MT approaches the BS, the weak signal quality handoff request gets improved, as mentioned by Ujarari and Kumar [194]. The mobile controlled handoff (MCHO) scheme provides the opportunity to the MT for choosing BS with good quality signal out of neighboring BSs in its vicinity. This type of handoff has a short reaction time and is easy to implement.

The evaluation of the performance of CAC schemes is usually done from the aspects of call blocking probability, call dropping probability and bandwidth utilization. The objective is to minimize new call blocking probability, to reduce handoff call dropping probability and to maximize the utilization of limited available bandwidth. A lot of re-

search has been carried out in the literature on the performance evaluation of prioritized CAC schemes, including the works of Vazquez-Avila et al. [62], Jain and Mittal [195], and Abdulova and Aybay [196]. However, the majority of CAC schemes proposed by researchers do not take the quality of signal into account. Goswami and Swain [197] investigated the effect of signal quality while evaluating the network performance but in a limited manner. In today's cellular networks, the quality of communication mainly refers to the measurement of a system with service availability, quality of signal captured and minimum delay. Since deterioration in the quality of received signals may result in call drop therefore, the significant impact of signal quality on the system performance cannot be ignored. Consequently, in a good CAC design, it is of paramount importance to control the quality of signal for better performance as suggested by Ahmed [68].

With this motivation, in this chapter, we propose two new CAC schemes to quantify the impact of signal quality on cellular networks. The analytical models are developed utilizing continuous-time Markov chains (CTMCs) to evaluate the proposed CAC schemes. We obtain explicit expressions for the stationary probabilities, which in turn, yield performance measures. Additionally, a non-linear optimization problem is formulated to determine the optimal value of new calls' acceptance probability while satisfying the QoS requirements simultaneously.

The rest of this chapter is organized as follows. In Section 3.2, the system description of cellular network together with the assumptions is given. Section 3.3 presents analytical Model I based on one-dimensional CTMC followed by important performance measures. In Section 3.4, we obtain the stationary solution and derive the performance measures for Model II based on two-dimensional CTMC. Following this, Section 3.5 describes an optimization problem and presents an algorithm to find the optimal value of acceptance probability. Numerical results are computed in Section 3.6 to validate the proposed models and to see the effect of various parameters on the performance measures. Finally, Section 3.7 concludes the chapter.

3.2 System Description and Assumptions

In cellular networks, a given geographical area is divided into a certain number of cells. We assume that the cells are statistically homogeneous. A particular cell with an infinite population of users from a homogeneous cellular network (i.e., cells in the network with the same capacity, performance and characteristics) is considered. The system model for a single cell is given in Fig 3.1.

Many CAC schemes have been proposed in the literature which varies in terms of complexity from simple non-prioritized schemes to more complex schemes that dynamically assigns the priorities based on the measurements (e.g., see Cruz-Pérez et al. [198]; Tung et al. [199]). Due to the fact that a fixed number of channels are reserved for handoff calls in the popular GC scheme, the handoff dropping and new call blocking probability vary largely with the change in the number of reserved channels. Whereas the fractional guard channel (FGC) policy also admits new calls with a certain probability depending on the number of busy channels and accepts a handoff call as long as there are some vacant channels available. That is, it provides suitable QoS for handoff users while keeping new call blocking probability at a reasonable level. Such a probabilistic call admission prevents the system from approaching congestion and we, therefore, adopt the FGC policy in the present work. Moreover, for more realistic performance evaluation models, the queueing scheme is coupled with the FGC policy. In case, all the channels are found to be occupied, the handoff calls that would otherwise be forcibly dropped are allowed to be queued with a certain probability for later transmissions and possibly served later. Moreover, to develop the analytical models, the following assumptions are made as the basis.

- The total S channels are allocated to each cell to serve the incoming calls and there is also the provision of finite buffer of size N to accommodate good quality handoff calls.

- It is assumed that both new calls and handoff calls are generated according to a