Modelling and Performance Analysis of Some Communication Related Queuing Systems

A thesis submitted

in partial fulfillment for the award of the degree of

Doctor of Philosophy

by

Sweta Dey

Department of Mathematics

Indian Institute of Space Science and Technology

Thiruvananthapuram, India

June 2020

Certificate

This is to certify that the thesis titled ***Modelling and Performance Analysis of Some Communication Related Queuing Systems*** submitted by **Sweta Dey**, to the Indian Institute of Space Science and Technology, Thiruvananthapuram, in partial fulfillment for the award of the degree of **Doctor of Philosophy** in **Mathematics**, is a bonafide record of the original work carried out by her under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Deepak T. G.
Associate Professor
Dept. of Mathematics

Dr. SABU. N.
Professor and Head
Dept. of Mathematics

Place: Thiruvananthapuram
Date: 12 June 2020

Declaration

I declare that this thesis titled ***Modelling and Performance Analysis of Some Communication Related Queuing Systems*** submitted in partial fulfillment for the award of the degree of **Doctor of Philosophy** in **Mathematics** is a record of original work carried out by me under the supervision of **Dr. Deepak T. G.**, and has not formed the basis for the award of any degree, diploma, associate-ship, fellowship, or other titles in this or any other Institution or University of higher learning. In keeping with the ethical practice in reporting scientific information, due acknowledgments have been made wherever the findings of others have been cited.

Place: Thiruvananthapuram
Date: June 2020

Sweta Dey
(SC15D020)

To my brother Bithin, for everything...

Acknowledgements

This thesis becomes a reality with the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I would like to start by thanking my parents for being with me at every moment, standing as my support system throughout my life. Nothing was possible without them. I owe my deepest gratitude to them. I would like to express my sincere gratitude to my supervisor Dr. Deepak T.G. for the continuous support during my research, for his patience, motivation, enthusiasm and immense knowledge. Without his guidance, fatherly affection this thesis would not have been possible. I could not have imagined having a better supervisor and mentor for my Ph.D. study.

Besides my supervisor, I would like to thank the rest of my Doctoral Committee members: Prof. A Krishnamoorthy, Prof. U.C. Gupta, Prof. Raju K George, Prof B.S. Manoj, and Prof Anil Kumar C.V. for their insightful comments and continuous assessment in the progress of this work. I should mention the motivation and support I have received from all the faculty members, all the office people, Department of Mathematics.

I am thankful to have awesome fellow researchers who were always ready to discuss many research problems and to share new ideas. I would like to thank all my hostel mates, all close friends specially Akhil, Sagnik, Jyoti, Nitesh to be always there when I needed them most.

My gratitude also goes to the authorities of IIST and Department of Space, Government of India, for providing all the facilities and financial support.

Thanks to almighty for giving all the strength and ability to understand, learn and complete this thesis.

Sweta Dey

Abstract

A queuing system can be described as arrival of customers for service; if the service is not immediate, then waiting for service and leaving the system after being served. The models that we consider here are intended for studying dynamics and performance characteristics of some queueing systems that are related to communication, specifically some wireless network and two-way communication (call centre situation) models.

In the first half of the thesis, we mainly deal with two wireless network models. The first model is developed to study the queueing characteristics of nodes in a wireless network, where the channel access is governed by binary exponential back off (BEB) rule based on CSMA/CA protocol. In the second model, apart from the assumptions put forth in the first model, we take into consider data packets that are of emergency in nature. The speciality here is that these packets have to be transmitted within a random amount of time after they are being generated. Otherwise, their relevance will be lost and hence such packets are assumed to be dropped.

The second half of the thesis deals with some models related to two-way communication, especially some call centre models. Some works have already been reported in this direction earlier. Here, we consider some variants of the said models by incorporating multi-class incoming calls with or without outgoing calls. A detailed analysis by regenerative approach has been carried out under general distribution assumptions for call processing times. Further, these models have been extended to multi-class orbital calls under balking set up and we have offered some rigorous mathematical treatment to derive some important system performance measures that are helpful for system design. In the final chapter, another variant of these models comprising server vacation is analysed and steady state distribution and some measures of effectiveness are computed by combining matrix analytic and regenerative approaches in Markovian set up.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Queuing theory

Queueing theory plays an important role in many areas for modelling systems where customers must line up (queue) for service (use of resource) such as restaurants (tables and seats), banks (tellers), networks (web server, router, WLAN), computer systems (CPU, disk I/O), etc. By spending each moment in a system waiting for any kind of service, a customer becomes part of a queue. The total time spent in the system includes the actual service time and the time spent before the customer gets into the service. The study of such waiting line problems comes under the topic of queuing theory.

Erlang's problem

At the beginning of the twentieth century, while working in a telephone company in Copenhagen, Danish Mathematician A.K. Erlang observed that a telephone system was generally characterized by either Poisson input, exponential holding times, multiple channels or Poisson inputs, constant holding times and a single channel. In 1909, he published his monumental work in this direction (see Erlang [34]).

During that time, a phone call was the realisation of a connection between a caller and receiver. The connection used to happen using a circuit board on the links between these two people involved in that particular call. When the call was in progress, the circuit remained occupied. When all the circuits were occupied, if a new call had been attempted, then this call was rejected. Local communities were connected by one board of circuits. Erlang's job was to determine the number of circuits to ensure a certain service level, given by the probability that a client is rejected. While designing the system, he published the historical work [34], which

showed that the number of calls coming in follows a Poisson distribution. This approximation is used to calculate system performance and is used until today in many areas. In continuation to this, Erlang published another remarkable work in 1917 (see Erlang [35]), where he observed that the duration of phone calls was exponentially distributed. Also, he was responsible for the concept of stationary equilibrium.

With this tremendous contribution, the revolution started in the history of queuing theory. In 1927, E.C. Molina [55] published a work related to telephone trunking problems, which was an extension of Erlang's work. In 1930, Felix Pollaczek [64] published some fundamental work on Poisson input, arbitrary output, and single/ multiple channel problems. In the same decade, Kolmogrov [49], Khintchine [45], and some other researchers took this a step forward. Kendall ([43], [44]) gave an insight into the stochastic process occurring in the theory of queues. In 1955, Cox [24] analyzed the congestion problems statistically. Mathematical methods in the theory of queues were discussed by Khintchine [46] in 1960.

After giving a brief history of the development of the queuing theory, we now present a detailed description of a queuing system.

1.1.1 Characteristics of queuing processes

The basic features that provide a description of queuing systems are as follows: (See Gross et al. [40])

a) **Arrival pattern:** In a queuing system, it is necessary to know in which pattern customers arrive into the system i.e. along with knowing the distribution of time between successive customers, we should also know whether the customers can arrive simultaneously or in batches.

Sometimes, even after seeing a long queue, the customer decides to wait in the queue. But in some cases, it may decide not to join the queue and leave the system. Such a customer is called a *balking customer*. Even after joining the queue, some lose their patience and leave the system. They are said to have *reneged*. If there are more queues, customers may have a tendency to switch from one queue to another, which is named as *jockeying*.

One important factor to be considered regarding the arrival pattern is the manner in which the pattern changes with time. If an arrival pattern does not change with time, it is called a stationary arrival pattern. The pattern which is not

time-independent is called non-stationary.

b) **Service pattern:** This indicates the manner in which service is rendered. A probability distribution is needed to describe service. Like arrivals, service can be provided in singles or batches. Generally, one customer is being served by a server. But there are cases where more than one customer can be served simultaneously by the same server (example- computer with parallel processing).

Sometimes, service process will be dependent on the number of customers waiting for service. In case of a long queue, server may work faster and similarly, server may become less efficient if there is less work. This kind of service is called state dependent service.

Like arrivals, service can be stationary and non-stationary with respect to time. A queuing system can be both non-stationary and state-dependent.

Services of customers may be delayed by waiting in the line even if the service rate is high. Generally, customers' arrival and departure happen at irregular intervals, so the queue length does not assume any definite pattern. In fact, the probability distribution for queue length is the result of both arrival and service process.

c) **Queue discipline:** This refers to the manner in which customers are selected for service. Some common disciplines are first come first served (FCFS), last come first served (LCFS), service in random order (SIRO) etc.

A variety of priority schemes are also included in queue discipline. The ones with higher priorities will be selected ahead of customers having lower priorities. If the customer with the highest priority is allowed to enter the service upon arrival, even if a lower priority customer is in service, it is called preemptive. In non-preemptive case, the highest priority customer will go to the head of the queue but cannot get the service until the service of the present customer is completed.

d) **System capacity:** The capacity of the system is the maximum number of customers that can be accommodated in the system. It can be finite or infinite. For a finite capacity system, no customer is allowed to enter the system once it reaches the maximum capacity.

e) **Number of service channels:** By specifying the number of service channels, we actually define the number of parallel servers that can serve the customers

simultaneously. There can be many variations of multi-channel systems.

Some basic concepts which we use to model a queuing system are given in the following sections.

1.2 Mathematical framework

Let $\{X_n : n \in N\}$ (discrete time) or $\{X_t : t \geq 0\}$ (continuous time) be a stochastic process defined on the probability space (Ω, U, P) assuming values in a countable set (i.e state space) E.

1.2.1 Markov process

Definition 1.2.1. *The discrete time Process $X = \{X_n : n \in N\}$ is called a Markov chain or discrete time Markov chain (DTMC) if*

$$P\{X_{n+1} = j | X_0, ..., X_n\} = P\{X_{n+1} = j | X_n\}$$

for all $j \in E$ and $n \in N$.

Hence a Markov chain is characterised by the property that its future probabilistic behaviour is conditionally independent of the past if the present is known. Transition probabilities of a time homogeneous Markov chain are defined as

$$P\{X_{n+1} = j | X_n = i\} = P(i, j), \quad i, j \in E.$$

If the state space is $E = \{0, 1, ...\}$ then we denote P as transition probability matrix whose (i, j)th entry is given by $P(i, j)$.
The probability that the chain moves from state i to state j in m steps is given by the (i, j)-entry of the mth power of the transition matrix P, i.e,

$$P\{X_{n+m} = j | X_n = i\} = P^m(i, j), \quad i, j \in E.$$

A state i is said to be *recurrent* if and only if starting from state i, the probability of returning to i after some finite time is 1. Otherwise, it is called *transient*. For a recurrent state, if the mean occurrence time is finite, it is called *positive recurrent*. Otherwise, it is called *recurrent null*.

The greatest common divisor of the recurrence times of a state is said to be its *period*. If the period is one, the state is said to be *aperiodic*. If all the states of a Markov chain are aperiodic, then the chain is said to be *aperiodic*.

A subset of the state space is said to be *closed* if no state outside it can be reached from any state in it. If a state forms a closed set by itself, then it is called *absorbing state*. If no proper subset of a closed set is closed, then that set is called *irreducible*. A Markov chain is called *irreducible* if its only closed set is the state space.

Theorem 1.2.1. *Suppose X is irreducible and aperiodic. Then all states are recurrent non-null iff the system of linear equations*

$$\begin{aligned} \pi(j) &= \sum_{i \in E} \pi(i) P(i,j), \quad j \in E \\ \sum_{j \in E} \pi(j) &= 1 \end{aligned}$$

has a unique solution π. If there exists a solution π, then it is strictly positive, there are no other solutions, and we have $\pi(j) = \lim_{n\to\infty} P^n(i,j)$ for all $i, j \in E$.

Definition 1.2.2. *The continuous time stochastic process $Y = \{Y_t : t \in R_+\}$is said to be a Markov process or continuous time Markov chain (CTMC) with state space E if for any $t, s \geq 0$ and $j \in E$,*

$$P\{Y_{t+s} = j | Y_u; u \leq t\} = P\{Y_{t+s} = j | Y_t\}.$$

The conditional probability stated above is dependent on both t and s. When

$$P\{Y_{t+s} = j | Y_t = i\} = P_s(i,j)$$

is independent of $t \geq 0$, for all $i, j \in E$, then the process Y is said to be a *time-homogeneous* Markov process. The function $t \to P_t(i,j)$ is called *transition function* of the Markov process Y.

Let us define W_t as the length of time the process Y is in the state which is being occupied at the instant t, i.e

$$W_t(\omega) = \inf \{s > 0 : Y_{t+s}(\omega) \neq Y_t(\omega)\}.$$

Theorem 1.2.2. *For any $i \in E$ and $t \geq 0$,*

$$P\{W_t > u | Y_t = i\} = e^{-\lambda(i)u}.$$

If $\lambda(i) = 0$, the state is called *absorbing* and *instantaneous* if $\lambda(i) = \infty$. Else the state will be called *stable* (i.e when $0 < \lambda(i) < \infty$).

Structure of a Markov process: Let Y be a Markov process with state space E and standard transition function P_t. Let $T_0, T_1, T_2, \ldots$ be the instants of state change for the process Y and $X_0, X_1, X_2, \ldots$ be the successive states visited by Y. If W_t denotes the waiting time from t until the next change of state, then

$$T_0 = 0; \quad T_{n+1} = T_n + W_{T_n}, n \in N$$

$$\text{and } X_n = Y(T_n), n \in N.$$

Theorem 1.2.3. $P\{X_{n+1} = j, T_{n+1} - T_n > u | X_0, ..., X_n = i; T_0, ..., T_n\}$
$= Q(i,j)exp(-\lambda(i)u)$, *where* $Q(i,j) = P\{X_{n+1} = j | X_n = i\}$

It shows that $\{X_n : n \in N\}$ is a Markov chain and from this we can show that

$$P\{T_{n+1} - T_n > u | X_n = i, X_{n+1} = j\} = e^{-\lambda(i)u}.$$

The generator/transition rate matrix of this Markov process $\{Y_t\}$will be given by

$$A(i,j) = \begin{cases} -\lambda(i) & if\ i = j \\ \lambda(i)Q(i,j) & if\ i \neq j \end{cases}$$

and from the relation $P_t = exp(tA)$ we can find the transition probability matrix of this Markov process.

If π is the steady-state vector (i.e $\pi(j) = \lim_{t\to\infty} P_t(i,j)$) of the Markov process $\{Y_t\}$, then

$$\pi A = \mathbf{0}, \pi \mathbf{e} = 1.$$

1.2.2 Renewal process

Definition 1.2.3. *Consider a fixed phenomenon, and let $W_1, W_2, ...$ be the times between its successive occurrences. Then we define the time of occurrences as*

$$S_0 = 0;\ S_{n+1} = S_n + W_{n+1}, \quad n \in N.$$

The sequence $S = \{S_n : n \in N\}$ is called a renewal process if $W_1, W_2, ...$ are independent and identically distributed non-negative random variables.

- If $S = \{S_n : n \in N\}$ is a renewal process and F is the distribution function of the inter-renewal times, then

$$P\{S_{n+m} - S_n \leq t | S_0, S_1, ..S_n\} = F^m(t), \mathrm{t} \geq 0$$

  where F^m is the m-fold convolution of F with itself.

- The number of renewals N_t in the interval $[0, t]$ is given by

$$N_t(\omega) = \sum_{n=0}^{\infty} I_{[0,t]}(S_n(\omega)),$$

  where $I_A(x) = 1$ or 0 according as $x \in A$ or not.

- If $R(t)$ denotes the expected number of renewals in $[0, t]$, then

$$R(t) = E[N_t] = \sum_{n=0}^{\infty} E[I_{[0,t]}(S_n)] = \sum_{n=0}^{\infty} P\{S_n \leq t\} = \sum_{n=0}^{\infty} F^n(t).$$

- A renewal process S is said to be *recurrent* if $W_n < \infty$ almost surely for every n. If S is recurrent, then

$$\lim_{t\to\infty} \frac{R(t)}{t} = \frac{1}{m},$$

  where m is the expected value of the times between successive renewals.

1.2.3 Markov renewal process

Suppose for each $n \in N$, a random variable X_n taking values in a countable set E and a random variable T_n taking values in $R_+ = [0, \infty)$ s.t $0 = T_0 \leq T_1 \leq T_2, ...$

Definition 1.2.4. *The stochastic process* $(X,T) = \{(X_n, T_n) : n \in N\}$ *is said to be a Markov renewal process if*

$$P\{X_{n+1} = j, T_{n+1} - T_n \leq t | X_0, ..., X_n; T_0, ..., T_n\} = P\{X_{n+1} = j, T_{n+1} - T_n \leq t | X_n\}$$

for all $n \in N, j \in E$ *and* $t \in R_+$.

- If the sojourn times at states are all equal to 1, it becomes a Markov chain.
- If the distribution of the sojourn times are all exponential and independent of the state next to be visited, it becomes a Markov process.
- If there is only one state, it becomes a renewal process.

A detailed discussion on Markov renewal theory is given in Cinlar [23].

1.3 Modeling tools

Here, we describe some tools that we have used in analyzing the models discussed in this thesis.

1.3.1 Exponential distribution

A random variable X is said to be exponentially distributed with parameter $\lambda > 0$ if it possesses the density function

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x},\ 0 \leq x < \infty \\ &= 0,\ x < 0. \end{aligned}$$

- The distribution function is given by $F(x) = 1 - e^{-\lambda x},\ x \geq 0$.
- The Laplace Transform function is $\frac{\lambda}{s+\lambda}$.
- The mean of this distribution is $\frac{1}{\lambda}$ and variance is $\frac{1}{\lambda^2}$.
- The moment generating function is $M_x(t) = (1 - \frac{t}{\lambda})^{-1}$.

Two important properties that make exponential distribution much useful in modelling queuing systems are the following:

- **Memoryless property:** Exponential distribution is the only distribution of the continuous type having this property. This property implies that if X denotes the duration of some activity, which is still going on, then the distribution of the remaining part of the activity is that of X, no matter when the activity has begun. That is,

$$P(X \geq x + y | X \geq x) = P(X \geq y).$$

- **Minimum of two exponential variates is exponential:** Let X_1 and X_2 be two exponential random variables with parameter λ_1 and λ_2 respectively, then $min(X_1, X_2)$ is exponential with parameter $\lambda_1 + \lambda_2$. Also $P(X_i < X_j) = \frac{\lambda_i}{\lambda_i + \lambda_j}$, $i, j = 1, 2$ and $i \neq j$.

1.3.2 Poisson process

A counting process $N = \{N_t : t \geq 0\}$ is called a Poisson process provided that the following axioms hold:
(a) for almost all ω, each jump of $t \to N_t(\omega)$ is of unit magnitude;
(b) for any $t, s \geq 0, N_{t+s} - N_t$ is independent of N_u; $u \leq t$;
(c) for any $t, s \geq 0$, the distribution of $N_{t+s} - N_t$ is independent of t.
Note that if the time between successive occurrence in a renewal process $S = \{S_n : n \in N\}$ are exponential, then S becomes the sequence of successive occurrences in a Poisson process.

- If N_t is a Poisson Process, we can prove that $P\{N_{t+s} - N_t = k\} = \frac{e^{-\lambda s}(\lambda s)^k}{k!}$ for some constant $\lambda \geq 0$.

- if $T_1, T_2, ..$ are the successive instants of occurrences, then for any $n \geq 0$,

$$P\{T_{n+1} - T_n \leq t \,|T_0, ..., T_n\} = 1 - e^{-\lambda t}.$$

- Let $L = \{L_t : t \geq 0\}$ and $M = \{M_t : t \geq 0\}$ be two Poisson processes independent of each other with rates λ and μ respectively. Now, define $N = \{N_t : t \geq 0\}$ as

$$N_t(\omega) = L_t(\omega) + M_t(\omega), \text{ for each } \omega \in \Omega.$$

Then N is a Poisson process with rate $\nu = \lambda + \mu$.

- Decomposition of a Poisson process results in Poisson processes which are independent of each other.

1.3.3 Phase type distribution

A Phase type (PH) distribution is defined as the distribution of time until absorption in a finite state continuous time Markov process, which has a finite number say, m of transient states and one absorbing state, to which absorption is certain.
Let the generator matrix of the Markov process be written in the form $\begin{bmatrix} T & T^0 \\ \mathbf{0} & 0 \end{bmatrix}$ where T is a $m \times m$ matrix corresponding to the transient states and T^0 is a column vector of order m. Then $T\mathbf{e} + T^0 = \mathbf{0}$. Let the initial distribution with which the process starts in states $j,\ j = 1, 2, ..., m+1$, where the first m states are transient and $m+1$ is absorbing, be $(\alpha_1, \alpha_2, ..., \alpha_{m+1})$. Then $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)$ and T define the distribution of the corresponding phase type variate and we use (α, T) as its representation.

- If F is the distribution function of a phase type variate having representation (α, T),
$$F(x) = 1 - \alpha e^{Tx}\mathbf{e},\ x \geq 0.$$
- The corresponding probability density function $f(.)$ is given by
$$f(x) = \alpha e^{Tx}T^0,\ x \geq 0.$$
- The Laplace- Stieltjes transform $f(s)$ of $F(.)$ is given by
$$f(s) = \alpha_{m+1} + \alpha(sI - T)^{-1}T^0, Re(s) \geq 0.$$
- The ith moment about zero, $\mu_i = i!\alpha(-T)^i\mathbf{e},\ \mathrm{i} = 1, 2, 3, ...$
- Phase type distributions form a dense family of distributions on $[0, \infty)$.

- For $X \sim PH(\alpha, T)$ and $Y \sim PH(\beta, S)$, $Z = X + Y$ is $PH(\gamma, L)$ where $\gamma = (\alpha, \alpha_{m+1}\beta)$ and $\begin{bmatrix} L & L^0 \\ \mathbf{0} & 0 \end{bmatrix} = \begin{bmatrix} T & T^0\beta & \beta_{k+1}T^0 \\ 0 & S & S^0 \\ \mathbf{0} & \mathbf{0} & 0 \end{bmatrix}$.

- Any finite convex mixture of phase type variates is also a phase type variate. If $X_i \sim PH(\alpha_i, T_i)$, $i = 1, .., k$ such that $Z = X_i$ with probability p_i, then $Z \sim PH(\gamma_{mix}, L_{mix})$, where $\gamma_{mix} = (p_1\alpha_1, p_2\alpha_2, ..., p_k\alpha_k)$ and $L_{mix} = \begin{bmatrix} T_1 & 0 & \cdots & \mathbf{0} \\ \mathbf{0} & T_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & T_k \end{bmatrix}$.

- For $X \sim PH(\alpha_x, T_x)$ of order k and $Y \sim PH(\alpha_y, T_y)$ of order m, $min(X, Y)$ is phase-type distributed with representation (γ_{min}, L_{min}) with $L_{min} = T_x \otimes I_y + I_x \otimes T_y$, where $\gamma_{min} = \alpha_x \otimes \alpha_y$ and $\otimes$ represents Kronecker product. Also, $max(X, Y)$ is phase-type distributed with representation (γ_{max}, L_{max}), where

$$L_{max} = \begin{bmatrix} T_x \otimes I_y + I_x \otimes T_y & I_x \otimes T_y^0 & T_x^0 \otimes I_y \\ \mathbf{0} & T_x & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & T_y \end{bmatrix}$$

and $\gamma_{max} = (\alpha_x \otimes \alpha_y, \alpha_x\alpha_{y,m+1}, \alpha_{x,k+1}\alpha_y)$.

Example 1.3.1. ***Exponential distribution :*** *The exponential variate with parameter λ is phase type with representation (α, T), where $\alpha = 1$ and $T = -\lambda$. (of course, the same exponential variate has other PH representations, since in general PH representation is not unique.)*

Example 1.3.2. ***Erlang distribution:*** *A random variable X is said to follow an Erlang-k distribution , k=1,2,... if it has the probability density function of the form*

$$f(x) = \frac{(\mu x)^{k-1}}{(k-1)!}\mu e^{-\mu x},\ x \geq 0.$$

One phase type representation of Erlang-k is (α, T), where

$$\alpha = (1, 0, 0, ..., 0)_{1\times k}$$

and

$$T = \begin{bmatrix} -\mu & \mu & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & -\mu & \mu & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & -\mu & \mu & 0 & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdot & \cdot & 0 & \cdot & -\mu & \mu \\ 0 & 0 & \cdot & \cdot & 0 & \cdot & 0 & -\mu \end{bmatrix}_{k \times k}$$

1.3.4 Fitting phase type distribution

Consider the problem of estimating the parameters of the PH distribution $PH(\beta, S)$ of order m from a random sample y_i , 1⩽ i⩽ n. There is a trajectory of the underlying Markov chain for each y_i, 1⩽ i⩽ n. If we can observe the entire trajectory, then the sample will be called *complete sample*, otherwise we call the sample as *incomplete sample*.

Complete sample: In connection with the above considered n trajectories, define the following statistics:

- B_i = number of trajectories that start in phase i, for i=1,2,..,m.
- N_i = number of trajectories for which absorption occurs from phase i, for i=1,2,..,m.
- N_{ij} = number of jumps from phase i to j, 1≤ i, j≤ m, i≠ j.
- Z_i = total sojourn time in phase i, for i=1,2,..,m.
  Then, the joint likelihood function of the sample is given by

$$L = \prod_{k=1}^{m} \beta_k^{B_k} \prod_{i=1}^{m} \prod_{j \neq i}^{m} S_{ij}^{N_{ij}} e^{-S_{ij} Z_i} \prod_{l=1}^{m} (s_l^0)^{N_l} e^{-s_l^0 Z_l}.$$

  The maximum likelihood estimators of the parameters β and S can be computed by using the above mentioned sufficient statistics, as shown below:

$$\hat{\beta}_i = \frac{B_i}{n}, \hat{S}_{ij} = \frac{N_{ij}}{Z_i}, \hat{s}_i^0 = \frac{N_i}{Z_i}, 1 \leqslant i, j \leqslant m, i \neq j, \tag{1.3.1}$$

  and we can set $\hat{S}_{ii} = -\sum_{j \neq i} \hat{S}_{ij} - \hat{s}_i^0$.

Incomplete sample: In practical cases, we will not be having the complete sample from the PH distribution. Instead, we will be dealing with a set of samples from the PH distribution, which gives us the total lifetime of the trajectories until absorption. In this case, we can consider the set of sufficient statistics as missing and can estimate them, based on whatever is observed, using the expectation-maximisation (EM) algorithm. For more details on this, refer Assmussen et al. [10]. A brief summary of this procedure is as follows:

Let

$$M(y,\beta,S) = \int_0^y e^{S(y-u)} s^0 \beta e^{Su} du, \quad \text{where } s^0 = -S\mathbf{e}.$$

Given a sample value y from $PH(\beta, S)$, conditional expectations of the sufficient statistics $\hat{B}_i, \hat{Z}_i, \hat{N}_i$ and $\hat{N}_{ij}$ can be computed as

$$\hat{B}_i(y,\beta,S) = \beta_i e_i' e^{Sy} s^0 / \beta e^{Sy} s^0, \tag{1.3.2}$$

$$\hat{Z}_i(y,\beta,S) = M_{ii}(y,\beta,S) / \beta e^{Sy} s^0, \tag{1.3.3}$$

$$\hat{N}_i(y,\beta,S) = s_i^0 \beta e^{Sy} e_i / \beta e^{Sy} s^0, \tag{1.3.4}$$

$$\text{and} \quad \hat{N_{ij}}(y,\beta,S) = S_{ij} M_{ji}(y,\beta,S) / \beta e^{Sy} s^0, i \neq j \text{ respectively}, \tag{1.3.5}$$

where e_i is a column vector of appropriate dimension with 1 in the ith place and 0 elsewhere.

Now, let $y_1, y_2, \ldots, y_n$ be a sample from the $PH(\beta, S)$ distribution and suppose that $B_i^{[k]}, N_i^{[k]}, N_{ij}^{[k]}$ and $Z_i^{[k]}$ are the statistics analogous to B_i, N_i, N_{ij} and Z_i defined above, related to the k^{th} trajectory. Then, we can write

$$B_i = \sum_k B_i^{[k]}, \; N_i = \sum_k N_i^{[k]}, \; N_{ij} = \sum_k N_{ij}^{[k]}, \text{ and } Z_i = \sum_k Z_i^{[k]}. \tag{1.3.6}$$

Here $B_i^{[k]}$ assumes the value 1 or 0 if the k^{th} process starts in phase i or not, and similarly $N_i^{[k]}$ takes the value 1 or 0 according as the k^{th} process gets absorbed from phase i or not.

The E-Step and the M-Step

To compute the ML estimates of the parameter $\theta = (\beta, S)$, we use the EM algorithm. The first step of each iteration of the EM algorithm, the E-step, consists of calculating the conditional expectation of the sufficient statistics , given the observed sample y and the current estimate of θ say, $\theta^{(k)}$, by using (1.3.2) – (1.3.5). Then in the M-step, these estimates of the sufficient statistics are used in (1.3.1) to get the new estimate of the parameter θ. That is, we get the new estimate of θ simply by replacing the statistics in (1.3.1) with their conditional expectations, given by (1.3.2) – (1.3.5), evaluated in the E-step. Then, these improved estimates of θ are used in E-step to get the new estimates of the sufficient statistics and so on. This process is repeated till we get reasonably good approximates for the parameter $\theta = (\beta, S)$.

1.3.5 Uniformization technique

Consider a finite space Markov process $\{X(t) : t \geq 0\}$ which has matrix Q as its generator. As the process has only finitely many states, there exists a constant c such that $|q_{ii}| \leq c < \infty \ \forall i \in E$. It is well known that $|q_{ii}|$ is the parameter of the exponentially distributed sojourn time in state i. The Matrix $K = \frac{1}{c}Q + I$ is stochastic.

Now, we take a Poisson process with rate λ and denote $t_0, t_1, ...$ as the renewal epochs of that process. Let us define a discrete time Markov Chain $\{Z_n : n \in N\}$ that has K as its transition matrix, and is independent of the Poisson process. Next, we define a process $\{Y(t) : t \geq 0\}$ such that

$$Y(t) = Z_n \quad \text{for} \quad t_n \leq t < t_{n+1} \quad \text{for} \quad n \geq 0.$$

Then $Y(t)$ is a Markov process having Q as its generator since the transition function

$$\begin{aligned} P_{ij}(t) &= P\{Y(t) = j | Y(0) = i\} \\ &= \sum_{n=0}^{\infty} P\{Y(t) = j | Y(0) = i, N(t) = n\} P\{N(t) = n\} \\ &= \sum_{n=0}^{\infty} P\{Z_k = j | Z_0 = i, N(t) = n\} \frac{e^{-\lambda t}(\lambda t)^n}{n!} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=0}^{\infty} K_{ij}^n \frac{e^{-\lambda t}(\lambda t)^n}{n!} \\
&= \left(\sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} K^n \right)_{ij} \\
&= \left(\sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} \sum_{0 \le k \le n} \binom{n}{k} (\frac{1}{\lambda} Q)^k \right)_{ij} \\
&= (\exp(Qt))_{ij}.
\end{aligned}
$$

So, this technique allows us to interpret a CTMC as a DTMC in which the constant times between any two transitions are replaced by independent exponential variables with the same parameter. Also, to evaluate the transition probabilities of $\{X(t)\}$, we don't need to use differential equations and Matrix exponential. For example, uniformization approach helps us to evaluate the matrix exponential $M(y, \beta, S)$, considered in the previous section by using the expression

$$
M(y, \beta, S) = \sum_{r=0}^{\infty} e^{-cy} \frac{(cy)^{r+1}}{(r+1)!} \sum_{m=0}^{r} P^m p \beta P^{r-m}
$$

where $c = \max(-S_{ii}),\ P = \frac{1}{c} S + I,\ p = \frac{1}{c} s^0$.

For more details on uniformization, see Latouche and Ramaswami [50].

1.3.6 Some processes helpful in modelling queueing systems

PH renewal processes:

- Inter-renewal time follows PH distributions.
- Associate a Markov process on $1, 2, .., n, n+1$ with initial distribution $(\alpha, 0)$ and infinitesimal generator

$$
Q = \begin{bmatrix} T & T^0 \\ \mathbf{0} & 0 \end{bmatrix}.
$$

- Start at time 0; evolve to absorption; instantaneously restart with a new state.
- Reinitialized time points form a PH renewal process.

The infinitesimal generator matrix of the PH renewal process is given by

$$D = T + T^0\alpha.$$

See Neuts [63] for more details on PH renewal process.

Markovian point processes:

Consider a PH renewal process with representation (α, T). Define N(t) as the number of renewals in (0,t], J(t) as the phase of the process at t. Then the process $\{(N(t), J(t)) : t > 0\}$ is called a Markovian Point Process.

Its infinitesimal generator is given by

$$Q = \begin{bmatrix} T & T^0\alpha & \mathbf{0} & \cdots \\ \mathbf{0} & T & T^0\alpha & \cdots \\ \mathbf{0} & \mathbf{0} & T & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

Markovian arrival processes (MAP):

Consider a directing process which is having D as its generator matrix .

- We can split D as $D = D_0 + D_1$, where D_0 is associated with phase transitions without arrivals and D_1 corresponds to the phase transitions with arrivals .
- Define $N(t)$ as the number of arrivals in (0,t] and $J(t)$ as the phase of the directing process at t. Then the process $\{(N(t), J(t)) : t > 0\}$ is called a Markovian arrival process.
- Its infinitesimal generator is given by

$$Q = \begin{bmatrix} D_0 & D_1 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & D_0 & D_1 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & D_0 & D_1 & \cdots \\ \cdot & \cdot & \cdot & D_0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

The process spends an exponential amount of time in state i with rate λ_i and moves to state j, accompanied by an arrival or not with probabilities $P_{ij}(1)$ or

$P_{ij}(0)$, respectively . Therefore, we can write

$$D_0 = \begin{bmatrix} -\lambda_1 & \lambda_1 P_{12}(0) & \lambda_1 P_{13}(0) & \cdots \\ \lambda_2 P_{21}(0) & -\lambda_2 & \lambda_2 P_{23}(0) & \cdots \\ \cdot & \cdot & -\lambda_3 & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}$$

and

$$D_1 = \begin{bmatrix} \lambda_1 P_{11}(1) & \lambda_1 P_{12}(1) & \lambda_1 P_{13}(1) & \cdots \\ \lambda_2 P_{21}(1) & \lambda_2 P_{22}(1) & \lambda_2 P_{23}(1) & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

- Fundamental arrival rate $= \pi D_0 \mathbf{e}$, where $\pi D = 0, \pi \mathbf{e} = 1$.

Batch Markovian arrival processes (BMAP):
A more general variant of MAP is BMAP, where a state transition can be accompanied even by the arrival of a batch of customers rather than a single one as in MAP. The generator matrix is then given by

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & D_4 & \cdots \\ \mathbf{0} & D_0 & D_1 & D_2 & D_3 & \cdots \\ \mathbf{0} & \mathbf{0} & D_0 & D_1 & D_2 & \cdots \\ \cdot & \cdot & \cdot & D_0 & D_1 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix}$$

where D_k corresponds to the phase transitions, accompanied by a batch of k customers. A detailed discussion on MAP and BMAP is given in Lucantoni [54].

1.4 Matrix analytic methods

Though queuing theory has lot of applications in areas like mobile phone communications, banks, computer networking etc., the usual existing methods like methods of generating functions, methods using transforms etc. fail to provide much tractability in the analysis of many queuing models especially when the distribution of inter-arrival time or service time is not exponential. Neuts [63] introduced and developed matrix analytic methods which gave us the ability to

analyse much complicated stochastic models in an algorithmic way and to numerically explore the problems more deeply. We use the matrix analytic methods to analyse quasi-birth-death (QBD) processes arising in the models in this thesis.

1.4.1 Level independent quasi-birth-and-death processes

Consider a Markov process $\{X(t) : t \in R_+\}$ with state space $\{(i,j) : i \geq 0\ , 1 \leq j \leq m\}$ which we again partition as $\cup_{n\geq 0} l(n)$ where $l(n) = \{(n,i) : 1 \leq i \leq m\}$. n is called the *level* and j is called the phase. Such Markov process is called a quasi-birth-and-death process (QBD) if one-step transitions from a level is restricted to states in the same level or in the two adjacent levels. If the transition probabilities are assumed as level independent, i.e $P\{X_1 = (n',j)|X_0 = (n,i)\}$ depends on $i, j, n' - n$, but does not depend on the specific values of n and n', then that process is called level independent quasi-birth-and-death process (LIQBD). The infinitesimal generator is the irreducible tridiagonal matrix Q, given by

$$Q = \begin{bmatrix} B_0 & A_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ B_1 & A_1 & A_0 & \mathbf{0} & \mathbf{0} & \cdot & \mathbf{0} \\ \mathbf{0} & A_2 & A_1 & A_0 & \mathbf{0} & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_2 & A_1 & A_0 & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

Then, we have the following theorem (Neuts [63]).

Theorem 1.4.1. *The QBD is positive recurrent iff the minimal non-negative solution R of the equation*

$$R^2 A_2 + R A_1 + A_0 = \boldsymbol{0} \tag{1.4.1}$$

has spectral radius less than one and there exists a positive vector x_0 that satisfies the following finite system of equations:

$$\begin{aligned} x_0(B_0 + RB_1) &= \boldsymbol{0} \\ x_0(I-R)^{-1}\boldsymbol{e} &= 1. \end{aligned} \tag{1.4.2}$$

The stationary probability vector $x = (x_0, x_1, ..)$ of the QBD is given by

$$x_i = x_0 R^i, \quad i \geq 1. \tag{1.4.3}$$

If the matrix $A = A_0 + A_1 + A_2$ is irreducible, then $sp(R) < 1$ if and only if $\pi A_0 \mathbf{e} < \pi A_2 \mathbf{e}$, where π is the stationary probability vector of matrix A.

Matrix R records the rate of sojourn in the states of $l(n+1)$ per unit local time of $l(n)$.

We define two matrices G and U as follows:

$G_{ij} = P\{\tau < \infty \,\&\, X(\tau) = (n-1, j) | X(0) = (n, i)\}$, where τ is the first passage time to a level

and

$$U = A_1 + A_0 G.$$

That is, U is the generator of the Markov process in the local time $l(n)$ before first visit to $l(n-1)$. We use Logarithmic Reduction Algorithm (See Latouche and Ramaswami [50]) to find the matrix G, which in turn can be used to find U and hence R, since $R = -A_0(A_1 + A_0 G)^{-1}$.

Special Structure: If the QBD is recurrent and that $A_2 = \mathbf{c.r}$, where $\mathbf{c}$ is a column vector and $\mathbf{r}$ is a row vector normalized by $\mathbf{r.e}$=1, then

$$G = \mathbf{e.r}$$

For details, see Latouche and Ramaswami [50].

1.4.2 Level dependent quasi-birth-and-death processes

A level dependent quasi-birth-and-death processes (LDQBD) is a Markov process with state space $\cup_{n\geq 0} l(n)$, where $l(n) = \{(n, i) : 1 \leq i \leq m\}$. The infinitesimal generator is given by

$$Q = \begin{bmatrix} A_{1,0} & A_{0,0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ A_{2,1} & A_{1,1} & A_{0,1} & \mathbf{0} & \mathbf{0} & \cdot & \mathbf{0} \\ \mathbf{0} & A_{2,2} & A_{1,2} & A_{0,2} & \mathbf{0} & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_{2,3} & A_{1,3} & A_{0,3} & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

The transitions are to the adjacent levels alone, but the transition rate will depend on the level at which the process is then. Assuming that the process is irreducible,

we have the following theorems (Latouche and Ramaswami [50]).

Theorem 1.4.2. *If a LDQBD is positive recurrent, its limiting probability vector* $\pi = (\pi_0, \pi_1, \pi_2, ...)$ *satisfies the relation*

$$\pi_n = \pi_{n-1} R_n, \quad n \geq 1$$

where the matrices R_n *are the minimal non-negative solutions of the system of equations*

$$R_n R_{n+1} A_{2,n+1} + R_n A_{1,n} + A_{0,n} = \boldsymbol{0}.$$

Theorem 1.4.3. *The LDQBD is positive recurrent if and only if there exists a strictly positive solution of the system*

$$\pi_0 = \pi_0 (A_{1,0} + R_1 A_{2,1}),$$

normalized by

$$\pi_0 \sum_{n \geq 0} \Big(\prod_{1 \leq k \leq n} R_k \Big) \boldsymbol{e} = 1.$$

1.5 Regenerative approach

The process $\{X(t) : t \in T\}$ is called regenerative if there exists a random epoch t_1 such that
(i) $\{X(t + t_1) : t \in T\}$ is independent of $\{X(t) : 0 \leq t < t_1\}$ and
(ii) $\{X(t + t_1) : t \in T\}$ has the same distribution as $\{X(t) : t \in T\}$.

Now, let us consider a class of regenerative queueing systems with renewal instants $\{t_n\}$ (for example arrival instants) with i.i.d inter-renewal times $\tau_n = t_{n+1} - t_n, n \geq 0$ independent of arrival instants t_0 of initial customer. The input process is called *zero-delayed* if $t_0 = 0$. If $t_0 > 0$, we call it *delayed* process. Let us take the input rate $\lambda = 1/E[\tau] \in (0, \infty)$ and $\{S_n, n \geq 0\}$ as service time process. Let us introduce the right-continuous queue size process $\nu = \nu(t)$, where at instant t, $\nu(t)$ denotes the number of customers in the system and let $\nu_n = \nu(t_n^-)$. Also, consider the right-continuous workload process $W = \{W(t)\}$, where $W(t)$ is the amount of time needed to complete service of all customers presented at t and define $W_n = W(t_n^-)$.

From *Lindley's recursion,* we know that

$$W_{n+1} = (W_n + S_n - \tau_n)^+,\ n \geq 0.$$

As it is obvious that $\{W_k = 0\} = \{\nu_k = 0\}$, the embedded sequences $\{\tau_n\}$ and $\{W_n\}$ regenerate simultaneously at the instants $\beta_n, n \geq 0$, which are defined by

$$\beta_{n+1} = min\{k > \beta_n \,|\nu_k = 0\},\quad n \geq 0,\ \text{with } \beta_0 = 0.$$

Then $\{T_n\}$, defined by $T_{n+1} = min\{t_k > T_n|\nu_k = 0\}$, are the regeneration instants of the processes τ and W, where $T_0 = t_0$. Clearly,

$$T_1 = \tau_0 + ... + \tau_{\beta_1 - 1}. \tag{1.5.1}$$

We call the corresponding queueing process zero-delayed if $\nu_0 = 0$. We use E_0 for expectation, $T_1 = T, \beta_1 = \beta$.
The renewal process $\{\beta_n\}$ is called positive-recurrent if

$$\beta_1 < \infty \ \text{ with probability 1 } \text{ and } \ E_0[\beta] < \infty.$$

Similarly, the renewal process $\{T_n\}$ is called positive-recurrent if

$$T_1 < \infty \ \text{ with probability 1 } \text{ and } \ E_0[T] < \infty.$$

For discrete-time-processes, we define the forward renewal time $\beta(n) = \min\{\beta_k - n|\beta_k - n > 0\}$ at instant n. For continuous-time-processes, we define the forward renewal time $T(t) = \min\{T_k - t|T_k - t > 0\}$ at instant t. In the zero-delayed case, $\beta(0) = \beta$, $T(0) = T$. Now, because of the asymptotic behaviour of $\beta(n)$ and $T(t)$, whatever be the initial values of $\beta(0)$ and $T(0)$, $\beta(n)$ and $T(t) \Rightarrow \infty$ if and only if $E[\beta] = \infty$ and $E[T] = \infty$ respectively, where $\rightarrow$ stands for convergence in probability.

Using Wald's identity, from equation (1.5.1), we get $E_0[T] = E[\tau]E_0[\beta]$. As the main condition to establish stability of a regenerative process is the finiteness of the mean regeneration period, it is sufficient to prove either $E_0[\beta] < \infty$ or $E_0[T] < \infty$ in order to show that the regenerative process is stable. The last step is to establish that regeneration period is aperiodic, which we can show by the convergence to the

limit distribution

$$\lim_{n\to\infty} P(X_n \in \cdot) = \frac{E_0[\sum_{k=0}^{\beta-1} \mathbf{1}_{(X_k \in \cdot)}]}{E_0[\beta]} = \pi(\cdot)$$

or

$$\lim_{t\to\infty} P(X(t) \in \cdot) = \frac{E_0[\int_0^T \mathbf{1}_{(X_k \in \cdot)}]}{E_0[T]} = \pi(\cdot)$$

as the case may be. Now to establish the finiteness of the mean regeneration period, it is sufficient to find constants $L < \infty, \epsilon > 0$ and a non-random sequence n_i, which will tend to ∞, such that $\inf_i P(\beta(n_i) \leq L) \geq \epsilon$ or $\inf_i P(T(n_i) \leq L) \geq \epsilon$. A detailed discussion on regenerative theory is given in Morozov [56].

1.6 Thesis outline and contributions overview

The models that we consider here are mainly intended for studying dynamics and characteristics of some queueing systems that are related to communication. We hope that their proper modelling and extensive analysis carried out here will help in the actual design of such real life systems.

In chapter 2, we propose a model to study the queueing characteristics of nodes in a wireless network in which the channel access is governed by the well known binary exponential back off (BEB) rule. By offering the general phase type (PH) distributional assumptions to channel idle and busy periods and assuming Poisson packet arrival processes at nodes, we represent the model as a quasi birth death process (QBD) and analyse it by using matrix analytic methods. Stability of the system is examined. Several important queueing characteristics that help in efficient design of such systems are derived. Extensive simulation analysis is performed to establish the validity of our theoretical results. It is shown that both the simulated and theoretical results agree on some important performance measures. Some real life data has been used to get approximate PH representations for channel idle and busy period variates, which in turn are used for numerical illustrations. Also, we use these results to compute the joint system size distribution of the network under some assumptions on routing and handling of packets.

In chapter 3, we consider a wireless sensor network model that handles emergency packets. If the packets cannot be transmitted within an amount of time after being generated, their relevance will be lost. Here also, we use the standard BEB scheme for

collision avoidance and take exponential distribution assumptions for channel busy and idle periods. Then, we derive distribution of time since a packet is ready for transmission till it successfully transmitted/ timed out and probability mass function of the number of collisions experienced by a packet. Some numerical illustrations are provided.

A single server retrial system with several classes of external customers (incoming calls) is considered in chapter 4. Arrival of these calls follow Poisson rule. Time between successive retrials, under constant-rate discipline, are assumed as exponential with class dependent parameter. In addition to this, there are different classes of outgoing calls (internal customers) which may occupy server for a random amount of time when it is idle. We consider two variations of the basic model. In the first model, if the server is busy, an arriving class-k customer joins the kth orbit and each orbit behaves like a separated queue. In the second model, all blocked customer joins a common orbit. Service times of each class are assumed to be iid. Regenerative analysis admits general service time distribution and by applying this approach, we derive some explicit expressions for the steady-state probabilities. All distributions are assumed to be exponential while applying matrix analytical approach. With the combination of these two methods, we derive the steady state distribution and some performance measures explicitly for Markovian model.

In chapter 5, a single-server retrial model with multiple classes of customers(calls) is considered. Arrival of customers follow independent Poisson rule. A new customer, facing a busy server upon its arrival, may join the corresponding (class-dependent) orbit queue with a class-dependent probability, or leaves the system forever (balks). The orbit queues follow constant retrial rate discipline, that is, only one (oldest) orbital customer of each orbit queue makes attempts to occupy the server, in a gap of class-dependent exponential times. Within each class, service times are assumed to be independent and identically distributed (iid). We show that this setting generalizes the two-way communication systems discussed in chapter 4. This multi-class system with general service time distributions is analyzed using regenerative approach. Necessary and sufficient stability conditions, as well as some explicit expressions for the basic steady-state probabilities, are obtained. A restricted, two-way communication model with exponential service time distributions, is analyzed by matrix-analytic method. Moreover, we combine both methods to efficiently derive explicit solutions for the restricted model. An extensive simulation analysis is performed to gain deep insight into the model stability and performance. It

is shown that both the simulated and exact results agree on some important measures for which analytical expressions are available, and hence establish the validity of our theoretical treatment. We numerically study the sophisticated dependence structure of the model to uncover the orbits interaction. We give further details and intuitive explanations for the system performance which complement the derived explicit expressions.

In chapter 6, we consider a single server retrial model with two streams of calls namely, incoming and outgoing. Each stream consists of multiple classes of calls. As part of the internal work load, presence of outgoing calls are always assumed in the system. Arrival of incoming calls obeys the Poisson law. Upon seeing a busy server at its arrival epoch, an incoming call will be directed to an orbit according to the class it belongs to and tries to get an idle sever in a gap of exponential amount of time, having class dependent mean. Similar kind of attempt is also being made by the outgoing calls to reach an idle server. Once the server becomes idle, if neither an incoming nor an outgoing call is being turned up for an exponential amount of time, the server goes for vacation and the vacation time is assumed to be exponential. Within each stream, service times of multiple classes of calls are assumed to be independent exponential with class dependent means. Matrix analytic method and regenerative approach are used to derive the explicit form of the steady state probabilities. Many performance measures are computed to analyse the system performance.

Finally, some conclusions have been drawn based on our contributions, and the thesis concludes with plan for future work.

Chapter 2

A Model to Study Some Queuing Characteristics of Nodes in a Wireless Network

2.1 Introduction

In recent times, much interest has been shown in setting up of wireless networks for local area communication due to the increase in infrastructure cost involved in establishing wired networks. Though wired network is usually faster than wireless network, the latter offers many privileges to its customers like accessing the network from anywhere in the range and sharing files and resources with other devices that are connected to the network, even without having a port.

In order to illustrate the dynamics and behaviour of nodes in a wireless network, we consider a simple network, having 4 nodes with gateway GW, as shown in figure 1. Among the 4 nodes, assume that nodes 1 and 3 are source nodes. That is, external arrivals can be generated only at these nodes. The entire route of the packets generated at each of the source nodes is also shown in figure 2.1.

A circle centred at a node defines the transmission range of that node. All nodes that are lying inside the transmission range of a particular node are called *one hop neighbours* of that node. All other nodes that are lying inside the circles centred at all one-hop neighbours of a node are called its *two-hop neighbours*. In figure 2.1, node 1 has only one one-hop neighbour which is node 2 but it has two two-hop neighbours namely, node 3 and 4. Node 2 has 3 one-hop neighbors 1,3 and 4, but it has no two-hop neighbour. Similarly for node 3, one-hop neighbour is node 2 and two hop

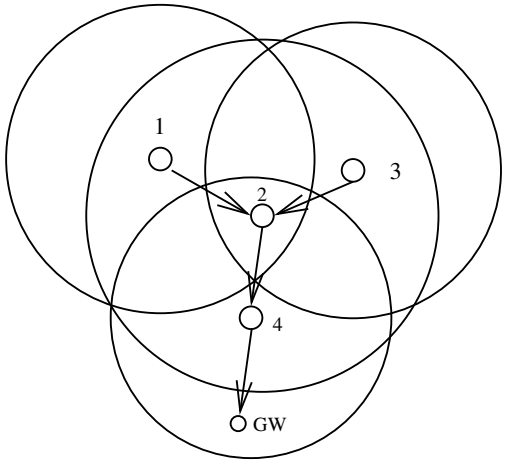

Figure 2.1: A general network

neighbours are 1 and 4. If a transmission is being taken place between two nodes, all one-hop neighbours of those two nodes will sense the channel as busy, but all two-hop nodes, being not in the transmission range of source node, will not be able to sense the channel and hence there can be a chance of collision due to the possible simultaneous transmission by these nodes.

Wireless medium is shared and scarce. Hence multiple nodes may transmit data packets at the same time over the same channel. So, multiple access protocols are needed to coordinate transmissions to avoid collisions. IEEE 802.11 protocol has been set up for fixing international standards for Wireless Local Area Networks (WLAN's). In the 802.11 protocol (see [74]), the fundamental mechanism to access the medium is called as the *distributed co-ordination function* (DCF). DCF is a random access scheme based on the carrier sense multiple access with collision avoidance (CSMA/CA) protocol.

The DCF mechanism, employed for channel access, is described as follows: consider a node in the network having a packet for transmission. First, the node senses the channel and if the channel is found idle for a pre-determined period, known as distributed inter-frame space (DIFS), it transmits the packet. If the channel is found busy at the instant at which it is monitored, the node has to undergo a back off period consisting of a random number of time slots, called *back off counters*, to minimize the collisions caused by packet transmissions from the other nodes . So, it initializes a back off timer and at each time instant at which the channel is monitored, the back off counter is decremented if the medium is found idle, and

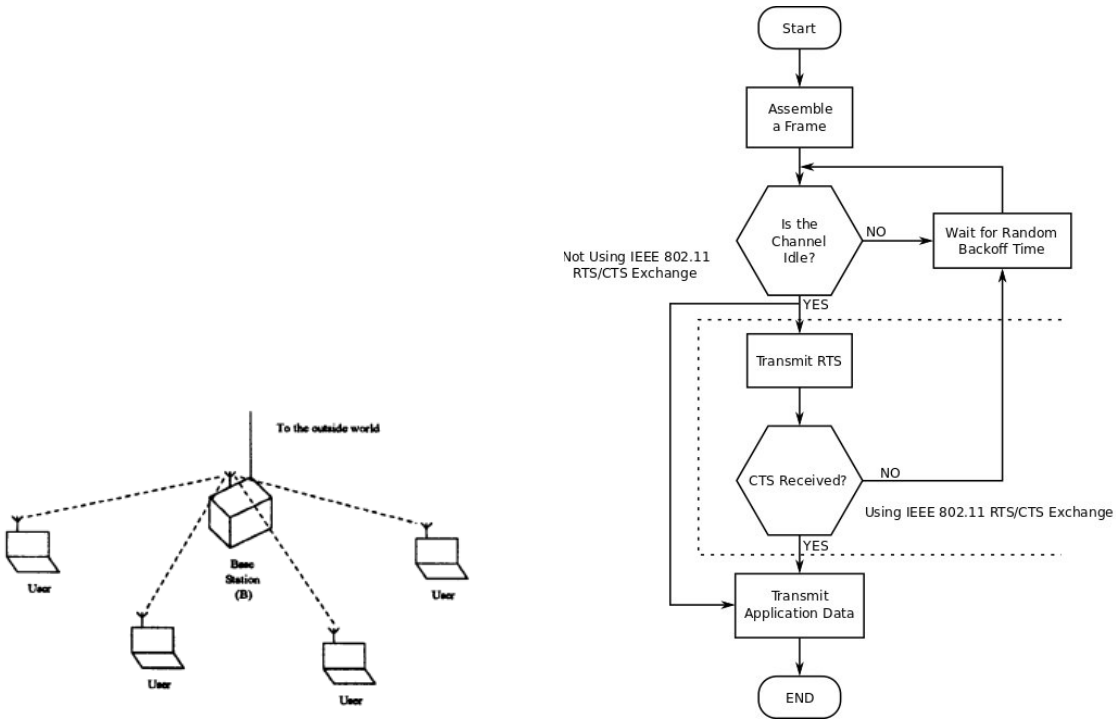

Figure 2.2: Wireless network

is frozen if it is sensed busy. In the latter case, the timer resumes only after the channel has been found idle for a period longer than DIFS. Among all nodes, the one for which the back off timer expires first begins transmission whereas the other nodes freeze their timers by sensing the channel busy at their respective channel monitoring instants. Once the current transmission gets completed for a node, the back off process, as explained above, is repeated. Upon the successful reception of a packet, the destination sends back an immediate acknowledgement (ACK) after a time interval equal to short inter-frame space (SIFS). In addition to the basic access mechanism, another optional method, called ready-to-send/clear-to-send (RTS/CTS) mechanism, is also adopted under DCF. Under this, a node, having a packet for transmission, reserves the channel by sending a special ready-to-send short frame and the destination node acknowledges the receipt of the same by sending back a clear-to send frame. After this, the normal packet transmission and ACK response occur. Since collision may occur only on the RTS frame, and if so the same is detected by the lack of CTS response, the RTS/CTS mechanism helps us to increase the system performance especially when lengthy messages are transmitted. In spite of which kind of collision avoidance scheme is being used for channel access, packet collision chances can not be ruled out completely. This is because, for several nodes, back off

periods may expire simultaneously (of course with small probability) and hence there are chances of simultaneous multiple transmissions. Also, the phenomena like *hidden node problem*, which was first mentioned by Tobagi and Kleinrock [70], may cause packet collisions. For more details on hidden node problem, refer Boroumand et al [21] also.

DCF employs a contention resolution scheme namely, binary exponential back-off (BEB) scheme, to reduce the probability of collisions that may occur due to multiple simultaneous transmissions. Under this rule, if a node is ready to transmit a packet, it fixes the contention window size as W and a random value for the number of back off counters is selected uniformly from 0, 1, 2,... $W-1$. After expiring the back off time (as per the procedure explained earlier), the node transmits the packet and if the packet meets with a collision in that attempt, the contention window size will be set as $W_1 = 2W$ and a value for number of back off counters to be undergone is selected uniformly from 0, 1, 2,... $W_1 - 1$. If the packet is again included in a collision on its next attempt, the contention window size will be doubled again and this will go on till the node experiences a maximum of m unsuccessful attempts. Even after this, if the node again fails to transmit the packet in its future attempts, after each trial, the contention window size will be fixed as $W_m = 2^m W$. In between, if an attempt results in successful transmission, the contention window size for the node will be reset as W. Hence, the minimum contention window size, $CW_{min} = W$ and the maximum contention window size, $CW_{max} = 2^m W$.

Phase type (PH) distributions introduced by Neuts [[62], [63]] form a dense subset (in the metric of weak convergence of distributions) of the family of all distributions defined on $[0, \infty)$ and hence they can be used as approximates to probability distributions of general non-negative random variates. A PH variate can be regarded as the time until absorption in a finite state Markov chain with one absorbing state into which absorption is certain. It finds applications in many areas where stochastic models can be effectively used for system analysis. Apart from their denseness property that makes them versatile as models, there are many other motivations for using PH distributions in statistical models, important among them are their connection with Markov chain and theory of matrices. For more details on PH distributions and their characteristics, see Latouche and Ramaswami [50] also.

Inspired by the standard BEB rule used for channel access in wireless network, here we propose a queueing model to study some major characteristics of packets waiting at an arbitrary node. The objective of the study is not to analyse the

characteristics of total delay experienced by a packet in the whole network as done by several researchers, rather we are more interested in finding the important statistical measures such as probability distributions of system size, waiting time of packets, number of collisions experienced by a packet at a node, and computation of their moments in a rigorous manner. Apart from these, we also make an attempt to derive approximate joint system size distributions at nodes in some kind of multi hop wireless networks. As mentioned earlier, several researchers have tried to analyse the throughput and packet delay occurring in wireless communication networks. Bisnik and Abouzeid [20] has made an attempt to compute the average end-to-end delay and maximum achievable per-node throughput in random access multi-hop wireless ad hoc network with stationary nodes. They modelled random access multi-hop wireless networks as open $G/G/1$ queueing networks and used the diffusion approximation (see Kobayashi[48]]) to derive closed form expressions for the average end-to-end delay. In Deepak [27], delay analysis in a multi-hop wire less network has been presented and probability distribution of the time spent by a packet at an arbitrary node from the epoch at which it is ready for transmission till it is successfully transmitted has been derived as a discrete PH distribution. In these articles, approximate forms of the pmf and average of the queue size at nodes have been computed by using diffusion approximation. But, here, we derive the probability distribution of the number of packets at nodes and its average, rigorously by using matrix analytic methods. Apart from these, we find out the probability distributions of the number of collisions experienced by a packet at a node, total waiting time at the node, and their moments by using the theory of continuous time phase type variates. Then, we use these results as well as the theoretical approach developed by Kelly [42] to compute the joint distribution of system size at all nodes in some multi hop networks, where processing of packets are governed by some specific queue disciplines.

Here, we use PH distributional assumptions for channel idle times and busy times (as sensed by the node under consideration) due to denseness property of the PH class. Later, we find out appropriate representation for these PH variates by collecting real data from a wireless network and computing the maximum likelihood estimates (MLE) of the concerned PH parameters by using the Expectation-Maximization (EM) approach. For more details on fitting PH distributions, see Asmussen et al. [10].

In this chapter, section (2.2) provides the mathematical formulation and analysis of the model in connection with the traffic flow at a node. Many important system performance measures that are useful in designing such systems are also derived in this

section. Section (2.3) deals with the joint system size distribution of a network under some assumptions on routing and handling of packets. In section (2.4), extensive simulation analysis is performed to establish the validity of our theoretical results, and some real life data has been used to get approximate PH representations for channel idle and busy period variates, which in turn are used for numerical illustrations.

2.2 The mathematical model

Based on the sequence of events that happen in connection with the traffic flow at a node and the rule governed by the standard BEB scheme, the following assumptions are made:

(i) Data packets arrive at the node and waiting in the queue till they are being considered for transmission. Let the arrival process be Poisson of rate λ.

(ii) At an instant at which a packet is considered for transmission, the back-off period for the node starts if the channel is sensed as idle, and a value for back-off counter is uniformly selected from 0, 1, 2,..., $W-1$. If the packet has already experienced j collisions, then the back off counter will be from 0, 1, 2,..., $W_j - 1$. Also, time spent on each of the back off counters are assumed to be independent and identically distributed exponential variates having mean $1/\mu$.

(iii) If the channel is found busy after expiring a back off counter time, the back off timer will be frozen and resume only after the channel is sensed as idle. The channel idle periods and busy periods are assumed to be independent PH variates having representations (α_1, T_1) and (α_2, T_2) of order n_1 and n_2 respectively.

(iv) When the back-off counter at a particular back-off stage becomes zero, the node starts transmission irrespective of the fact that if the channel is busy or idle. Packet transmission times are assumed to be independent and identical exponential variates having mean $1/\gamma$.

(v) A transmission results in collision with probability p and is successful with probability $1-p$.

For nodes in the network that relay messages, the arrivals are of two types- those primarily generated at the concerned nodes and the ones transmitted from the other neighbouring nodes- so that in strict sense the aggregate arrival stream may not be Poisson due to the correlation among the arrivals. However, Zhou and Mitchell [76] and Wang et al [73] have justified the Poisson arrival assumption through some experimental measurements. Similarly, a justification to the exponential assumption

for back off periods and transmission times are also given in Zhou and Mitchell [76] . However, our back off period distributional assumption is more general since it consists of a random number of exponential back off slots and hence the entire back off period acts almost like a PH variate.

Throughout this thesis it is assumed that **e** stands for a column vector of 1's and I represents unit matrix, of appropriate dimensions.

In relation to a particular node, let the state variables be defined as follows:

- $N(t)$, the system size at time t.
- $J(t)$, 0 or 1 according as the channel is idle or busy, as sensed by the node, at t.
- $S(t)$, the back-off stage at t.
- $B(t)$, the back off counter at t. Note that when $S(t) = i$, $B(t)$ assumes one of the values uniformly from $0, 1, 2, \ldots, W_i - 1$, where $W_i = 2^i W$.
- $P(t)$, the phase of the idle period or busy period at t (depending on if the channel is idle or busy at t).

Then the process $\{X(t) : t \geq 0\} = \{(N(t), J(t), S(t), B(t), P(t)) : t \geq 0\}$ is a continuous time Markov chain with state space $E = \cup_i \cup_j E_{ij}$, where $E_{ij} = \{0, 1, 2, 3, ...\} \times \{j\} \times \{i\} \times \{0, 1, 2, ..., W_i - 1\} \times \{1, 2, ..., n_{j+1}\}$ for $i = 0, 1, 2, \ldots, m$ and $j = 0, 1$. Assume that the states are arranged lexicographically. If we term the system size as level, then the one step transitions of the above process from the states in a level are restricted either to the states in the same level or to the states in the adjacent levels.

Transitions among states in the same level, say, level n, $n > 0$ can be considered as four cases as given below:

If $\overline{(n, i)}, i = 0, 1$ represents the state vector corresponding to the system size n, and the channel is idle or busy respectively, then transitions from level n to itself consists of those from $\overline{(n, 0)}$ to $\overline{(n, 0)}$, $\overline{(n, 0)}$ to $\overline{(n, 1)}$, $\overline{(n, 1)}$ to $\overline{(n, 0)}$, and $\overline{(n, 1)}$ to $\overline{(n, 1)}$.

Transitions rates from $\overline{(n, 0)}$ to $\overline{(n, 0)}$ are given by

$$\begin{bmatrix} D_0 & B_1 & 0 & 0 & \ldots 0 & 0 \\ 0 & D_1 & B_2 & 0 & \ldots 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & \ldots D_{m-1} & B_m \\ 0 & 0 & 0 & 0 & \ldots 0 & D_m + B_m' \end{bmatrix}$$

where

$$D_i = \left[\begin{array}{c|ccccc} & (n,*,i,0) & \overline{(n,0,i,1)} & \overline{(n,0,i,2)} & \cdots & \overline{(n,0,i,W_i-1)} \\ \hline (n,*,i,0) & -\gamma & 0 & 0 & \cdots & 0 \\ \overline{(n,0,i,1)} & \mu e & T_1-\mu I & 0 & \cdots & 0 \\ \overline{(n,0,i,2)} & 0 & \mu I & T_1-\mu I & \cdots & 0 \\ . & . & . & . & \cdots & . \\ \overline{(n,0,i,W_i-1)} & 0 & 0 & 0 & \cdots & T_1-\mu I \end{array}\right] \tag{2.2.1}$$

for $i = 0, 1, 2, \ldots m$ and

$$B_i = \left[\begin{array}{c|ccccc} & (n,*,i,0) & \overline{(n,0,i,1)} & \overline{(n,0,i,2)} & \cdots & \overline{(n,0,i,W_i-1)} \\ \hline (n,*,i-1,0) & p\gamma/W_i & p\gamma\alpha_1/W_i & p\gamma\alpha_1/W_i & \cdots & p\gamma\alpha_1/W_i \\ \overline{(n,0,i-1,1)} & 0 & 0 & 0 & \cdots & 0 \\ \overline{(n,0,i-1,2)} & 0 & 0 & 0 & \cdots & 0 \\ . & . & 0. & . & \cdots & . \\ \overline{(n,0,i-1,W_{i-1}-1)} & 0 & 0 & 0 & \cdots & 0 \end{array}\right] \tag{2.2.2}$$

for $i = 1, 2, \ldots m$.

$B_m^{'}$ has the same structure as B_m with the only exception that it reflects the transitions from back off stage m to itself.

Also note that, as per the back off rule, when the back off counter reaches zero, the node transmits the packet. So, in this case it is not relevant to mention the idle or busy status of the channel, due to which the state at this instant is represented by $(n, *, i, 0)$ corresponding to system size n and back off stage i.

In D_i, the transitions among the states in the state vector $\overline{(n,0,i,k)}$ are just transitions among intermediate states of the channel idle phase process without decrementing the back off counter and the transitions from the states $\overline{(n,0,i,k)}$ to $\overline{(n,0,i,k-1)}$ correspond to decrement in the back off counter by 1 with rate μ.

B_i defines the transitions when a packet, already experienced $i-1$ collisions, is again involved in another collision (with probability p). In this case, the back off stage of the node is changed to i and back off counter can be chosen as one of the values uniformly from $0, 1, 2, ..., W_i - 1$.

Transition rates corresponding to $\overline{(n,0)}$ to $\overline{(n,1)}$ are given by

$$\begin{bmatrix} E_0 & 0 & 0 & \cdots & 0 \\ 0 & E_1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & E_m \end{bmatrix}$$

where

$$E_i = \left[\begin{array}{c|cccc} & \overline{(n,1,i,1)} & \overline{(n,1,i,2)} & \cdots & \overline{(n,1,i,W_i-1)} \\ \hline (n,*,i,0) & 0 & 0 & \cdots & 0 \\ \overline{(n,0,i,1)} & T_1^0\alpha_2 & 0 & \cdots & 0 \\ \overline{(n,0,i,2)} & 0 & T_1^0\alpha_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots \quad \cdot \\ \overline{(n,0,i,W_i-1)} & 0 & 0 & \cdots & T_1^0\alpha_2 \end{array}\right] \tag{2.2.3}$$

for $i = 0, 1, \ldots, m$. Here, the entries of E_i signify the rates in connection with the absorption of the idle time and consequent beginning of the busy period.

In a similar manner, rates of transitions among the states in $\overline{(n,1)}$ to those in $\overline{(n,0)}$ are given by

$$\begin{bmatrix} F_0 & 0 & 0 & \cdots & 0 \\ 0 & F_1 & & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & F_m \end{bmatrix}$$

where

$$F_i = \left[\begin{array}{c|ccccc} & (n,*,i,0) & \overline{(n,0,i,1)} & \overline{(n,0,i,2)} & \cdots & \overline{(n,0,i,W_i-1)} \\ \hline \overline{(n,1,i,1)} & 0 & T_2^0\alpha_1 & 0 & \cdots & 0 \\ \overline{(n,1,i,2)} & 0 & 0 & T_2^0\alpha_1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \overline{(n,1,i,W_i-1)} & 0 & 0 & 0 & \cdots & T_2^0\alpha_1 \end{array}\right] \tag{2.2.4}$$

for $i = 0, 1, \ldots, m$.

Also, transitions among the states $\overline{(n,1)}$ to themselves yield

$$\begin{bmatrix} G_0 & 0 & 0 & \cdots & 0 \\ 0 & G_1 & & \cdots & 0 \\ . & . & . & \cdots & . \\ 0 & 0 & 0 & \cdots & G_m \end{bmatrix}$$

where

$$G_i = \left[\begin{array}{c|ccccc} & \overline{(n,1,i,1)} & \overline{(n,1,i,2)} & \overline{(n,1,i,3)} & \cdots & \overline{(n,1,i,W_i-1)} \\ \hline \overline{(n,1,i,1)} & T_2 & 0 & 0 & \cdots & 0 \\ \overline{(n,1,i,2)} & 0 & T_2 & 0 & \cdots & 0 \\ \overline{(n,1,i,3)} & 0 & 0 & T_2 & \cdots & 0 \\ . & . & . & . & \cdots & . \\ \overline{(n,1,i,W_i-1)} & 0 & 0 & 0 & \cdots & T_2 \end{array}\right] \tag{2.2.5}$$

for $i = 0, 1, \ldots, m$.

Transition from level n to level $n-1$ can happen only with the successful transmission of a packet. In order to avoid channel capture, the node has to wait a back off time after a transmission even though the channel is idle immediately after a successful transmission. Hence, the said transitions are confined only from the states in $\overline{(n,0)}$ to those in $\overline{(n-1,0)}$. After a successful transmission - which happens with rate $(1-p)\gamma$ - the node will set 0 as its back off stage and choose back-off counter as one of the values uniformly from $0, 1, 2, W-1$. Thus the transition rates from the states $\overline{(n,0)}$ to $\overline{(n-1,0)}$ are given by

$$Z_0 = \begin{bmatrix} C_0 & 0 & \cdots & 0 \\ C_1 & 0 & \cdots & 0 \\ . & . & . & \cdots & . \\ C_m & 0 & \cdots & 0 \end{bmatrix} \tag{2.2.6}$$

where

$$C_i = \left[\begin{array}{c|ccccc} & (n-1,*,0,0) & \overline{(n-1,0,0,1)} & \overline{(n-1,0,0,2)} & \cdots & \overline{(n-1,0,0,W-1)} \\ \hline (n,*,i,0) & (1-p)\gamma/W & (1-p)\gamma\alpha_1/W & (1-p)\gamma\alpha_1/W & \cdots & (1-p)\gamma\alpha_1/W \\ \overline{(n,0,i,1)} & 0 & 0 & 0 & \cdots & 0 \\ \overline{(n,0,i,2)} & 0 & 0 & 0 & \cdots & 0 \\ . & . & . & . & \cdots & . \\ \overline{(n,0,i,W_i-1)} & 0 & 0 & 0 & \cdots & 0 \end{array}\right]$$

for $i = 0, 1, \ldots, m$.

Clearly, the transitions from level n to level $n+1$ can occur only by an arrival of a packet. In this case, transitions can happen from the states in $\overline{(n,0)}$ to those in $\overline{(n+1,0)}$, and from those in $\overline{(n,1)}$ to $\overline{(n+1,1)}$, with the rate given by the matrix λI in each case.

Now, we consider transitions among the boundary states. Transition rates from the states $\overline{(1,0)}$ to $\overline{(0,0)}$ are given by

$$Z_1 = \left[\ U_0, U_1, \ldots U_m\ \right]^T, \tag{2.2.7}$$

where

$$U_i = \left[\ (1-p)\gamma\alpha_1, 0, \ldots 0\ \right]^T,$$

which is having $W_i - 1$ blocks, and 0 is a zero matrix of order n_1.

Rate at which transitions occurring from the states $\overline{(0,0)}$ to $\overline{(1,0)}$ are given by

$$Z_2 = \left[\ H_0 \quad 0 \quad 0 \quad \cdots \quad 0\ \right], \tag{2.2.8}$$

where

$$H_0 = \left[\ \lambda e/W \quad \lambda I/W \quad \ldots \quad \lambda I/W\ \right].$$

Similarly, rates of transition from $\overline{(0,1)}$ to $\overline{(1,1)}$ are given by

$$Z_3 = \left[\ H_1 \quad 0 \quad 0 \quad \cdots \quad 0\ \right], \tag{2.2.9}$$

where

$$H_1 = \left[\ \tfrac{\lambda I}{W-1} \quad \tfrac{\lambda I}{W-1} \quad \cdots \quad \tfrac{\lambda I}{W-1}\ \right].$$

Note that, since the states $\overline{(0,0)}$ and $\overline{(0,1)}$ are those corresponding to an empty node (the one having no packet for transmission) and no back off stage is being set for such nodes, the states representing channel idle and busy periods are only figured under these tags.

Thus, the infinitesimal generator of the continuous time Markov process $\{X(t) : t \geq 0\}$ is

$$Q = \begin{bmatrix} L & M & 0 & 0 & 0 & \cdots & 0 \\ K & A_1 & A_0 & 0 & 0 & \cdot & 0 \\ 0 & A_2 & A_1 & A_0 & 0 & \cdot & 0 \\ 0 & 0 & A_2 & A_1 & A_0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \tag{2.2.10}$$

where

$$L = \begin{bmatrix} T_1 - \lambda I & T_1^0 \alpha_2 \\ T_2^0 \alpha_1 & T_2 - \lambda I \end{bmatrix}$$

$$K = \begin{bmatrix} Z_1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$M = \begin{bmatrix} Z_2 & 0 \\ 0 & Z_3 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} D_0 - \lambda I & B_1 & 0 & \cdots & 0 & E_0 & 0 & 0 & \cdots & 0 \\ 0 & D_1 - \lambda I & B_2 & \cdots & 0 & 0 & E_1 & 0 & \cdots & 0 \\ 0 & 0 & D_2 - \lambda I & \cdots & 0 & 0 & 0 & E_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & D_m + B_m^{'} - \lambda I & 0 & 0 & 0 & \cdots & E_m \\ F_0 & 0 & 0 & \cdots & 0 & G_0 - \lambda I & 0 & 0 & \cdots & 0 \\ 0 & F_1 & 0 & \cdots & 0 & 0 & G_1 - \lambda I & 0 & \cdots & 0 \\ \cdot & \cdot & F_2 & \cdots & 0 & 0 & 0 & G_2 - \lambda I & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & F_m & 0 & 0 & 0 & \cdots & G_m - \lambda I \end{bmatrix} \tag{2.2.11}$$

$$A_2 = \begin{bmatrix} Z_0 & 0 \\ 0 & 0 \end{bmatrix} \tag{2.2.12}$$

and

$$A_0 = \lambda I. \tag{2.2.13}$$

Note that the matrices D_i, B_i, E_i, F_i, G_i, Z_0, Z_1, Z_2, and Z_3 are given by eqns. (2.2.1) - (2.2.9) respectively.
Next, we explore the condition under which the steady state distribution exists for the irreducible Markov process $\{X(t) : t \geq 0\}$.

2.2.1 Stability condition

By Neuts [63], $\{X(t) : t \geq 0\}$ is positive recurrent (and hence steady state distribution exists) if and only if

$$\pi A_0 e < \pi A_2 e \tag{2.2.14}$$

where π is the stationary distribution associated with the generator $A = A_0 + A_1 + A_2$. Let π be partitioned as $\pi = [\pi_{00}, \pi_{01}, ..., \pi_{0m}, \pi_{10}, \pi_{11}...\pi_{1m}]$ based on if the channel is idle or busy.
Now

$$A = \begin{bmatrix} D_0 + C_0 & B_1 & 0 & \cdots & 0 & E_0 & 0 & 0 & \cdots & 0 \\ C_1 & D_1 & B_2 & \cdots & 0 & 0 & E_1 & 0 & \cdots & 0 \\ C_2 & 0 & D_2 & \cdots & 0 & 0 & 0 & E_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ C_m & \cdot & \cdot & \cdots & D_m + B_m' & 0 & 0 & 0 & \cdots & E_m \\ F_0 & 0 & 0 & \cdots & 0 & G_0 & 0 & 0 & \cdots & 0 \\ 0 & F_1 & 0 & \cdots & 0 & 0 & G_1 & 0 & \cdots & 0 \\ \cdot & \cdot & F_2 & \cdots & 0 & 0 & 0 & G_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & F_m & 0 & 0 & 0 & \cdots & G_m \end{bmatrix}$$

Then $\pi A = 0$ and $\pi e = 1$ yield:

$$\pi_{*i0} = \pi_{*00} p^i \text{ for } 1 \leq i \leq m-1 \tag{2.2.15}$$

$$\pi_{*m0} = \pi_{*00}(p^m / 1 - p) \tag{2.2.16}$$

$$\pi_{0ik} = \pi_{*00}\frac{\gamma p^i}{W_i}\alpha_1 \sum_{j=1}^{W_i-k}(-1)^j\mu^{j-1}(H)^{-j} \text{ for } 0 \le i \le m, 1 \le k \le W_i - 1 \tag{2.2.17}$$

$$\pi_{1ik} = \pi_{0ik}T_1^0\alpha_2(-T_2)^{-1} \text{ for } 0 \le i \le m, 1 \le k \le W_i - 1 \tag{2.2.18}$$

$$\pi_{*00} = \Big[\frac{1}{1-p}+\sum_{i=0}^{m}\sum_{k=1}^{W_i-1}\frac{p^i\gamma}{\mu}\frac{(W_i-k)}{W_i}+\sum_{i=0}^{m}\sum_{k=1}^{W_i-1}\sum_{j=1}^{W_i-k}\frac{\gamma p^i}{W_i}\alpha_1(-1)^j\mu^{j-1}(H)^{-j}T_1^0\alpha_2(-T_2)^{-1}e\Big]^{-1} \tag{2.2.19}$$

where

$$H = T_1 - \mu I + T_1^0\alpha_2(-T_2)^{-1}T_2^0\alpha_1;$$

$$\pi_{0i} = [\pi_{*i0}, \pi_{0i1}, ..., \pi_{0iW_i-1}],$$

and

$$\pi_{1i} = [\pi_{1i1}, \pi_{1i2}...., \pi_{1iW_i-1}]$$

for $i = 0, 1, \ldots, m$. Then, eqn.(2.2.14) implies

$$\lambda < \gamma\pi_{*00} \tag{2.2.20}$$

where π_{*00} is given by eqn.(2.2.19).

2.2.2 Steady state distribution of $\{X(t) : t \ge 0\}$

Under the assumption that eqn.(2.2.20) holds good, here we compute the steady state distribution of the number of packets at the node by using matrix analytic methods. Since $\{X(t) : t \ge 0\}$ is a level independent quasi birth death process (LIQBD), its steady state distribution (if exists) has a matrix geometric form. Please refer Neuts [63] and Latouche and Ramaswami [50] for details on the matrix geometric solution of QBD processes.

If q_i denotes the stationary probability vector of the above LIQBD process corresponding to the level (number of packets) i, we have

$$q_i = q_1R^{i-1} \text{ for } i \ge 2, \tag{2.2.21}$$

where R is the minimal non negative (matrix) solution to the equation

$$R^2 A_2 + RA_1 + A_0 = 0$$

and the vectors q_0 and q_1 are obtained by solving the system

$$q_0 L + q_1 K = 0 \tag{2.2.22}$$

$$q_0 M + q_1 (A_1 + RA_2) = 0 \tag{2.2.23}$$

subject to the normalizing condition

$$(q_0 + q_1 (I - R)^{-1}) e = 1. \tag{2.2.24}$$

For the computation of the matrix R, we use the logarithmic reduction algorithm proposed by Latouche and Ramaswami [50]. Once $q = [q_0, q_1, \ldots, q_n, \ldots]$ is available, $[q_0 e, q_1 e, \ldots, q_n e, \ldots]$ define the pmf of the queue size of the packets waiting at the node.
So, the mean queue size at a node

$$E[N] = \sum_{n=1}^{\infty} n q_n e = q_1 (I - R)^{-2} e \tag{2.2.25}$$

and its variance

$$Var[N] = 2q_1 R(I - R)^{-3} e + E[N](1 - E[N]).$$

In the sections to follow, we derive probability distributions of some important variates and other measures which may help us in understanding the dynamics and behaviour of the system more rigorously.

2.2.3 Distribution of the time since a packet is picked for transmission till it is successfully transmitted

As we discussed, even though a packet at a node is considered for transmission, it can not be transmitted immediately but it has to undergo a random back off time before it is getting transmitted. The transmission may result in a collision and in

that case it has to undergo another random back off time and so on. So, naturally we are interested in knowing the probability distribution of the time a packet spends at a node after it is picked for transmission. Let U be the time a packet spends at a node from the instant at which it is picked for transmission till it is successfully transmitted. Then U is the time till absorption in a finite state Markov chain, with absorption corresponds to the successful transmission, starting from any of the states corresponding to the case where the channel is idle and the back off stage is 0. Hence U is a PH variate with representation (β, S) where

$$\beta = \begin{bmatrix} \frac{1}{W} & \frac{\alpha_1}{W} & \cdots & \frac{\alpha_1}{W} & 0 & 0 & \dots & 0 \end{bmatrix} \tag{2.2.26}$$

and

$$S = \begin{bmatrix} D_0 & B_1 & 0 & \cdots & 0 & E_0 & 0 & 0 & \cdots & 0 \\ 0 & D_1 & B_2 & \cdots & 0 & 0 & E_1 & 0 & \cdots & 0 \\ 0 & 0 & D_2 & \cdots & 0 & 0 & 0 & E_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \cdot & \cdots & D_m + B_m' & 0 & 0 & 0 & \cdots & E_m \\ F_0 & 0 & 0 & \cdots & 0 & G_0 & 0 & 0 & \cdots & 0 \\ 0 & F_1 & 0 & \cdots & 0 & 0 & G_1 & 0 & \cdots & 0 \\ \cdot & \cdot & F_2 & \cdots & 0 & 0 & 0 & G_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & F_m & 0 & 0 & 0 & \cdots & G_m \end{bmatrix}$$

Hence the density of U is

$$f(u) = \beta e^{Su}(-S)e, \qquad 0 < u < \infty \tag{2.2.27}$$

So, the average time a packet spends at a node after it is selected for transmission,

$$E[U] = \beta(-S)^{-1}e \tag{2.2.28}$$

and its variance

$$Var[U] = 2\beta(S)^{-2}e - (\beta(-S)^{-1}e)^2.$$

2.2.4 Probability mass function of the number of collisions experienced by a packet

In order to find out the pmf of the number of collisions experienced by a packet, we consider the Markov process $\{(J(t), B(t), S(t), P(t)) : t \geq 0\}$ with $\triangle$ as the absorbing state, which corresponds to the successful transmission of the packet. It is to be noted that for a node, back off stage k means the packet that is being considered for transmission has already experienced k collisions. By arranging all the transient states lexicographically and listing the absorbing state $\triangle$ as the last one, we get the generator

$$\hat{Q} = \begin{bmatrix} K_0 & J_1 & 0 & 0 & \cdots & 0 & L_0 \\ 0 & K_1 & J_2 & 0 & \cdots & 0 & L_1 \\ 0 & 0 & K_2 & J_3 & \cdots & 0 & L_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & K_m + J_m^{'} & L_m \end{bmatrix}$$

where

$$K_i = \begin{bmatrix} D_i & E_i \\ F_i & G_i \end{bmatrix}, L_i = \begin{bmatrix} C_i \\ 0 \end{bmatrix} \text{ for } i = 0, 1, \ldots m;$$

$$J_i = \begin{bmatrix} B_i & 0 \\ 0 & 0 \end{bmatrix} \text{ for } i = 1, \ldots m,$$

$$\text{and } J_m^{'} = \begin{bmatrix} B_m^{'} & 0 \\ 0 & 0 \end{bmatrix}.$$

Let y_k be the probability that the packet experiences exactly k collisions before its successful transmission. Then

$$y_0 = \beta(-K_0)^{-1}L_0$$

$$y_k = \beta \prod_{j=1}^{k} (-K_{j-1})^{-1} J_j (-K_k)^{-1} L_k \text{ for } k = 1, 2 \ldots m-1$$

and

$$y_{k+m} = \beta \prod_{j=1}^{m} (-K_{j-1})^{-1} J_j (-(K_m + J_m^{'}))^{-(k+1)} L_m \text{ for } k \geq 0,$$

where β is given by eqn.(2.2.26). Therefore, expected number of collisions experienced

by the packet,

$$E[C] = \sum_{k=1}^{\infty} k y_k.$$

and its variance,

$$Var[C] = \sum_{k=1}^{\infty} k^2 y_k - E[C]^2.$$

2.2.5 Waiting time distribution of a tagged packet

In this section, we attempt to derive the distribution of the time since a packet joins the queue at a node till it is successfully transmitted. Assume that the packet under consideration joins the queue say, as the rth unit, $r > 0$. Now, we consider the Markov process $\{Y(t) = (R(t), J(t), S(t), B(t), P(t) : t \geq 0\}$, where $R(t)$ is the rank of the said packet at time t and $J(t), S(t), B(t)$, and $P(t)$ are same as those defined earlier. The rank R(t) of the packet is assumed to be i if it is the ith unit in the queue at time t. it's rank will be decremented by 1 after each successful transmission from that node. Since the packets that arrive after the tagged packet can not affect its rank, level changing transitions in $W(t)$ can take place in only one side of the diagonal of the generator. Clearly, $R(t)$ can assume one of the values $r, r-1, ..., 1$. Let $\overline{0}$ be the absorbing state denoting the successful transmission of the tagged packet.

The infinitesimal generator $\bar{Q}$ of $\{Y(t) : t \geq 0\}$ assumes the form

$$\bar{Q} = \left[\begin{array}{c|cccccc} & \overline{r} & \overline{r-1} & \overline{r-2} & \cdots & \overline{1} & \overline{0} \\ \hline \overline{r} & \overline{A_1} & A_2 & 0 & \cdots & 0 & 0 \\ \overline{r-1} & 0 & \overline{A_1} & A_2 & \cdots & 0 & 0 \\ \overline{r-2} & 0 & 0 & \overline{A_1} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \overline{2} & 0 & 0 & 0 & \cdots & A_2 & 0 \\ \overline{1} & 0 & 0 & 0 & \cdots & \overline{A_1} & K \\ \overline{0} & 0 & 0 & 0 & \cdots & 0 & 0 \end{array}\right]$$

where $\overline{A_1}$ is obtained from A_1 by omitting the λI term ,

$$K = \left[\begin{array}{c} Z_1 \\ 0 \end{array}\right],$$

and Z_1 is given by eqn. (2.2.7).
That is,

$$\bar{Q} = \begin{bmatrix} \overline{T} & \overline{T}^0 \\ 0 & 0 \end{bmatrix},$$

where $\overline{T}$ is the part of the generator corresponding to the transient states $\overline{r}, \overline{r-1}, ..., \overline{1}$.
Hence the waiting time W of a packet that joins a node as the rth unit in the queue is a PH $(\gamma, \overline{T})$ variate with $\gamma = (\overline{q_r}, 0, 0, ..., 0)$, where $\overline{q_r} = q_r/q_r e$ and q_r are given by eqns. (2.2.21)-(2.2.24).
Using the uniformization approach (refer Latouche and Ramaswami [50] for details), the distribution function of the waiting time of such a packet can be computed as

$$W(t) = 1 - \sum_{k=0}^{\infty} e^{-ct} \frac{(ct)^k}{k!} \gamma \bar{P} e,$$

where

$$\bar{P} = \frac{1}{c}\overline{T} + I$$

and c is the maximum of the negative of the diagonal elements of $\overline{T}$.
Also, the average waiting time of such a packet

$$E[W] = \gamma(-\overline{T})^{-1}e = -\overline{q_r}(\overline{A_1})^{-1}[I + \sum_{i=1}^{r-1}(-A_2(\overline{A_1})^{-1})^i].$$

2.2.6 Estimate of the conditional collision probability p

As per our assumption, a node transmits a packet at the epoch at which its back off period expires irrespective of the status of the channel.
Therefore, the steady state probability that a node transmits a packet or transmission probability of an arbitrary node

$$\tau = \sum_{i=0}^{m} \pi_{*i0} = \frac{\pi_{*00}}{1-p} \tag{2.2.29}$$

by using eqns.(2.2.15) and (2.2.16). Note that π_{*00} is given by eqn. (2.2.19). So, if we assume that there are n nodes in the network, including the one being considered, which may possibly transmit simultaneously, then as reasoned in Bianchi [19], we

have

$$p = 1 - (1 - \tau)^{n-1}. \tag{2.2.30}$$

The above class of nodes are generally called *interfering neighbours* of the node under consideration.

Eqns. (2.2.29) and (2.2.30) represent a system of two non-linear equations in two unknowns, which can be solved numerically to get p.

By fixed point iteration, we can suggest a scheme to get p iteratively given by

$$p^{(k+1)} = p^{(k)} + \delta \bar{G}(p^{(k)}), \tag{2.2.31}$$

where $\bar{G}(p) = p + \left(1 - \frac{\pi_{*00}}{1-p}\right)^{n-1} - 1$ and δ is a number such that $\delta \bar{G}'(p) < 0$. Note that $\bar{G}'$ represents the derivative of $\bar{G}$ with respect to p.

2.3 Joint system size distribution

Network queues correspond to systems which consist of many queues with different types of customers moving from one queue to another in their routes. The route of a customer through the queues of the system may be fixed or random. Several researchers produced equilibrium system size distribution in product form for such networks based on the assumption that amounts of service required by a customer at successive queues along its route are independent and exponentially distributed. This assumption forced the said authors to demand that knowledge of the past route of a customer in a queue is of no use in predicting its future route. However, Kelly [42] conjectured that if the queues of the network were of a certain form, then even with the assumptions that the amount of service required by a customer at a queue in its route was almost arbitrarily distributed and depended on its route and the amount of service required by it at other queues along its route, the equilibrium system size distribution could be found in an analytical form. Later Barbour [17] proved this conjecture.

Kelly [42] dealt with an open system and used a customer's type to determine not only its route through the system but also the distribution of the amount of service it required at each queue along that route. The following were the main assumptions made by Kelly [42] and Barbour [17].

- The queueing network consists of J nodes.
- Customers of type i $(i = 1, 2, .., I)$ enter the system in a Poisson stream at rate $\nu(i)$ and pass through the sequence of queues $r(i,1), r(i,2), \ldots, r(i,S(i))$ before leaving the system, where $S(i)$ denotes the number of stages a customer of type i visits along its route.
- A type i customer at its stage s needs a random amount of service Q_{is}.
- Total service effort offered by a single server when there are n_j customers in queue j is $\phi_j(n_j)$.
- A customer in mth position of jth queue will be given a proportion $g_j(m, n_j)$ of this effort, where $1 \leq m \leq n_j$.
- When a customer arrives at queue j, it moves into position m $(1 \leq m \leq n_j + 1)$ with probability $g_j(m, n_j + 1)$.

Then, Kelly [42] conjectured and Barbour [17] later proved that $n(t) \equiv \{n_1(t), n_2(t), ..., n_J(t)\}$ has a limiting distribution $P(n)$ such that

$$P(n) \propto \prod_{j=1}^{J} \frac{a_j^{n_j}}{\prod_{m=1}^{n_j} \phi_j(m)}, \tag{2.3.1}$$

where

$$a_j = \sum_{n=1}^{I} \nu(i) \sum_{s=1}^{S(i)} I_{[r(i,s)=j]} E[Q_{is}], \tag{2.3.2}$$

provided

$$M = \sum_{n} P(n) < \infty.$$

Note that the usage of the same function g in the last two assumptions listed above is very essential, without which the existence of the equilibrium distribution of the joint system size given by eqns (2.3.1) and (2.3.2) will not be valid for network models bearing non-exponential service time distributional assumptions. For a detailed discussion on this, refer Kelly [42] and Barbour [17].

Now we use eqns (2.3.1) and (2.3.2) to determine the joint distribution of the number of packets waiting at nodes in some special type of wireless networks. Let us consider a network with nodes having identical features like the same number of one-hop and two-hop neighbours. Because of this, we can assume that the distribution

of the amount of time the channel is sensed as busy by each of the nodes are identically distributed. In a similar manner, channel idle times sensed by all nodes can also be assumed to be distributed identically. Hence, the distribution of the time from the instant at which a packet is ready to the instant at which it is successfully transmitted from each node are also identically distributed. Its density and mean are defined by eqns (2.2.27) and (2.2.28) respectively. Hence $E[Q_{is}]$ corresponding to our model can assumed to be the same for all i and s, and is given by

$$E[Q_{is}] = \beta(-S)^{-1}e. \tag{2.3.3}$$

Now consider the routing probability matrix as

$$\bar{R} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}.$$

As we assumed earlier, type of a customer will be decided by the route along which it may traverse. Hence, we can have a maximum of $I = n!$ types of customers. Suppose that the total external packet generation to the system obey a Poisson rule of parameter λ and a q_i proportion of these is of type i for $i = 1, 2, \ldots, I$ so that $\sum_{i=1}^{I} q_i = 1$. As we assumed, average time that any type of customer takes at any node in its route is $E[Q_{is}]$, and is given by eqn (2.3.3.)

As per the the two important assumptions made by Kelly [42] and Barbour [17], which are listed as the last two assumptions given above in this section, we should also use the same function g in our model due to the non-exponential variate Q_{is}. Hence, we assume two cases here namely,

case 1

Selection of packets for transmission at nodes is done by LCFS and the new packet always joins at the end of the queue.

Then we have

$$\begin{aligned} g_j(m, n_j) &= 1 \quad if\ m = n_j \\ &= 0 \quad if\ m \neq n_j. \end{aligned}$$

case 2

Selection of packets for transmission is done uniformly from the waiting line and also the customer joins a position randomly (as per uniform law) upon its arrival at a

node along its route. In this case, we have

$$g_j(m, n_j) \quad = \frac{1}{n_j} \quad \text{for} \quad m = 1, 2, 3, ...n_j.$$

In both cases, we have

$$a_j = \sum_{i=1}^{I} q_i \lambda \sum_{s=1}^{S(i)-1} p_{r(i,s),r(i,s+1)=j} \beta(-S)^{-1} e. \tag{2.3.4}$$

Hypothetically, since we have only one server at each node, $\phi_j(m) = 1$ for $m = 1, \cdots, n_j$ and $j = 1, \cdots, J$.
Therefore,

$$P(n) \equiv \prod_{j=1}^{n} \frac{a_j^{n_j}}{\prod_{m=1}^{n_j} \phi_j(m)}$$

$$\equiv \prod_{j=1}^{n} \left(\lambda \beta(-S)^{-1} e \sum_{i=1}^{I} q_i \sum_{s=1}^{S(i)-1} p_{r(i,s),r(i,s+1)=j} \right)^{n_j}. \tag{2.3.5}$$

In the above, $p_{r(i,s),r(i,s+1)=j}$ represents the routing probability of a packet of type i, which is currently at the s th stage of its route, moving to node j at the next stage.

2.4 Numerical illustration

2.4.1 Simulation results

(a) Exponential idle and busy periods (b) Erlang idle and busy periods

Figure 2.3: Difference between theoretical and simulated results

In this section, we present a simulation study of the system under consideration to see whether the theoretical results that we derived via equations (2.2.21) - (2.2.25) match with the simulation results. Single trace of the simulation is obtained for

each value of event corresponding to $N = 1000, 2000, 3000, 4000, 5000, 6000$, where N stands for the total number of packets used in simulation study. From each such trace, estimates of the probabilities of queue being empty (sum of the entries of q_0), queue having one customer(sum of the entries of q_1) and mean queue size (EN) at node are obtained. Absolute difference between these simulated values, which are denoted by q_{0e}, q_{1e}, and EN_e respectively, and the corresponding theoretical values are computed and compared.

The first experiment (see Fig.2.3 (a)) is conducted by assuming exponential channel idle and busy times with rates 1.5 and 1 respectively. Transmission times are exponentially distributed with $\gamma = 0.15$. Time spent on each back off counter is assumed to be exponentially distributed with parameter $\mu = 0.05$. We take minimum contention window size $W = 2$, maximum back off stage $m = 3$, collision probability $p = 0.1$ and arrival rate $\lambda = 0.01$. The parameters are chosen so as to satisfy the system stability condition.

The second experiment (see Fig.2.3 (b)) is carried out by taking Erlang assumptions, which is a particular case of PH distribution, for idle and busy periods of the channel. Idle times and busy periods are assumed to follow Erlang(3,4.5) and Erlang(3,3) respectively of order 3. Also, $\gamma = 0.15$, $\mu = 0.05$, $W = 2$, $m = 2$, $p = 0.1$, and $\lambda = 0.01$.

In both cases, it can be seen that the differences between the theoretical values and simulated values of the above stated measures converge to zero for large values of n, the number of packets used in simulation study. This illustrates the validity of our theoretical results.

2.4.2 Numerical example

In this section, we use our model to study the characteristics of a real time network. We collect around 300000 observations from our institute network (IIST campus network), which is being governed by BEB scheme under 802.11 MAC specification, and use wireshark software to analyse the data. Activities at one of the nodes are monitored and the observations corresponding to the events like arrivals of data packets at the node and amount of channel idle and busy times sensed by that node

are compiled and displayed by using the histograms shown in Fig 2.4, Fig 2.5, and Fig 2.6 respectively.

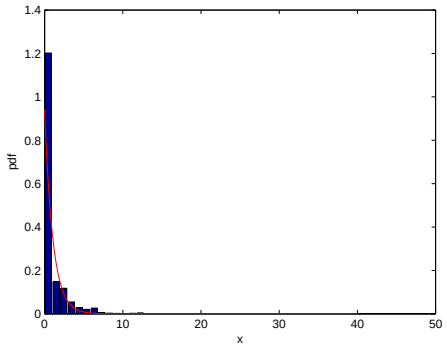

Figure 2.4: Inter-arrival times fitted with $exp(1.0629)$

In Fig 2.4, a histogram with 50 bins is used to represent packet inter arrival times obtained from our real data and surprisingly we can see that the inter arrival time can very well be approximated by an exponential variate of parameter $\lambda = 1.0629/\text{ms}$. The density of the fitted distribution is also shown in Fig 2.4. To some extent, this justifies our Poisson packet arrival process assumption.
In figure 2.5, the channel idle time observations are shown by a histogram having 50 bins. We try to fit this with a PH variate of order 3. By using the EM approach for fitting PH variates, developed by Asmussen et al [10], the approximate ML estimates of the corresponding PH parameters are obtained as

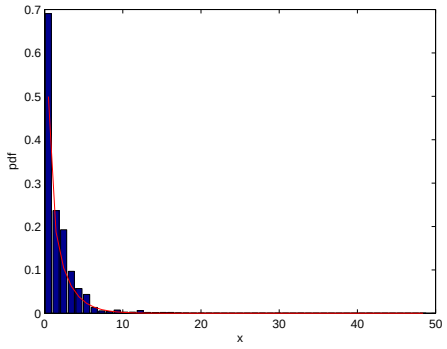

Figure 2.5: Channel idle times fitted with $PH(\alpha_1, T_1)$ of order 3

$$\alpha_1 = \begin{bmatrix} 0.7530 & 0.0766 & 0.1704 \end{bmatrix}, T_1 = \begin{bmatrix} -1.8097 & 0.2397 & 0.6790 \\ 0.1939 & -1.3306 & 0.5483 \\ 0.1847 & 0.5014 & -1.1274 \end{bmatrix}.$$

The density of the fitted PH variate is also exhibited in Fig 2.5.

In Fig 2.6, the histogram used for representing channel busy period observations

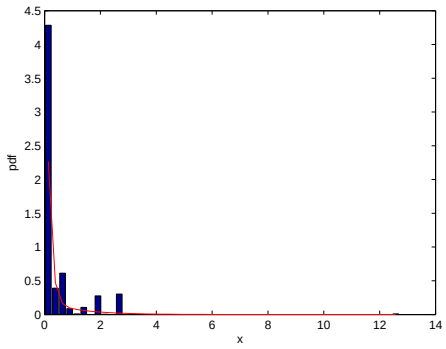

Figure 2.6: Channel busy times fitted with $PH(\alpha_2, T_2)$ of order 3

shows that it is not easy to fit this data with a PH variate of small order. But, going for a representation of higher order will put all numerical computation in chaos. So, for our numerical illustration purpose, we are forced to stick on with a PH approximate of order 3, having parameters as

$$\alpha_2 = \begin{bmatrix} 0.8823 & 0.0307 & 0.0870 \end{bmatrix}, T_2 = \begin{bmatrix} -7.2175 & 0.1960 & 0.5386 \\ 0.2835 & -1.5407 & 0.5142 \\ 0.2729 & 0.5297 & -1.3043 \end{bmatrix}.$$

The density of the PH approximate shown in Fig 2.6 signifies that the fit is far away from a good fit. However, due to the aforementioned reasons, we proceed with this representation for channel busy time.

From the data, we get an estimate of the transmission rate from a node as $\gamma = 2.4856/\text{ms}$. Time spent on each back off counter is assumed to be exponentially distributed with parameter $\mu = 1.5\text{ms}$. We take minimum contention window size $W = 2$ and maximum back off stage $m = 3$. By using the fixed point iteration scheme rendered by eqn. (2.2.31), we get an estimate of the conditional collision probability as $p = .0537$ by taking the number of interfering neighbours as 20. For

the above fixed set of parameters, we compute various measures of effectiveness given by $E[N] = 73.3913, Var[N] = 4700.4, E[U] = 0.6034, Var[U] = .3555, E[C] = 5.0190$ and $Var[C] = 2.8353$.

Now, the probabilities that the waiting times of tagged packets, that join the node at various positions, do not exceed some pre-determined values, are computed and shown in table 2.1. Note that the last row of the table represents average waiting times of such tagged packets.

	r=2	3	4	5	6
t=1	0.2276	0.0806	0.0224	0.0051	0.0009
2	0.4762	0.2896	0.1471	0.0633	0.0203
3	0.6274	0.4795	0.3242	0.1930	0.1014
4	0.7195	0.6140	0.4840	0.3480	0.2269
5	0.7809	0.7058	0.6065	0.4889	0.3662
6	0.8250	0.7705	0.6961	0.6022	0.4936
7	0.8596	0.8181	0.7621	0.6891	0.5998
8	0.8870	0.8546	0.8116	0.7553	0.6840
9	0.9094	0.8836	0.8499	0.8060	0.7497
10	0.9279	0.9069	0.8802	0.8456	0.8011
$E[W]$	3.5025	4.4266	5.3508	6.2750	7.1991

Table 2.1: Waiting time distribution of tagged packets

In order to illustrate the theoretical results, established in section 2.3, numerically, we consider a network model with nodes having equal number of one-hop and two-hop neighbours, as shown in figure 2.7. Here node 1 and 2 are assumed as source nodes and GW is the gateway. The matrices $\bar{R}, \bar{F}$ and $\bar{N}$ exhibit the details of routing of packets, one-hop, and two-hop neighbours of each node respectively. That is, $\bar{F}_{ij}$ =1 if j is a one-hop neighbour of node i. Similarly, $\bar{N}_{ij}$ =1 if j is a two-hop neighbour of i.

$$\bar{R} = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 \\ 4 & 0 & 0 & 0 & 0 & 1 \\ 5 & 0 & 0 & 0 & 0 & 0 \end{array}$$

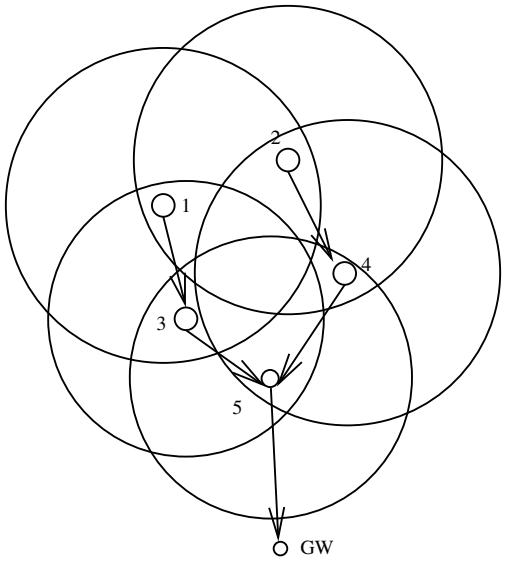

Figure 2.7: A particular network

$$\bar{F} = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ 3 & 1 & 0 & 0 & 0 & 1 \\ 4 & 0 & 1 & 0 & 0 & 1 \\ 5 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\bar{N} = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 0 & 0 & 1 & 1 \\ 2 & 0 & 0 & 1 & 0 & 1 \\ 3 & 0 & 1 & 0 & 1 & 0 \\ 4 & 1 & 0 & 1 & 0 & 0 \\ 5 & 1 & 1 & 0 & 0 & 0 \end{array}$$

As approximate phase-type representations of the distributions for channel busy time and idle time sensed by each node, we use the same representations that have been estimated in this section. Hence, we have $E[Q_{is}]$=0.6034.

In the present example, there are two types of packets namely, the one that traverses the route $1 \to 3 \to 5 \to$ GW and the other having the route $2 \to 4 \to 5 \to$ GW. Suppose that the inflow of packets to the system obey Poisson rule of rate $\lambda = 1.0629$, of which both types claim the same proportion. That is, $q_i = \frac{1}{2}$ for $i = 1, 2$. Table 2.2

presents a few values for the joint system size probabilities of packets at nodes in our model, under both case 1 and case 2 discussed in the previous section.

n	$P(n)$	n	$P(n)$
(1,2,1,1,2)	0.000106	(1,1,1,3,2)	0.000034
(1,1,2,2,1)	0.000053	(3,1,1,2,1)	0.000017
(2,2,1,1,2)	0.000034	(1,1,2,2,3)	0.000022
(1,1,1,3,3)	0.000022	(3,1,1,1,2)	0.000034
(1,1,3,2,1)	0.000017	(1,1,1,1,2)	0.000332
(1,4,1,1,3)	0.000007	(2,1,1,2,2)	0.000034
(1,1,2,2,2)	0.000034	(3,1,1,1,3)	0.000022
(2,1,1,2,3)	0.000022	(1,2,2,1,2)	0.000034
(1,1,2,2,1)	0.000053	(1,1,1,3,3)	0.000022
(2,1,1,2,3)	0.000022	(1,1,1,3,1)	0.000053

Table 2.2: Joint system size probabilities

For computing the joint system size probabilities, as displayed in table 2.2, the normalization constant is taken as the sum of the probabilities corresponding to state vectors $n = (n_1, n_2, n_3, n_4, n_5)$ for each n_i varies over 0 to 50.

Chapter 3

A Queuing Model for Wireless Network Handling Packets of Emergency in Nature

3.1 Introduction

Research on technological support in disaster and other emergency situations has become increasingly common. Wireless sensor networks has gained wonderful attention both in academia and industry because of its massive application potential. One of the critical areas of research related to wireless sensor network is to develop energy-saving techniques, which can extend the lifetime of a wireless sensor network as much as possible. Many applications (such as intrusion detection, emergency alarm, fire alarm detection, SOS messages) require immediate, and guaranteed deliverability of the information otherwise the information looses its relevance. Therefore, in an Internet of thing (IoT) environment, sensor networks provide reliable and secure data transportation. In a resource constraint environment such as wireless sensor networks, it is important to propagate the information as efficiently as possible among the neighboring nodes. The energy-consumption effects other factors such as end-to-end delay, hence plays a crucial role.

When an emergency packet occurs in one node, that node has to wait for its assigned slot to transmit such a life-critical packet which causes unbearable delay. Li et al. [53] proposed a new scheme named Quasi-Sleep-Preemption-Supported (QS-PS) to tackle this problem. Duy and Castro [33] proposed a network architecture that uses cellular networks and WiFi connections to deliver large files in emergency scenarios

under the impairments of wireless channel such as packet losses and intermittent connection issues. Ameen [2] described a on-demand emergency packet transmission scheme for wireless body area networks.

Inspired by this assumption, here we consider a wireless sensor network model that handles emergency packets. Like in the case of the model studied in the previous chapter, here also we use standard BEB scheme for collision avoidance. However, we take exponential distribution assumptions for channel busy and idle periods unlike in the previous case, where we have used more general PH distributions for the same.

Here, we are not attempting to compute the steady state distribution of the system size. Rather, we just focus our attention on deriving the probability distributions of packets at nodes after being selected for transmission and the number of collisions experienced by them before they are successfully transmitted/ timed out.

3.2 Mathematical model

Based on the sequence of events that happen in connection with the traffic flow at a node and the rule governed by the standard BEB scheme, the following assumptions are made :

- Data packets arrive at the node and waiting in the queue till they are being considered for transmission. Let the arrival process be Poisson of rate λ.

- At an instant at which a packet is considered for transmission, the back-off period for the node starts if the channel is sensed as idle, and a value for back-off counter is uniformly selected from 0, 1, 2,... $W-1$. If the packet has already experienced j collisions, then the back off counter will be from 0, 1, 2,..., $W_j - 1$ where $W_j = 2^j W$. Also, time spent on each of the back off counters are assumed to be independent and identically distributed exponential variates having mean $1/\mu$.

- If the channel is found busy after expiring a back off counter time, the back off timer will be frozen and resume only after the channel is sensed as idle. The channel idle periods and busy periods are assumed to be independent exponential variates having mean $\frac{1}{\theta}$ and $\frac{1}{\delta}$ respectively.

- When the back-off counter at a particular back-off stage becomes zero, the node starts transmission irrespective of the fact that if the channel is busy or

idle. Packet transmission times are assumed to be independent and identical exponential variates having mean $1/\gamma$.

- A transmission results in collision with probability p and is successful with probability $1 - p$.
- We assume that packets are of emergency in nature. So, if the packets cannot be transmitted within an amount of time after being generated (for convenience, we call this *packet life times*), their relevance will be lost. Life time of packets are assumed as PH distributed having representation (α, T) of order n.

The waiting time process of a packet at a node, after being considered for transmission, can be viewed as a CTMC. In relation to a particular node, let the state variables be defined as follows:

- $J(t)$, 0 or 1 according as the channel is idle or busy at t, as sensed by the node.
- $S(t)$, the back-off stage at t.
- $B(t)$, the back off counter at t. Note that when $S(t) = i$, $B(t)$ assumes one of the values uniformly from $0, 1, 2, \ldots, W_i - 1$, where $W_i = 2^i W$.
- $Z(t)$, the phase of life time period of the packet at t.

Then, the process $\{\phi(t) : t \geq 0\} = \{(J(t), S(t), B(t), Z(t)) : t \geq 0\}$ is a continuous time Markov chain with state space $E = \cup_i \cup_j E_{ij}$, where $E_{ij} = \{j\} \times \{i\} \times \{0, 1, 2, ..., W_i - 1\} \times \{1, 2, ..., n\}$ for $i = 0, 1, 2, \ldots, m$ and $j = 0, 1$.
In detail, the states are defined as $\{\overline{(0, i)}, \overline{(1, j)}\}$, where $i, j = 0, ..., m$. Now,$\overline{(0, i)} = \{\overline{(0, i, k)} : k = 0, 1, .., W_i - 1\}$ and $\overline{(1, j)} = \{\overline{(1, j, l)} : l = 0, 1, .., W_j - 1\}$.
For our convenience, we define $\overline{(0, i, 0)} = \overline{(i, 0)}$. So, $\overline{(i, 0)} = \{\overline{(i, 0, 1)}, .., \overline{(i, 0, n)}\}$ for $0 \leq i \leq m$. Assume that the states are arranged lexicographically.

Let Q be the infinitesimal generator of the process $\phi(t)$. Then Q can be written

as

$$Q = \left[\begin{array}{c|ccccccccccc} & \overline{(0,0)} & \overline{(0,1)} & \overline{(0,2)} & \cdots & \overline{(0,m)} & \overline{(1,0)} & \overline{(1,1)} & \overline{(1,2)} & \cdots & \overline{(1,m)} \\ \hline \overline{(0,0)} & D_0+C_0 & B_1 & 0 & \cdots & 0 & E_0 & 0 & 0 & \cdots & 0 \\ \overline{(0,1)} & C_1 & D_1 & B_2 & \cdots & 0 & 0 & E_1 & 0 & \cdots & 0 \\ \overline{(0,2)} & C_2 & 0 & D_2 & \cdots & 0 & 0 & 0 & E_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \overline{(0,m)} & C_m & \cdot & \cdot & \cdots & D_m+B_m' & 0 & 0 & 0 & \cdots & E_m \\ \overline{(1,0)} & F_0 & 0 & 0 & \cdots & 0 & G_0 & 0 & 0 & \cdots & 0 \\ \overline{(1,1)} & 0 & F_1 & 0 & \cdots & 0 & 0 & G_1 & 0 & \cdots & 0 \\ \overline{(1,2)} & \cdot & \cdot & F_2 & \cdots & 0 & 0 & 0 & G_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \overline{(1,m)} & \cdot & \cdot & \cdot & \cdots & F_m & 0 & 0 & 0 & \cdots & G_m \end{array}\right].$$

Different block matrices appearing in Q are explained and exhibited as follows.

$$D_i = \left[\begin{array}{c|ccccc} & \overline{(i,0)} & \overline{(0,i,1)} & \overline{(0,i,2)} & \cdots & \overline{(0,i,W_i-1)} \\ \hline \overline{(i,0)} & T+T^0\alpha-\gamma I & 0 & 0 & \cdots & 0 \\ \overline{(0,i,1)} & \mu I & T+T^0\alpha-(\theta+\mu)I & 0 & \cdots & 0 \\ \overline{(0,i,2)} & 0 & \mu I & T+T^0\alpha-(\theta+\mu)I & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \overline{(0,i,W_i-1)} & 0 & 0 & 0 & \cdots & T+T^0\alpha-(\theta+\mu)I \end{array}\right]$$

In D_i, the transitions among the states in the state vector $\overline{(0,i,k)}$ are just transitions among intermediate states of the life time phase process without decrementing the back off counter and the transitions from the states $\overline{(0,i,k)}$ to $\overline{(0,i,k-1)}$ correspond to decrement in the back off counter by 1 with rate μ.

$$B_i = \left[\begin{array}{c|cccccc} & \overline{(i,0)} & \overline{(0,i,1)} & \overline{(0,i,2)} & \cdots & \overline{(0,i,W_i-1)} \\ \hline \overline{(i-1,0)} & \frac{p\gamma e\alpha}{W_i} & \frac{p\gamma e\alpha}{W_i} & \frac{p\gamma e\alpha}{W_i} & \cdots & \frac{p\gamma e\alpha}{W_i} \\ \overline{(0,i-1,1)} & 0 & 0 & 0 & \cdots & 0 \\ \overline{(0,i-1,2)} & 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & 0\cdot & \cdot & \cdots & \cdot \\ \overline{(0,i-1,W_{i-1}-1)} & 0 & 0 & 0 & \cdots & 0 \end{array}\right].$$

B_i defines the transitions when a packet, already experienced $i-1$ collisions, is again involved in another collision (with probability p). In this case, the back off stage of the node is changed to i and back off counter can be chosen as one of the values uniformly from $0, 1, 2, \ldots, W_i - 1$.

$$C_i = \left[\begin{array}{c|ccccc} & \overline{(0,0)} & \overline{(0,0,1)} & \overline{(0,0,2)} & \cdots & \overline{(0,0,W-1)} \\ \hline \overline{(i,0)} & \frac{(1-p)\gamma e\alpha}{W_i} & \frac{(1-p)\gamma e\alpha}{W_i} & \frac{(1-p)\gamma e\alpha}{W_i} & \cdots & \frac{(1-p)\gamma e\alpha}{W_i} \\ \overline{(0,i,1)} & 0 & 0 & 0 & \cdots & 0 \\ \overline{(0,i,2)} & 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & 0\cdot & \cdot & \cdots & \cdot \\ \overline{(0,i,W_i-1)} & 0 & 0 & 0 & \cdots & 0 \end{array}\right].$$

C_i defines the transitions corresponding to the case when a packet experience successful transmission at back off stage i.

$$F_i = \left[\begin{array}{c|ccccc} & \overline{(i,0)} & \overline{(0,i,1)} & \overline{(0,i,2)} & \cdots & \overline{(0,i,W_i-1)} \\ \hline \overline{(1,i,1)} & 0 & \delta I & 0 & \cdots & 0 \\ \overline{(1,i,2)} & 0 & 0 & \delta I & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \overline{(1,i,W_i-1)} & 0 & 0 & 0 & \cdots & \delta I \end{array}\right]$$

The entries of F_i signify the rates in connection with the completion of the channel busy time and subsequent beginning of the channel idle period.

$$E_i = \left[\begin{array}{c|ccccc} & \overline{(1,i,0)} & \overline{(1,i,1)} & \overline{(1,i,2)} & \cdots & \overline{(1,i,W_i-1)} \\ \hline \overline{(i,0)} & 0 & 0 & 0 & \cdots & 0 \\ \overline{(0,i,1)} & \theta I & 0 & 0 & \cdots & 0 \\ \overline{(0,i,2)} & 0 & \theta I & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \overline{(0,i,W_i-1)} & 0 & 0 & 0 & \cdots & \theta I \end{array}\right]$$

Here, the entries of E_i represent the rates associated with the expiry of the channel idle time and hence beginning of its busy period.

$$G_i = \left[\begin{array}{c|ccccc} & \overline{(1,i,1)} & \overline{(1,i,2)} & \overline{(1,i,3)} & \cdots & \overline{(1,i,W_i-1)} \\ \hline \overline{(1,i,1)} & T+T^0\alpha-\delta I & 0 & 0 & \cdots & 0 \\ \overline{(1,i,2)} & 0 & T+T^0\alpha-\delta I & 0 & \cdots & 0 \\ \overline{(1,i,3)} & 0 & 0 & T+T^0\alpha-\delta I & \cdots & 0 \\ . & . & . & . & \cdots & . \\ \overline{(1,i,W_i-1)} & 0 & 0 & 0 & \cdots & T+T^0\alpha-\delta I \end{array}\right]$$

In G_i, the transitions among the states in the state vector $\overline{(1,i,k)}$ are just transitions among intermediate states of the life time phase process.

Then steady state distribution of the CTMC $\{\phi(t) : t \geq 0\}$ is obtained as

$$\pi_{i0} = \pi_{00}p^i \text{ for } 1 \leq i \leq m-1, \tag{3.2.1}$$

$$\pi_{m0} = \pi_{00}(p^m/1-p), \tag{3.2.2}$$

$$\pi_{0ki} = \frac{(-1)^{i-1}}{\mu^i}\pi_{k0}C_k^{'}(B_2^{'}+\delta\theta B_3^{'})^{i-1} + \sum_{l=1}^{i-1}\frac{(-1)^l}{\mu^l}\frac{\pi_{k0}\gamma e\alpha}{W_k}(B_2^{'}+\delta\theta B_3^{'})^{l-1}, \tag{3.2.3}$$

$$\text{and } \pi_{1ki} = \theta\pi_{0ki}B_3^{'} \text{ for } 1 \leq i \leq W_k-1, 0 \leq k \leq m \tag{3.2.4}$$

where $C_k^{'} = \gamma(I - \frac{e\alpha}{W_k}) - T - T^0\alpha, B_1^{'} = C_0^{'}, B_2^{'} = T + T^0\alpha - (\theta+\mu)I$ and $B_3^{'} = (\delta I - T - T^0\alpha)^{-1}$.
Now, π_{00} can be calculated by the normalizing condition

$$\sum_{k=0}^{m}\sum_{i=1}^{W_k-1}\pi_{0ki}\mathbf{e} + \sum_{k=0}^{m}\sum_{i=1}^{W_k-1}\pi_{1ki}\mathbf{e} + \sum_{k=0}^{m}\pi_{k0}\mathbf{e} = 1.$$

Let τ be the probability that a node transmits a packet (before the packet is being

timed out) at an arbitrary point of time.Then

$$\tau = \gamma \sum_{i=0}^{m} \pi_{i0} b, \tag{3.2.5}$$

where b is a column vector having n components with $b_j = e_j^T(\gamma I - T)^{-1}\mathbf{e}$.

3.2.1 Distribution of time since a packet is ready for transmission till it successfully transmitted/ timed out

As we discussed, even though a packet at a node is being considered for transmission, it can not be transmitted immediately but it has to undergo a random back off time before it gets transmitted. Then the transmission may result in a collision and in that case it has to undergo another random back off time and the process continues in this manner. During this process , even the life time of the packet may get over before its successful transmission and in this case the packet will be dropped. So, naturally we are interested in knowing the probability distribution of the time since a packet is ready for transmission till it is transmitted/ timed out.

Let U be the duration of time from the instant at which a packet (at an arbitrary node) is ready for transmission till it is either dropped or transmitted. Then, U is a phase-type random variate with representation (β, S).
Now,

$$S = \left[\begin{array}{c|cccccccccc}
 & \overline{(0,0)} & \overline{(0,1)} & \overline{(0,2)} & \cdots & \overline{(0,m)} & \overline{(1,0)} & \overline{(1,1)} & \overline{(1,2)} & \cdots & \overline{(1,m)} \\ \hline
\overline{(0,0)} & D_0' & B_1 & 0 & \cdots & 0 & E_0 & 0 & 0 & \cdots & 0 \\
\overline{(0,1)} & 0 & D_1' & B_2 & \cdots & 0 & 0 & E_1 & 0 & \cdots & 0 \\
\overline{(0,2)} & 0 & 0 & D_2' & \cdots & 0 & 0 & 0 & E_2 & \cdots & 0 \\
\cdots & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\
\overline{(0,m)} & 0 & \cdot & \cdot & \cdots & D_m' + B_m' & 0 & 0 & 0 & \cdots & E_m \\
\overline{(1,0)} & F_0 & 0 & 0 & \cdots & 0 & G_0' & 0 & 0 & \cdots & 0 \\
\overline{(1,1)} & 0 & F_1 & 0 & \cdots & 0 & 0 & G_1' & 0 & \cdots & 0 \\
\overline{(1,2)} & \cdot & \cdot & F_2 & \cdots & 0 & 0 & 0 & G_2' & \cdots & 0 \\
\cdots & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\
\overline{(1,m)} & \cdot & \cdot & \cdot & \cdots & F_m & 0 & 0 & 0 & \cdots & G_m'
\end{array}\right]$$

where E_i, F_i are the same as those defined earlier and $D_i^{'}$ and $G_i^{'}$ are defined as

$$D_i^{'} = \begin{bmatrix} T-\gamma I & 0 & \cdots & 0 & 0 \\ \mu I & T-(\theta+\mu)I & \cdots & 0 & 0 \\ . & . & . & \cdots & . \\ 0 & 0 & \cdots & \mu I & T-(\theta+\mu)I \end{bmatrix}, G_i^{'} = I_{W_i-1} \otimes (T-\delta I_n).$$

If the initial probability vector β is partitioned as $\beta = (\overline{Y_{00}}, \overline{Y_{01}}, ..., \overline{Y_{0m}}, \overline{Y_{10}}, \overline{Y_{11}}, .., \overline{Y_{1m}})$ corresponding to the lexicographical ordering of the states as explained earlier, then

$$\begin{aligned} \overline{Y_{00}} &= \pi_{00}(e-\gamma b)\alpha + \sum_{i=0}^{m} \pi_{i0} \frac{\gamma(1-p)b\alpha}{W} \\ \overline{Y_{00l}} &= \pi_{00l}(e-(\theta+\mu)d)\alpha + \sum_{i=0}^{m} \pi_{i0} \frac{\gamma(1-p)b\alpha}{W}, \text{ for } 1 \le l \le w-1 \\ \overline{Y_{i0}} &= \pi_{i0}(e-\gamma b)\alpha, \text{ for } 1 \le i \le m \\ \overline{Y_{0il}} &= \pi_{0il}(e-(\theta+\mu)d)\alpha \\ \overline{Y_{1il}} &= \pi_{1il}(e-\delta f)\alpha, \text{ for } 1 \le i \le m \text{ and } 1 \le l \le w_i - 1 \end{aligned} \tag{3.2.6}$$

where $d_j = e_j^T((\theta+\mu)I - T)^{-1}e$ and $f_j = e_j^T(\delta I - T)^{-1}e$.

Hence, the density of U is

$$f(u) = \beta e^{Su}(-S)e, \qquad 0 < u < \infty. \tag{3.2.7}$$

So, the average time a packet spends at a node after being selected for transmission,

$$E[U] = \beta(-S)^{-1}e \tag{3.2.8}$$

with the variance

$$Var[U] = 2\beta(S)^{-2}e - (\beta(-S)^{-1}e)^2.$$

3.2.2 Probability mass function of the number of collisions experienced by a packet

In order to find out the pmf of the number of collisions experienced by a packet, we consider the Markov process $\{(J(t), B(t), S(t), Z(t)) : t \geq 0\}$ with $\triangle$ as the absorbing state, which corresponds to the successful transmission/ time out of the packet. It is to be noted that for a node, back off stage k means the packet that is being considered for transmission has already experienced k collisions. By arranging all the transient states lexicographically and listing the absorbing state $\triangle$ as the last one, we get the generator

$$\hat{Q} = \begin{bmatrix} K_0 & J_1 & 0 & 0 & \cdots & 0 & L_0 \\ 0 & K_1 & J_2 & 0 & \cdots & 0 & L_1 \\ 0 & 0 & K_2 & J_3 & \cdots & 0 & L_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & K_m + J_m^{'} & L_m \end{bmatrix}$$

where

$$K_i = \begin{bmatrix} D_i^{'} & E_i \\ F_i & G_i^{'} \end{bmatrix}, L_i = \begin{bmatrix} H_i \\ N_i \end{bmatrix} \text{ for } i = 0, 1, \ldots m;$$

$$H_i = e^{(W_i)} \otimes T^0 + e_1^{(W_i)} \otimes \gamma e^{(n)} + e^{(W_i)} \otimes (1-p)\gamma e_1^{(n)}, N_i = e^{(W_i - 1)} \otimes T^0$$

$$J_i = \begin{bmatrix} B_i & 0 \\ 0 & 0 \end{bmatrix} \text{ for } i = 1, \ldots m,$$

$$\text{and } J_m^{'} = \begin{bmatrix} B_m^{'} & 0 \\ 0 & 0 \end{bmatrix}.$$

Here $e^{(j)}$ represents a column vector of order j and $e_i^{(j)}$ denotes a vector of order j having only non-zero component 1 at the ith position.

Let y_k be the probability that the packet experiences exactly k collisions before its successful transmission. Then

$$y_0 = \beta(-K_0)^{-1} L_0$$

$$y_k = \beta \prod_{j=1}^{k} (-K_{j-1})^{-1} J_j (-K_k)^{-1} L_k, \text{ for } k = 1, 2 \ldots m-1$$

and

$$y_{k+m} = \beta \prod_{j=1}^{m} (-K_{j-1})^{-1} J_j (-(K_m + J_m^{'}))^{-(k+1)} L_m, \text{ for } k \geq 0,$$

where β is given by eqn.(3.2.6). Therefore, expected number of collisions experienced by the packet

$$E[C] = \sum_{k=1}^{\infty} k y_k$$

and its variance

$$Var[C] = \sum_{k=1}^{\infty} k^2 y_k - E[C]^2.$$

3.3 Numerical illustration

In order to illustrate the performance of the system, we are presenting some numerical results based on our theoretical findings. In table 3.1, we exhibit the probabilities that the waiting time of a packet does not exceed some pre-determined values t till it is successfully transmitted/ timed out, against the variation of θ, the idle time parameter. For the illustration, we take $n = 3$, $\mu = 1$, $\gamma = 1$, $w = 2$, $p = 0.05$, $m = 3$, $\delta = 0.6$, $\alpha = [1/3, 1/2, 1/6]$, and

$$T = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1.2 & 0 \\ 0 & 0 & -0.8 \end{bmatrix}$$

	θ=0.2	0.4	0.6	0.8
t=2	0.9522	0.9517	0.9511	0.9470
3	0.9874	0.9872	0.9865	0.9848
4	0.9966	0.9965	0.9961	0.9954
5	0.9991	0.9990	0.9988	0.9985

Table 3.1: Cumulative distribution function of waiting time of a packet since it is ready for transmission till it is successfully transmitted/ timed out vs θ

From table 3.1, it can be seen that, for a fixed t, the above said probabilities are decreasing with the increase in values of θ whereas, for a fixed θ, they are increasing

with the increase in values of t. This is due to the fact that as θ increases, probability that the node senses a busy channel increases so that the proportion of time a packet waits a time less than a pre-determined value decreases.

By taking the same parameters as considered as above and assuming $\theta = 0.6$, in table 3.2, we exhibit the probabilities that the waiting time does not exceed some pre-determined values till it is successfully transmitted/ timed out corresponding to different values of δ, the channel busy parameter.

	δ=0.2	0.4	0.6	0.8
t=2	0.9503	0.9507	0.9511	0.9522
3	0.9862	0.9863	0.9865	0.9868
4	0.9957	0.9959	0.9961	0.9963
5	0.9986	0.9987	0.9988	0.9989

Table 3.2: Cumulative distribution function of waiting time of a packet since it is ready for transmission till it is successfully transmitted/ timed out vs δ

Table 3.2 shows the variation of the said probabilities against that of δ and t. It can be seen that for a fixed value of t, both δ and the probabilities are moving in the same direction. A similar relation exists between t and the probabilities, corresponding to a fixed value of δ.

p	$E[C]$	p	$E[C]$
0.01	0.7965	0.04	1.2964
0.02	0.8712	0.05	1.6827
0.03	1.0214	0.06	2.003

Table 3.3: Expected number of collisions experienced by a packet verses collision probability

Table 3.3 displays the variation of $E[C]$, the mean number of collisions experienced by a packet, against the variation of collision probability p. As is expected, $E[C]$ increases with the increase in values of p.

Chapter 4

Two-way Communication Orbit Queue Model with Constant Retrial Rates

4.1 Introduction

Loss models are those where a customer, upon seeing the server as busy at its arrival, takes a decision to leave the system. These customers are called blocked customers. There are other cases where a customer sees the server as busy at its arrival gets the service at a later time by waiting for its turn in an infinite buffer in the system. But many other real life situations are there in which the blocked customers are not patient enough to wait and may decide to leave the system initially, but may try after some random time to get a free server from outside the system. In such cases, we may assume a blocked customer waits in a virtual waiting space (infinite capacity *orbit*) outside the system before retrying to get the server back. These queueing situations are mainly modeled as retrial queues.

Retrial queues are broadly used in modelling many practical problems such as those related to call centres, computer networks, cellular networks, medium access protocols in wired and wireless networks etc. A detailed review of retrial queueing literature can be found, for example, in [3, 4, 38, 39, 47, 75]. Specific features of the retrial queue model such as retrial discipline, customer patience, number of servers, number of customer classes etc. may dramatically complicate the analysis.

The two main retrial disciplines considered in the literature are *classical retrial* (where blocked orbital customers retry independently) and *constant retrial* (where the retrial rate is fixed). In the latter case, the retrying customers may be considered as waiting in the *orbit queue*, with only the oldest customer retrying to approach the server. It is worth mentioning that the stability conditions of these two models are

quite different. The system with classical retrial discipline is stable under the same conditions as the corresponding buffered system due to the so-called *asymptotically work conserving property* [57]. At the same time, constant retrial rate discipline makes stability analysis much more involved. Here, we focus our attention on the latter discipline.

A significant part of the literature related to retrial queues addresses the situation where the server remains idle, after a service completion, until a new arrival or a retrial attempt happens. From a call centre illustration perspective, this means that, after serving an incoming call, the server always waits for the next call to come. But in many applications, apart from attending the incoming calls, the server (during idle periods) may need to perform some internal activity (maintenance, vacation) or initiate outgoing calls. Models associated with this kind of situations are usually referred as *two-way communication* models and such centres are called *blended call centres*. (It is more natural to think of a customer retrying from orbit, whereas from the server's perspective it might be thought as an incoming/outgoing call. Thus, hereafter we use the words *call* and *customer* as synonyms, adopting the name *service time* to designate call duration.)

To begin with, we mention some recent results for multi-class systems governed by *classical retrial* discipline. Despite a major difference in retrial disciplines for models with classical and constant retrial rates, performance analysis and some analytic results are similar for both the systems, see Morozov and Phung-Duc [60] for details. Also, in [60], necessary and sufficient stability conditions for GI/G/c-type retrial system with outgoing calls and feedback, were obtained. Shin and Moon [66] derived stability criterion for M/M/c-type system by producing approximations to stationary performance measures. Avrachenkov et al. [13] studied a single server multi-class retrial model with marked Markovian arrival process (MMAP) and two classes of customers with different service time distributions and retrial rates (from two separate orbits), however, with one orbit of limited capacity.

References [14, 16, 58] are dedicated to stability analysis of multi-class systems with *constant retrial rate*. Avrachenkov et al. [14] considered GI/G/c-type and M/G/1-type systems, and necessary stability conditions were derived. Avrachenkov et al. [16] obtained sufficient stability conditions for M/G/c-type system. Morozov and Dimitriou [58] addressed an M/G/1-type model with two *coupled queue orbits*, for which necessary and sufficient stability conditions were obtained. We also mention the recent works [15, 68], though they consider a simplified system with service rate,

independent of customer class.

Now, we look into the literature concerning *two-way communication models*. Bhulai and Koole [18] proposed a multi-server queuing model with two-way communication by assuming identical service time distributions for both incoming and outgoing calls. Deslauriers et al. [28] developed five Markovian queueing models for blended call centres, where incoming and outgoing calls are distinguished as well as undistinguished. But, in both these papers, retrial assumptions were not taken into account. Falin [37] and Choi et al. [22] found some explicit expressions for performance measures of an M/G/1/1 retrial system with two-way communication, in which incoming calls and outgoing calls were assumed to follow the same service time distribution. Later, Artalejo and Resing [7], by using mean value analysis, succeeded in getting some measures for M/G/1/1 retrial queues with classical retrial discipline, and having different service time distributions for incoming and outgoing calls. Avrachenov et al. [13] studied a single server retrial model with marked Markovian arrival process (MMAP) and two classes of customers with different service times and retrial discipline. Later, Artajelo and Phung-Duc [8, 9] analysed the steady state behaviour of a single server classical retrial queue with two-way communications with exponential and general service times, respectively. Further, Sakurai and Phung-Duc [69] extended the study to two-way communication retrial models having multiple classes of outgoing calls with class-dependent exponential service times. Phung Duc et al. [65] studied both single-server and multi-server retrial models with constant retrial rate and short term balanced call blending. A Poisson input, general service time, single-server, two-way communication model with constant retrial rate was studied by Aissani and Phung-Duc [12], for which stability conditions as well as some performance measures were obtained.

The model which is the most similar to the one considered in this chapter is studied in the recent work by Morozov and Phung-Duc [61], where a M/G/1/1-type constant retrial rate model with two-way communication is considered. This is a model with single class of arriving customers and multiple classes of outgoing calls. In another closely related work by Sakurai and Phung-Duc [69], a similar M/M/1/1-type model is analysed.

It is to be noted that in the present chapter, we extend the models studied by Morozov and Phung-Duc [61] and Sakurai and Phung-Duc [69] to multi-class multi-orbit single server system. In section 4.2, we highlight the novelty of results obtained here in comparison to [61, 69]. It is worth mentioning that extension to

multiple orbits dramatically complicates the analysis not only due to complicated state space, but also because of a dependence between the orbits. We elaborate more on this aspect in sections (4.2) and (4.6).

The contribution of this chapter is threefold. First, we extend the model studied by Morozov and Phung-Duc [61] to multiple classes of incoming customers joining multiple orbits with class-dependent probabilities, which makes the analysis more complex. In particular, we develop a new coupling-based approach which allows to connect the process of retrial attempts from each orbit with an independent Poisson process. Using the new approach, combined with PASTA property, we obtain some stationary performance measures which have not been available for such general retrial systems. After obtaining performance measures, we address the stability conditions yet unknown, to the best of our knowledge, for the considered system. Secondly, using a restricted single-class model studied by Sakurai and Phung-Duc [69], we demonstrate a methodological advantage that might be used to explicitly obtain the stationary probabilities of system states by combining the regenerative and matrix-analytic methods. Finally, we extend the applicability of the stationary performance analysis by performing extensive numerical experiments

The structure of the chapter is as follows. In section (4.2), we introduce the single-class single-orbit system with some preliminary results obtained earlier. In section (4.3), we extend this model to a multi-orbit model with no outgoing calls (assuming general service time) and discuss some performance measures along with necessary and sufficient stability conditions of the model, where coupling method is used extensively. In section (4.4), a model with different classes of outgoing calls has been discussed. Section (4.5) deals with a Markovian model where all blocked customers join the same orbit. Finally in section (4.6), we validate our theoretical results with extensive simulation performed in R.

4.2 Preliminary results

We start our analysis with a single server, two-way communication system studied in [61] and [69], where a *single class* of outgoing calls was being considered. We begin with general service time case and then refine the results for exponentially distributed service times. Both cases are necessary for methods of exposition.

A single server accepts input of incoming calls arriving at epochs of Poisson process of rate λ, with generally distributed iid service times with mean $E[S] = 1/\mu$. Once the

server is busy, a customer is routed to a single orbit-queue (or, shortly, orbit). Orbital (retrial) customers try to access the server in gap of exponential amount of time with constant retrial rate $\lambda^{(r)}$. When the server is idle, it may initiate an outgoing call having exponential duration (with mean $E[Z] = 1/\mu^{(o)}$) after waiting for some exponentially distributed (with rate $\lambda^{(o)}$) period. We put superscripts $(i), (r)$ and (o) to designate the quantities related to incoming/retrying/outgoing call, respectively. Note that, at server, we do not distinguish the primary calls and retrying orbital customers, since both types have the same (general) service time distribution and thus are later referred as *incoming* calls.
Let

$$\rho = \frac{\lambda}{\mu} \text{ and } \rho^{(o)} = \frac{\lambda^{(o)}}{\mu^{(o)}}.$$

Now, let us introduce the following performance measures of which we are interested in:

$P_0 (P_b = P_b^{(i)} + P_b^{(o)})$ – idle (busy) probability of server;

$\pi_0^{(r)} (\pi_b^{(r)})$ – probability of empty (non-empty) orbit;

$P_{0,0}$ – empty system probability;

$P_{0,b}$ – probability of idle server together with non-empty orbit;

$P_{b,0}$ – probability of busy server and empty orbit;

$P_{b,b}$ – probability of busy server and non-empty orbit.

We also note the following interrelations between the aforementioned measures:

$$\begin{array}{ccccc} P_0 & = & P_{0,0} & + & P_{0,b} \\ + & & + & & + \\ P_b & = & P_{b,0} & + & P_{b,b} \\ \| & & \| & & \| \\ 1 & = & \pi_0^{(r)} & + & \pi_b^{(r)} \end{array} \tag{4.2.1}$$

The following results are adopted from Morozov and Phung-Duc [61]:

$$P_0 = \frac{1-\rho}{1+\rho^{(o)}} = 1 - P_b;$$

$$P_b^{(i)} = \rho; P_b^{(o)} = \rho^{(o)} \frac{1-\rho}{1+\rho^{(o)}};$$
$$P_{0,0} = \frac{1-\rho-\lambda/\lambda^{(r)}(\rho+\rho^{(o)})}{1+\rho^{(o)}};$$
$$P_{0,b} = \frac{\lambda}{\lambda^{(r)}} \frac{\rho+\rho^{(o)}}{1+\rho^{(o)}}.$$

The main methods used in Morozov and Phung-Duc [61] to obtain these performance measures were *regenerative approach*, balance equations for cumulative processes, and coupling technique. We apply similar techniques here to extend these results into multi-orbit system set up in section (4.3).

4.2.1 Exponentially distributed service times

Note that, in the particular case of exponentially distributed service times, the performance measures referred above are derived in Sakurai and Phung-Duc [69] by using the probability generating function approach as:

$$\begin{aligned}
P_b &= \rho = 1 - P_0; \\
P_{0,b} &= \rho \frac{\lambda}{\lambda^{(r)}}; \\
P_{0,0} &= P_0 - P_{0,b}; \\
P_{b,0} &= \rho P_{0,0}; \\
P_{b,b} &= \rho^2 \left(1 + \frac{\lambda}{\lambda^{(r)}}\right); \\
\pi_0^{(r)} &= (1+\rho) P_{0,0}; \\
\pi_b^{(r)} &= 1 - \pi_0.
\end{aligned}$$

Now, we demonstrate how the steady state distribution for the single incoming, single outgoing class Markovian orbit queue model, which is a particular case of the model discussed in Sakurai and Phung-Duc [69], can be computed by combining both *regenerative* and *matrix-analytic* approach. The same methodology will also be used in section (4.5) to derive the steady state system size distribution for a particular case of our model (a Markovian model where all multi-class incoming calls joining a single orbit). Note that our general model in this chapter is Non-Markovian (due to general service time assumptions) and incoming calls belonging to different class join in the corresponding class specific orbit.

In connection with the above referred particular case of the model discussed in Sakurai and Phung-Duc [69], we consider the two-dimensional Markov process $\{X(t) = (N(t), Q(t)) : t \geqslant 0\}$, where $N(t)$ (the *level*) is the number of customers in the orbit and $Q(t)$ (the *phase*) is the server state, which is encoded as follows: 0 — server idle, 1 — server busy with incoming call, 2 — server busy with outgoing call (routine).

Since, by assumption, the process $\{N(t)\}$ may change only by ± 1 at each time epoch, the process $\{X(t) : t \geqslant 0\}$ is Quasi-Birth-Death (QBD) with infinitesimal generator having the block-tridiagonal form, given by

$$Q = \begin{pmatrix} A^{0,0} & A_0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \ddots \\ 0 & 0 & A_2 & A_1 & \ddots \\ 0 & 0 & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4.2.2}$$

where the matrix $A^{0,0}$ corresponds to boundary states at zero level , and square matrices A_i, $i = 0, 1, 2$ of order 3 correspond to non-boundary states. We define these matrices explicitly as follows:

$$A_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \quad A_1 = \begin{pmatrix} -\lambda - \lambda^{(o)} - \lambda^{(r)} & \lambda & \lambda^{(o)} \\ \mu & -\lambda - \mu & 0 \\ \mu^{(o)} & 0 & -\lambda - \mu^{(o)} \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0 & \lambda^{(r)} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A^{0,0} = \begin{pmatrix} -\lambda - \lambda^{(o)} & \lambda & \lambda^{(o)} \\ \mu & -\lambda - \mu & 0 \\ \mu^{(o)} & 0 & -\lambda - \mu^{(o)} \end{pmatrix}.$$

Note that A_0 corresponds to arrival of customers *into the orbit*, and A_2 is the departure of customers *from the orbit*, which is possible only when the server becomes idle. We also note that A_1 corresponds to server occupation by external customer arrivals, as well as by service completion, while the diagonal elements of A_1 are derived from balance condition $A\mathbf{e} = \mathbf{0}$, where

$$A = A_0 + A_1 + A_2 = \begin{pmatrix} -\lambda - \lambda^{(o)} - \lambda^{(r)} & \lambda + \lambda^{(r)} & \lambda^{(o)} \\ \mu & -\mu & 0 \\ \mu^{(o)} & 0 & -\mu^{(o)} \end{pmatrix}. \tag{4.2.3}$$

We also stress that A_2 is of rank 1.

The condition under which the system is stable is given by the formula (see Neuts [63])

$$\alpha A_2 \mathbf{e} > \alpha A_0 \mathbf{e}, \tag{4.2.4}$$

where α is approximately the vector of probabilities of phases at higher levels (refer He [41]), derived from the system

$$\begin{cases} \alpha A = \mathbf{0} \\ \alpha \mathbf{e} = 1. \end{cases} \tag{4.2.5}$$

From (4.2.3) and (4.2.5), it is easy to obtain α, after some algebra, as

$$\alpha = \left(\mu\mu^{(o)}, (\lambda + \lambda^{(r)})\mu^{(o)}, \mu\lambda^{(o)}\right) \left((\mu + \lambda + \lambda^{(r)})\mu^{(o)} + \mu\lambda^{(o)}\right)^{-1}. \tag{4.2.6}$$

Thus, the stability criterion (4.2.4) reduces to

$$\frac{\lambda(\lambda + \lambda^{(r)})}{\mu\lambda^{(r)}} + \frac{\lambda\lambda^{(o)}}{\lambda^{(r)}\mu^{(o)}} < 1, \tag{4.2.7}$$

which is

$$\rho + \frac{\lambda}{\lambda^{(r)}}(\rho + \rho^{(o)}) < 1, \tag{4.2.8}$$

where $\rho = \lambda/\mu$ and $\rho^{(o)} = \frac{\lambda^{(o)}}{\mu^{(o)}}$.

Now, we note that $A_2 = cr$, where $c = (\lambda^{(r)}, 0, 0)$ is a column vector, and $r = (0, 1, 0)$ is a row vector. Then it follows from Latouche and Ramaswami [50] that $G = \mathbf{e}r$, and

$$R = -A_0(A_1 + A_0 G)^{-1}. \tag{4.2.9}$$

After some algebra, one obtains

$$R = \frac{\lambda}{\mu\lambda^{(r)}(\lambda + \mu^{(o)})} \begin{pmatrix} 0 & 0 & 0 \\ \mu(\lambda + \mu^{(o)}) & \lambda(\lambda + \lambda^{(o)} + \lambda^{(r)} + \mu^{(o)}) + \lambda^{(r)}\mu^{(o)} & \mu\lambda^{(o)} \\ \mu(\lambda + \mu^{(o)}) & \lambda(\lambda + \lambda^{(o)} + \lambda^{(r)} + \mu^{(o)}) & \mu(\lambda^{(o)} + \lambda^{(r)}) \end{pmatrix}. \tag{4.2.10}$$

Then the vector $\pi = (\pi_0, \pi_1, \dots)$, where $\pi_i = (\pi_{i,0}, \pi_{i,1}, \pi_{i,2})$ is the stationary probability of having i customers in the orbit, with phase $0, 1, 2$, respectively, is

obtained in the matrix-geometric form

$$\pi_i = \pi_0 R^i, \quad i \geqslant 1, \tag{4.2.11}$$

where π_0 is obtained from the system

$$\begin{cases} \pi_0(A^{0,0} + RA_2) = \mathbf{0} \\ \pi_0(I - R)^{-1}\mathbf{e} = 1. \end{cases} \tag{4.2.12}$$

After some straightforward calculations, from the first equation of the above system, one obtains

$$\pi_0 = \pi_{0,0}\left(1, \rho\big(1 + \frac{\lambda^{(o)}}{\lambda + \mu^{(o)}}\big), \frac{\lambda^{(o)}}{\lambda + \mu^{(o)}}\right),$$

whereas the second equation provides

$$\pi_{0,0} = \frac{1 - \rho - \frac{\lambda}{\lambda^{(r)}}(\rho + \rho^{(o)})}{1 + \rho^{(o)}}.$$

We note that the solution of the system (4.2.12) requires matrix inversion, which might be avoided by considering the following alternative system

$$\begin{cases} \xi_0(I - R)(A^{0,0} + RA_2) = \mathbf{0} \\ \xi_0\mathbf{e} = 1, \end{cases} \tag{4.2.13}$$

where $\xi_0 = \pi_0(I - R)^{-1}$. Then π_0 might be obtained by matrix multiplication $\pi_0 = \xi_0(I - R)$. Moreover, the vector ξ_0 itself is a performance measure of the system. In fact, $\xi_{0,i}$ is the stationary probability of server being in state $i = 0, 1, 2$, irrespective of the orbit size. Due to the model properties, the system (4.2.13) might be further transformed into the system

$$\begin{cases} \xi_0(A + (I - R)M) = \mathbf{0} \\ \xi_0\mathbf{e} = 1, \end{cases} \tag{4.2.14}$$

where $M = A^{0,0} + A_0 - A$ has non-zero elements only in the first row and, moreover, $M\mathbf{e} = \mathbf{0}$. This allows us to easily obtain the vector ξ_0 as

$$\xi_0 = \left(\frac{1 - \rho}{1 + \rho^{(o)}}, \rho, \frac{\rho^{(o)}(1 - \rho)}{1 + \rho^{(o)}}\right).$$

We note that $\xi_{0,0}$ is the probability that the server is idle.

4.3 Multi-orbit model with no outgoing calls

Now, we turn onto our model under consideration in this chapter. Initially, we assume that our model doesn't take into account any class of outgoing calls. Let us consider a single server retrial queue with M classes of (primary) customers. Class-k customers follow a Poisson input flow with rate $\lambda_k \in (0, \infty)$ and have independent identically distributed (iid) service times $\{S_n^{(k)}, n \geq 1\}$ with a general distribution with mean $E[S^{(k)}] = 1/\mu_k < \infty$. By seeing a busy server, a class-k customer joins the k-orbit, which follows a constant retrial policy. In other words, duration of successive retrial attempts from k-orbit follows exponential distribution with rate $\lambda_k^{(r)}$, and independent of the orbit size (the number of customers in the orbit).

Let

$$\lambda = \sum_k \lambda_k \text{ and } p_k = \frac{\lambda_k}{\lambda}, \ k = 1, \ldots, M.$$

Now, let us denote by $V(t)$ the summary work arrived in the system, $B(t)$ the busy time of the server, and $I(t)$ the idle time of the server, in interval $[0, t)$. Then, $B(t) + I(t) = t$. Also, let $V_k(t)$ be the summary work of class-k customers arrived in interval $[0, t)$ so that $V(t) = \sum_k V_k(t)$. Let $N_k(t)$ denotes the number of the orbital customers and $W_k(t)$ denotes the workload (remaining work) in k-orbit queue, at instant t^-. Let $S(t)$ be the remaining service time at instant t^-, i.e. the time the current customer, being in service, departs the server. (By definition, $S(t) = 0$ if the server is free at instant t.)

We consider the basic *non-Markovian* summary queue size process as defined as follows:

$$X(t) = \sum_k N_k(t) + Q(t), \ t \geq 0,$$

where $Q(t) \in \{0, 1\}$ is the number of calls in the server at instant t^-. Let $\{t_n, n \geq 1\}$ be the instants of the superposed input (Poisson) process with the rate λ. Denote $X(t_n) = X_n$, $n \geq 1$. Let $T_0 = 0$ and $T_n, n \geq 1$ be defined recursively by

$$T_{n+1} = \inf\Big(t_k > T_n : X_k = 0\Big), \ n \geq 0.$$

We assume zero initial state, in which case the first customer arrives in the idle

system at instant $t = 0$, and, with probability p_k, this is a class-k customer. It is easy to see that $\{T_n\}$ are classical regenerations of basic process X, with the iid *regeneration periods* $T_{n+1} - T_n$ (and with generic period T). It is well-known (for instance, Assmussen [11]) that if mean generic period $E[T] < \infty$, then the process $\{X(t)\}$ (and the basic system) is *positive recurrent* and there exists the weak limit $X(t) \Rightarrow X$, which is the stationary number of customers in the system.

Besides the performance measures that are already introduced in section (4.2), let us define $P_b^{(k)}$ as the probability of server being busy with class $k = 1, \ldots, M$ call; $P_{0,\,b}^{(k)}$ as the stationary probability that the server is idle and k-orbit is non empty; $P_{0,\,0}^{(k)}$ as the stationary probability that the server is idle and k-orbit is empty; $P_{b,\,b}^{(k)}$ as the stationary probability that the server is busy and k-orbit is non-empty and $P_{b,0}^{(k)}$ as the stationary probability that server is busy and k-orbit is empty.

Also, the load related to each class, and the summary load are given by

$$\rho_k = \frac{\lambda_k}{\mu_k} \quad \text{and} \quad \rho = \sum_{k=1}^{M} \rho_k$$

respectively.
Below, we obtain, by a simple regenerative approach, some important steady-state performance measures of the model, which have been found in Phung-duc et al [65], Sakurai and Phung-Duc [69] by the detailed Kolmogorov equations approach, for pure Markovian set up.

4.3.1 Analysis of steady-state regime

Theorem 4.3.1. *If the basic system is positive recurrent, then*

$$P_0 = 1 - \rho = 1 - P_b, \tag{4.3.1}$$

and for each $k = 1, \ldots, M$,

$$P_b^{(k)} = \rho_k, ; \tag{4.3.2}$$

$$P_{0,\,b}^{(k)} = \frac{\lambda_k}{\lambda_k^{(r)}}\rho; \tag{4.3.3}$$

$$P_{0,\,0}^{(k)} = 1 - (1 + \frac{\lambda_k}{\lambda_k^{(r)}})\rho. \tag{4.3.4}$$

Proof. . Let $A_k(t)$ denotes the number of class-k arrivals in interval [0, t). Then, we have the balance equation

$$V(t) \;=\; \sum_{k=1}^{M}\sum_{i=1}^{A_k(t)} S_i^{(k)} = S(t) + \sum_{k=1}^{M} W_k(t) + t - I(t). \tag{4.3.5}$$

Note that the remaining service time $S(t)$ may relate to each class-k customer, however, $S(t) = o(t)$ (see Morozov [56]) and, because of the assumed stability, $\sum_{k=1}^{M} W_k(t) = o(t)$ as $t \to \infty$ with probability 1 (w. p. 1) as well, see Smith [67], Morozov and Delgado [59]. Note that, for each k, with probability 1

$$\frac{1}{t}\sum_{i=1}^{A_k(t)} S_i^{(k)} = \frac{1}{A_k(t)}\sum_{i=1}^{A_k(t)} S_i^{(k)} \cdot \frac{A_k(t)}{t} \to \rho_k. \tag{4.3.6}$$

Then, as $t \to \infty$,

$$\lim_{t\to\infty} \frac{V(t)}{t} \;=\; \rho. \tag{4.3.7}$$

On the other hand,

$$\lim_{t\to\infty} \frac{1}{t}\Big(S(t) + \sum_{k=1}^{M} W_k(t) + t - I(t)\Big) = 1 - P_0. \tag{4.3.8}$$

Note that

$$\lim_{t\to\infty} \frac{I(t)}{t} = P_0 = P(X = 0)$$

exists w. p. 1 (because of the stationarity) where X , the weak limit of $X(t)$, denotes the stationary number of calls in the system (see Smith [67]). Now equation (4.3.1) follows from equations (4.3.5)-(4.3.8).
Since the work generated by class-k customers in [0, t) is given by $V_k(t) = \sum_{i=1}^{A_k(t)} S_i^{(k)}$, we have

$$V_k(t) = S_k(t) + B_k(t), \tag{4.3.9}$$

where $S_k(t)$ is the remaining service time of a class-k customer, provided it is with the server at t ($S_k(t) = 0$, if server is free or busy by other class customers), and $B_k(t)$

is the busy time the server devotes to class-k customers in interval $[0,\, t)$. Then

$$\frac{V_k(t)}{t} \to \rho_k, \;\; \frac{B_k(t)}{t} \to P_b^{(k)} \text{ as } t \to \infty,$$

and equation (4.3.2) follows.

It is more challenging to find $P_{0,b}^{(k)}$. Let $D_k(t)$ denotes the number of retrial class-k customers which depart k-orbit in $[0,\, t)$. Also, let $\hat{D}_k(t)$ be the number of renewals in interval $[0,\, t)$ generated by the iid exponential variables with parameter $\lambda_k^{(r)}$ (with $\hat{D}_k(0) = 0$). To connect $D_k(t)$ and $\hat{D}_k(t)$, we use a coupling as follows: We start at instant $t = 0$ and sample the process $\hat{D}_k$ until the 1st class-k customer joins the orbit at some arrival instant $t_n^{(k)}$ (of the Poisson input of class-k customers). For convenience, we denote $t_n^{(k)} = z_1^{(k)}$. From $z_1^{(k)}$ onwards, this customer begins to make retrial attempts. At instant $z_1^{(k)}$, we re-sample remaining renewal (exponential) time in the renewal process $\hat{D}_k$ and then synchronize inter renewal intervals in both processes D_k and $\hat{D}_k$ until the successful attempt of the 1st customer happens, at some instant $v_1^{(k)}$. If k-orbit remains non-empty, that is if $N_k(v_1^{(k)} + 0)) > 0$, then we continue to sample the identical inter renewal times (actually, intervals between the attempts) in both processes D_k and $\hat{D}_k$ until the successful attempt of the next class-k customer happens, at some instant $u_1^{(k)}$. If k-orbit becomes empty after instant $v_1^{(k)}$, that is if $N_k(v_1^{(k)} + 0)) = 0$, then we continue to sample only intervals in the process $\hat{D}_k$ until the next orbital class-k customer appears, at some instant $z_2^{(k)}$. At that, the process of real departures D_k remains "frozen" until instant $z_2^{(k)}$. Then, at instant $z_2^{(k)}$, we resample remaining interval in the process $\hat{D}_k$ and synchronize next inter renewal times in both processes as above, until successful attempt occurs at instant $u_2^{(k)}$, etc. By construction, the instants of the appearance of class-k customers and successive attempts of the (top) orbital customers in the process D_k are a subsequence of the renewal instants of the process $\hat{D}_k$ with resampling. (We will keep the same notation $\hat{D}_k$ for this "resampled" process which is stochastically equivalent to originally defined process $\hat{D}_k$.) Also, we keep notation $t_n^{(k)}$ for the instants of the appearance of the events in the (modified) process $\hat{D}_k$. Let us denote

$$Q(t_i^{(k)} - 0) = Q_i^{(k)}, \text{ and } N(t_i^{(k)} + 0) = N_i^{(k)}.$$

Then, we can define the number of customers which depart k-orbit in interval $[0,\, t)$

as

$$D_k(t) = \sum_{i=1}^{\hat{D}_k(t)} \mathbf{1}(Q_i^{(k)} = 0,\, N_i^{(k)} > 0), \tag{4.3.10}$$

where $\mathbf{1}$ stands for the indicator function. By construction, $D_k(t)$ is the number of Poisson arrivals (with rate $\lambda_k^{(r)}$) which meet *idle server and non-empty k-orbit*. Then $\lim_{t\to\infty} \frac{\sum_{i=1}^{\hat{D}_k(t)} \mathbf{1}(Q_i^{(k)}=0,\, N_i^{(k)}>0)}{\hat{D}_k(t)}$, which exists (w.p. 1 and in the mean) by the positive recurrence, is equal to $P_{0,b}^{(k)}$ since, by PASTA property, $P_{0,b}^{(k)}$ is the limiting fraction of the time when the server is free and simultaneously k-orbit is occupied. Thus, by equation (4.3.10)

$$\lim_{t\to\infty} \frac{D_k(t)}{\hat{D}_k(t)} = P_{0,b}^{(k)}. \tag{4.3.11}$$

Now, we introduce $\hat{A}_k(t)$, the number of class-k customers that join the orbit in interval $[0,\, t)$. Denote $\hat{Q}(t_i^-) = \hat{Q}_i$, the state of the server *just before* the ith arrival in the superposed input process. Let $\mathbf{1}(\zeta_i = k) = 1$ if the ith customer is class-k one. Then, we have

$$\hat{A}_k(t) = \sum_{i=1}^{A(t)} \mathbf{1}(\hat{Q}_i = 1,\, \zeta_i = k),$$

where $A(t)$ denotes the number of primary arrivals occurring in $[0, t)$. Since $E[\mathbf{1}(\zeta_i = k)] = \lambda p_k = \lambda_k$, the limit

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} \mathbf{1}(\hat{Q}_i = 1,\, \zeta_i = k)}{A(t)} = \lim_{t\to\infty} \frac{\hat{A}_k(t)}{A(t)} = p_k P_b = p_k \rho \tag{4.3.12}$$

exists, where $P_b = \rho$ is the stationary busy probability of the server, and we take into account the independence between $\hat{Q}_i$ and indicator $\mathbf{1}(\zeta_i = k)$. Clearly,

$$\hat{A}_k(t) = N_k(t) + D_k(t). \tag{4.3.13}$$

Note that $N_k(t) = o(t)$ as $t \to \infty$, by the positive recurrence, and that, by the renewal

theory, w.p.1,

$$\frac{\hat{D}_k(t)}{t} \to \lambda_k^{(r)},\ k = 1, \ldots, M. \tag{4.3.14}$$

Then, from equation (4.3.11),

$$\frac{D_k(t)}{t} \to \lambda_k^{(r)} P_{0,b}^{(k)},\ k = 1, \ldots, M. \tag{4.3.15}$$

Again

$$\frac{\hat{A}_k(t)}{t} = \frac{\hat{A}_k(t)}{A_(t)} \frac{A_(t)}{t} \to \lambda_k \rho \tag{4.3.16}$$

by equation (4.3.12) as $t \to \infty$. Then, from (4.3.15) and (4.3.16) we get (4.3.3).

Note that, for each k,

$$P_0 = 1 - \rho = P_{0,b}^{(k)} + P_{0,0}^{(k)},\ k = 1, \ldots, M. \tag{4.3.17}$$

Then, (4.3.4) follows from (4.3.1) and (4.3.3).

□

4.3.2 Stability analysis

In this section, we formulate and discuss stability conditions of the basic process X.

For an easy reading, we recall some previous results in this direction obtained, in particular, in [12, 14, 16, 65, 69].

First of all, we recall the (necessary) stability conditions for the M-orbit model *with no outgoing calls*, found in Avrachenkov et al. [16],

$$\lambda_k P_b < \lambda_k^{(r)} (1 - P_b),\ k = 1, \ldots, M, \tag{4.3.18}$$

where $P_b = \rho = \sum_{k=1}^{M} \rho_k$.

Remark 4.3.1. *It is worth mentioning that conditions* (4.3.18) *can be easily obtained*

from the expression

$$P_{0,\,b}^{(k)} = \frac{\lambda_k}{\lambda_k^{(r)}}\rho = \frac{\lambda_k}{\lambda_k^{(r)}}P_b, \quad \textit{given by } (4.3.3).$$

Indeed, since each orbit k is visited by class-k customers with a positive probability within a regeneration cycle, $\max_k P_{0,\,b}^{(k)} < P_0$ *so that*

$$P_0 > \frac{\lambda_k}{\lambda_k^{(r)}}P_b.$$

Hence the equation (4.3.18) *follows.*

Now, we discuss a *sufficient stability condition* for the model under consideration in this section. This form of condition is motivated by the requirement to have negative drift of the workload process.

We note that the idle time of the server (when there exists orbital customers) between two successive arrivals, happens after departure of a customer from the server. Note that the remaining time to next arrival (at the departure instant) is distributed as interarrival time τ. Let $\xi_0^{(i)}$ be an exponential variable with parameter $\lambda_i^{(r)}$, which describes the time between attempts from orbit i. Then, provided orbit i is non-empty, the idle time after departure is upper bounded by the variable $\min(\tau,\, \xi_0^{(i)})$ with the mean $1/(\lambda + \lambda_i^{(r)})$. Note that this idle time (delay) can be treated as a lost part of server capacity. On the other hand, it can be treated as an additional "service time" of the customer leaving server. To obtain the sufficient stability condition, we must take into account the "worst" case, means the maximum possible delay. Let $I_i = 1$ if a new arrival is class i one, and $I_i = 0$, otherwise, that is $E[I_i] = p_i$ (here and below we deal with generic variables). Now we consider an enlarged "service time" $\hat{S}$ which includes the above mentioned maximum delay so that

$$\hat{S} = \sum_{i=1}^{M} I_i\Big(S^{(i)} + \min(\tau,\, \xi_0^{(i)})\Big). \tag{4.3.19}$$

To obtain the negative drift of the workload process, we must have $E[\hat{S}] < E[\tau] = 1/\lambda$ (see Morozov [59]). Now,

$$E[\hat{S}] \;=\; \sum_{i=1}^{M} p_i E[S^{(i)}] + \sum_{i=1}^{M} p_i E[\min(\tau,\, \xi_0^{(i)})]$$

$$= \sum_{i=1}^{M} \frac{\lambda_i}{\lambda} \frac{1}{\mu_i} + \sum_{i=1}^{M} \frac{p_i}{\lambda + \lambda_i^{(r)}}$$

so that $E[\hat{S}] < \frac{1}{\lambda}$ implies

$$\sum_{i=1}^{M} \rho_i + \lambda \sum_{i=1}^{M} \frac{p_i}{\lambda + \lambda_i^{(r)}} < 1. \tag{4.3.20}$$

.

By using $\sum_i p_i = 1$ and slightly strengthening the latter inequality, we obtain a sufficient stability condition as explained below.

The left hand side of (4.3.20) yields

$$\begin{aligned} &\sum_{i=1}^{M} \rho_i + \lambda \sum_{i=1}^{M} \frac{p_i}{\lambda + \lambda_i^{(r)}} \\ &\leq \sum_{i=1}^{M} \rho_i + \max_{1 \leq i \leq M} \frac{\lambda}{\lambda_i^{(r)} + \lambda} < 1. \end{aligned} \tag{4.3.21}$$

Hence, this is a sufficient stability condition. We can rewrite equation (4.3.21) as

$$\rho = \sum_i \rho_i < 1 - \max_{1 \leq i \leq M} \frac{\lambda}{\lambda_i^{(r)} + \lambda} = \min_i \left(\frac{\lambda_i^{(r)}}{\lambda + \lambda_i^{(r)}} \right). \tag{4.3.22}$$

Also, it is to be noted that the necessary conditions given by (4.3.18) yield

$$\rho < \min_i \frac{\lambda_i^{(r)}}{\lambda_i + \lambda_i^{(r)}}. \tag{4.3.23}$$

Observe that the function $f(x) = \frac{x}{(\lambda + x)}$ is monotonically increasing in x and $\max_i \lambda_i < \lambda$. Then, we obtain

$$\min_i \frac{\lambda_i^{(r)}}{\lambda + \lambda_i^{(r)}} < \min_i \frac{\lambda_i^{(r)}}{\lambda_i + \lambda_i^{(r)}}. \tag{4.3.24}$$

It allows us to calculate the difference between the upper bounds of ρ appeared in equation (4.3.22) and (4.3.23), which are sufficient and necessary conditions respectively, of the system stability.

This difference is given by

$$\Delta = \min_i \frac{\lambda_i^{(r)}}{\lambda_i + \lambda_i^{(r)}} - \min_i \frac{\lambda_i^{(r)}}{\lambda + \lambda_i^{(r)}} > 0. \tag{4.3.25}$$

We study this difference numerically. It is worth mentioning that $\Delta > 0$ in any case, and it shows that our approach can not provide a necessary and sufficient condition for stability. Moreover, it is easy to see that in the particular case that $\lambda_i^{(r)} = \mu$ and $\lambda_i = \lambda/M$,

$$\Delta = \frac{\mu}{\lambda/M + \mu} - \frac{\mu}{\lambda + \mu}$$

and $\Delta = 0$ if and only if $M = 1$.

The behaviour of Δ corresponding to variation of different class arrival probabilities, both in symmetric (equal probabilities) as well as in asymmetric (unequal probabilities) cases, are discussed numerically in section (4.6).

4.4 Multi-orbit model with L classes of outgoing calls

Now, we turn our attention onto a model as similar to the one discussed in section (4.3) with the additional assumption that the new model consists of multiple classes of outgoing calls. The analysis presented in the previous section can be extended to study the model considering in this section. Here, we consider a single server system with M classes of incoming customers, obeying Poisson arrival rule with rates λ_k, $k = 1, \ldots, M$ and L classes of internal (outgoing) customers. We assume that when the server becomes empty, an internal class-j call makes attempts to capture the server after exponential amount of time with rate $\lambda_j^{(o)}$, $j = 1, \ldots, L$. The class-j call have iid service times $\{Z_n^{(j)}, n \geq 1\}$ with general distribution and service rate $\mu_j^{(o)} = 1/E[Z^{(j)}]$. Thus, there is a competition between external and internal calls to capture the server. Also, note that the basic process $\{X(t)\}$ remains the same as in the previous case since $Q(t)$ defines only status of the server and does not describe which type of customer being served.

Let $\hat{G}_j(t)$ denotes the renewal process generated by attempts of the class-j outgoing customers to get the server (successful or not) in interval $[0, t)$. That is, outgoing customers make attempts regardless of whether the server is free or busy, however such an attempt may be successful only when the server is free. Denote

by $\{v_i^{(j)},\, i \geq 1\}$ the renewal points of the process $\hat{G}_j$ and let $\hat{Q}_i^{(j)} = Q(v_i^{(j)} - 0)$ be the state of server when the ith customer of the j-outgoing call class (in the renewal process $\hat{G}_j$) appears, $j = 1, \ldots, L$. Then the actual (successful) number of class-j customers, $G_j(t)$, appearing in interval $[0,\, t)$, is

$$G_j(t) = \sum_{i=1}^{\hat{G}_j(t)} \mathbf{1}(\hat{Q}_i^{(j)} = 0),$$

and the summary work $U_j(t)$ generated by class-j outgoing customers in interval $[0, t)$ is

$$U_j(t) = \sum_{i=1}^{\hat{G}_j(t)} \mathbf{1}(\hat{Q}_i^{(j)} = 0) Z_i^{(j)},\ j = 1, \ldots, L. \tag{4.4.1}$$

Then, by keeping all other notations as such in the previous section, the balance equation (4.3.5) assumes the form

$$V(t) \;=\; \sum_{k=1}^{M} \sum_{i=1}^{A_k(t)} S_i^{(k)} + \sum_{j=1}^{L} U_j(t) = S(t) + \sum_{k=1}^{M} W_k(t) + t - I(t), \tag{4.4.2}$$

where the remaining service time $S(t)$ may relate to any customer (incoming or outgoing), which is being served at instant t, if any. Let $\rho_j^{(o)} = \lambda_j^{(o)}/\mu_j^{(o)}$, $j = 1, \ldots, L$. By the strong law of large numbers,

$$\frac{U_j(t)}{t} \to P_0 \rho_j^{(o)} \tag{4.4.3}$$

due to the independence between $\mathbf{1}(\hat{Q}_i^{(j)})$ and $Z_i^{(j)}$. Here $P_0 = E[\mathbf{1}(\hat{Q}^{(j)} = 0)]$ is the stationary probability that a j-outgoing customer meets an idle server (and hence captures it), where the weak limit

$$\mathbf{1}(\hat{Q}_i^{(j)} = 0) \Rightarrow \mathbf{1}(\hat{Q}^{(j)} = 0) \text{ as } i \to \infty$$

exists. By PASTA property, $E[\mathbf{1}(\hat{Q}^{(j)} = 0)]$ is also the limiting fraction of time when the server is free, and this is the reason for using the notation P_0. Then, as in

equations (4.3.5)-(4.3.8), we obtain

$$\lim_{t\to\infty}\frac{V(t)}{t}=\rho+\sum_{j=1}^{L}\rho_j^{(o)}P_0 \tag{4.4.4}$$

and

$$\lim_{t\to\infty}\frac{1}{t}\Big(S(t)+\sum_{k=1}^{M}W_k(t)+t-I(t)\Big)=1-P_0, \tag{4.4.5}$$

which yield the expression for the stationary idle probability of the server as

$$P_0=\frac{1-\sum_{k=1}^{M}\rho_k}{1+\sum_{j=1}^{L}\rho_j^{(o)}}=1-P_b. \tag{4.4.6}$$

The stationary probability that the server is occupied by a class-j outgoing customer is indeed given in (4.4.3). In fact, an important observation here is that the workload $U_j(t)$, can not be accumulated, unlike the work generated by external customers, and equals the busy time which the server devotes to class-j outgoing calls in $[0,\, t)$. Then, from (4.4.3),

$$\lim_{t\to\infty}\frac{U_j(t)}{t}=P_b^{(j)}=\rho_j^{(o)}P_0,\;\; j=1,\ldots,L. \tag{4.4.7}$$

We note that

$$\sum_{j=1}^{L}P_b^{(j)}=P_0\sum_{j=1}^{L}\rho_j^{(o)}$$

is the stationary probability that the server is busy with an outgoing customer. Thus the unconditional busy probability is

$$P_b=\rho+P_0\sum_{j=1}^{L}\rho_j^{(o)}, \tag{4.4.8}$$

which is consistent with (4.4.6).

4.4.1 Stability analysis

In this section, we present a brief stability analysis for some particular cases of our model by taking specific values for M and L, the number of classes of incoming

and outgoing calls respectively. Then we use these results to formulate a conjecture, concerning stability of the system, in a general set up.

To start with, consider a model with $M = 1$ and $L = 1$. In this case, to obtain negative drift (stability) condition, we must take into account the lost capacity of the server (or lost working time) caused by the outgoing calls. To this end, we note that with probability

$$q = \frac{\lambda^{(o)}}{\lambda^{(o)} + \lambda^{(r)}},$$

the server is captured by an outgoing call, which requires a mean service time $1/\mu_1^{(o)}$. Then, with probability q^2, it happens two times and so on. Thus, the total mean time devoted to outgoing calls between departure of a retrial call and the next successive attempt equals $\lambda^{(o)}/(\mu_1^{(o)}\lambda^{(r)})$. Adding the unit time portion of this quantity to the left hand side of the inequality obtained by taking $M = 1$ in (4.3.18), we get

$$\rho + \frac{\lambda}{\lambda^{(r)}}\rho + \frac{\lambda\lambda^{(o)}}{\mu_1^{(o)}\lambda^{(r)}} < 1, \tag{4.4.9}$$

which coincides with the necessary stability condition derived in Aissani and Phung-Duc [12]. Recall that for $M = 1$ and $L = 2$ classes of calls , a necessary stability condition in the form

$$\rho + \frac{\lambda}{\lambda^{(r)}}\rho + \lambda\Big[\frac{\lambda^{(o)}}{\mu_1^{(o)}(\lambda_2^{(o)} + \lambda^{(r)})} + \frac{\lambda_2^{(o)}}{\mu_2^{(o)}(\lambda^{(o)} + \lambda^{(r)})}\Big] < 1 \tag{4.4.10}$$

is given in Morozov and Phung-Duc [61].

Continuing in the same way, we consider the model with L classes of outgoing customers. Denote

$$q_j = \frac{\lambda_j^{(o)}}{\sum_{i=1}^{L}\lambda_i^{(o)} + \lambda^{(r)}},\ j = 1, \ldots, L,$$

the probability that a class-j outgoing customer captures an idle server. Then, as above, we find that the the mean time devoted to class-j calls by the server before a successful attempt of an incoming customer is

$$\frac{1}{\mu_j^{(o)}}\sum_{i=1}^{\infty} q_j^i = \frac{1}{\mu_j^{(o)}}\frac{q_j}{(1-q_j)} = \frac{1}{\mu_j^{(o)}}\frac{\lambda_j^{(o)}}{\sum_{i\neq j}\lambda_i^{(o)} + \lambda^{(r)}}.$$

Hence, the mean total time spent by the server to serve outgoing calls, that are coming

in between two incoming calls is given by

$$W = \sum_{j=1}^{L} \frac{1}{\mu_j^{(o)}} \frac{\lambda_j^{(o)}}{\sum_{i \neq j} \lambda_i^{(o)} + \lambda^{(r)}}. \tag{4.4.11}$$

This leads to the following conjecture.
Conjecture:

$$\rho + \frac{\lambda}{\lambda^{(r)}} \rho + \lambda W < 1 \tag{4.4.12}$$

is a necessary stability condition for the single-class model with multiple classes of outgoing calls, and general service times for both incoming and outgoing calls.

4.5 Single orbit model with multiple classes of incoming and outgoing calls

Two major characteristics of the models that are discussed in section (4.3) and section (4.4) are (i) general service time assumptions for both incoming and outgoing calls and (ii) class specific orbits for incoming calls. However, in this section we present a model by compromising on both these characteristics by assuming exponential service times for both type of calls and a common orbit for all class of incoming calls. Consider a model with M classes of incoming calls. However, it is assumed that the class of a call is defined just before the service starting epoch. Let λ be the arriving call rate (with exponentially distributed interarrival times). If at arrival the server is free, the arriving call defines its class as i w.p. $p_i,\ i = 1, \ldots, M$ and occupies the server for exponentially distributed service time with rate μ_i. Otherwise, the call joins the *common* orbit and *does not define its class*. When the orbit is not empty, the (oldest) call makes attempts to enter the server with exponentially distributed retrial times (of rate $\lambda^{(r)}$) and, upon successful attempt, right before service starts, defines its class as i w.p. $p_i^{(r)}, i = 1, \ldots, M$. The service rate is μ_i as defined earlier, for class i call. Finally, when being idle, the server waits for exponentially distributed time (with rate $\lambda^{(o)}$) and if no customer arrives/retries before this timer expires, the server starts some routine (or makes an outgoing call) of class j w.p. $p_j^{(o)}$, which makes it busy for exponentially distributed time with rate $\mu_j^{(o)}, j = 1, \ldots, L$.

We consider the process $\{X(t) = (N(t), Q(t)); t \geqslant 0\}$, where $N(t)$ is the orbit size

at time t and $Q(t) = 0$ if the server is idle, $Q(t) = i \in \{1, \ldots, M\}$ if the server is busy serving incoming call i and $Q(t) = M + j,\ j = 1, \ldots, L$, if the server is performing outgoing call of class j. For convenience, we use the same numbering $0, 1, \ldots, M + L$ in the rows and columns of matrices, and see that the process $X(t)$ is indeed a QBD. We define the square matrices (of order $M+L+1$) for the block-tridiagonal generator, given in (4.2.2) , as follows:

$$A_0 = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & \lambda & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & \lambda^{(r)}p_1^{(r)} & \ldots & \lambda^{(r)}p_M^{(r)} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix},$$

(4.5.1)

$$A_1 = \begin{pmatrix} -\lambda - \lambda^{(r)} - \lambda^{(o)} & \lambda p_1 & \lambda p_2 & \ldots & \lambda p_M & \lambda^{(o)}p_1^{(o)} & \ldots & \lambda^{(o)}p_L^{(o)} \\ \mu_1 & -\lambda - \mu_1 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ \mu_2 & 0 & -\lambda - \mu_2 & \ldots & \ldots & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \mu_M & 0 & 0 & \ldots & -\lambda - \mu_M & \ldots & \ldots & 0 \\ \mu_1^{(o)} & 0 & 0 & \ldots & 0 & -\lambda - \mu_1^{(o)} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_L^{(o)} & 0 & 0 & \ldots & 0 & 0 & \ldots & -\lambda - \mu_L^{(o)} \end{pmatrix},$$

(4.5.2)

$$A^{0,0} = A_1 + \begin{pmatrix} \lambda^{(r)} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}. \tag{4.5.3}$$

For convenience, we use the following notations:

$$\rho_i = \lambda \frac{p_i}{\mu_i},\ i = 1, \ldots, M;$$

$$\rho = \sum_{i=1}^{M} \rho_i,$$

$$\rho_i^{(r)} = \lambda^{(r)} \frac{p_i^{(r)}}{\mu_i};\ i = 1, \ldots, M;$$

$$\rho^{(r)} = \sum_{i=1}^{M} \rho_i^{(r)},$$

$$\rho_j^{(o)} = \lambda^{(o)} \frac{p_j^{(o)}}{\mu_j^{(o)}},\ j = 1, \ldots, L;$$

$$\rho^{(o)} = \sum_{j=1}^{L} \rho_j^{(o)}.$$

Special structure of the matrix $A = A_0 + A_1 + A_2$ allows to obtain the vector α, appeared in (4.2.4), as

$$\alpha = \alpha_0 \left(1, \rho_1 + \rho_1^{(r)}, \ldots, \rho_M + \rho_M^{(r)}, \rho_1^{(o)}, \ldots, \rho_L^{(o)}\right),$$

where the second equation in (4.2.5) provides

$$\alpha_0 = \left[1 + \rho + \rho^{(r)} + \rho^{(o)}\right]^{-1}.$$

Thus, it is easy to obtain from (4.2.4) the following stability criterion

$$\lambda(\rho + \rho^{(r)} + \rho^{(o)}) < \lambda^{(r)}. \tag{4.5.4}$$

Interestingly, (4.5.4) provides a second degree polynomial in λ (since $\rho = \lambda \sum_{i=1}^{M} p_i/\mu_i$) which can be shown to have two distinct roots. Let the largest one be denoted by λ^*. Then,

$$\lambda^* = \frac{-\rho^{(r)} - \rho^{(o)} + \sqrt{(\rho^{(r)} + \rho^{(o)})^2 + 4\lambda^{(r)} \sum_{i=1}^{M} p_i/\mu_i}}{2\sum_{i=1}^{M} p_i/\mu_i}.$$

So, for any value of λ, which is smaller than λ^* will satisfy (4.5.4) Thus, the system is stable for $\lambda < \lambda^*$ with the other parameters fixed.

It only remains to note that $A_2 = cr$, where $c = (\lambda^{(r)}, 0, 0)$ is a column vector and $r = (0, p_1^{(r)}, \ldots, p_M^{(r)}, 0, \ldots, 0)$ is a row vector. Then, it follows from Latouche and Ramaswami [50] that $G = \mathbf{e}r$, and the rate matrix R is obtained explicitly as

$$R = -A_0(A_1 + A_0 G)^{-1}.$$

Recall that the vector $\pi = (\pi_0, \pi_1, \dots)$, where $\pi_i = (\pi_{i,0}, \pi_{i,1}, \dots, \pi_{i,M+L})$ is the stationary probability of having i customers in the orbit, with server being idle, busy with incoming or outgoing customer of corresponding class, may be obtained by the following matrix-geometric solution

$$\pi_i = \pi_0 R^i, \quad i \geqslant 1, \tag{4.5.5}$$

where π_0 is obtained from the system

$$\begin{cases} \pi_0(A^{0,0} + RA_2) = \mathbf{0} \\ \pi_0(I - R)^{-1}\mathbf{e} = 1 \end{cases} \tag{4.5.6}$$

Thus, the stationary system state probabilities are obtained explicitly.

4.5.1 Explicit expression for idle probability

Now, we demonstrate a methodologically interesting approach to obtain the steady-state probabilities explicitly. To simplify the scope, we assume $M = 1$, which puts us into the framework of the model considered in Sakurai and Phung-Duc [69].

Obtaining π_0 requires matrix inversion and in general is complicated. So, consider the following equivalent system

$$\begin{cases} \xi_0(I - R)(A^{0,0} + RA_2) = \mathbf{0} \\ \xi_0\mathbf{e} = 1, \end{cases}$$

where $\xi_0 = \pi_0(I-R)^{-1}$. Then, π_0 might be obtained by matrix multiplication as $\pi_0 = \xi_0(I-R)$, and $\xi_0 = (P_0, P_b^1, P_{b(1)}^{(o)}, \dots, P_{b(L)}^{(o)})$, where P_b^1 denotes the probability of server being busy with primary customer and $P_{b(j)}^{(o)}$ denotes the probability of server being busy with jth type of outgoing call, may be obtained using regenerative approach.

For example, for the model with $M = 1$ and $L \geq 1$, we get

$$\xi_0 = \left(\frac{1-\rho}{1+\rho^{(o)}}, \rho, \rho_1^{(o)} \frac{1-\rho}{1+\rho^{(o)}}, \dots, \rho_L^{(o)} \frac{1-\rho}{1+\rho^{(o)}} \right)$$

by using regenerative approach discussed earlier. Then, from the relation $\pi_0 = \xi_0(I - R)$, we get

$$P_{0,0} = \pi_{0,0} = \frac{1 - \rho - \frac{\lambda}{\lambda^{(r)}} \left(\rho + \rho^{(o)}\right)}{1 + \rho^{(o)}}.$$

Thus, in this section, we used both matrix analytic method and regenerative approach to get the steady state distribution and some important performance measures for the particular case that $M = 1, L \geq 1$.

4.6 Simulation results

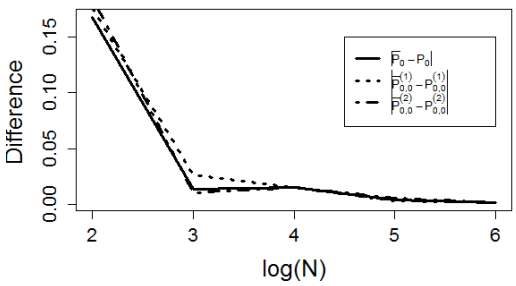

Figure 4.1: Estimated vs. exact values of P_0 and $P_{0,0}^{(k)}$

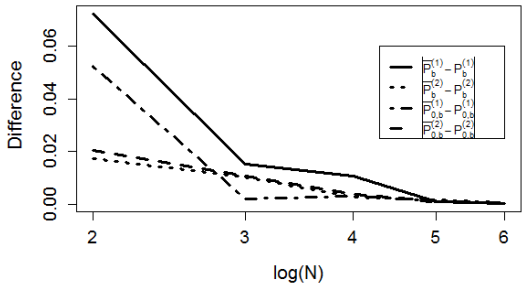

Figure 4.2: Estimated vs. exact values of $P_b^{(k)}$ and $P_{0,b}^{(k)}$

In order to validate our theoretical results, we perform some simulation in R. In the first experiment, we establish the convergence of estimates of those performance measures for which exact values are available via (4.3.1)–(4.3.4). The single trace of simulation run is obtained for each value of number of events, $N = 10^i$, for $i = 2, \ldots, 6$. From each such trace, estimates of the probabilities P_0, P_b, $P_b^{(k)}$, $P_{0,0}^{(k)}$, and $P_{0,b}^{(k)}$ are obtained. Absolute difference between these simulated estimates (denoted by 'bar') and exact values (obtained from (4.3.1)–(4.3.4)), for a system with $M = 2$ classes of incoming customers, is computed (see Fig. 4.1 and 4.2). Values of p_i , $\lambda_i^{(r)}$ and μ_i, for $i = 1, 2$, are taken arbitrarily. Service times are assumed to be exponential. Note that, multiple experiments are performed for various parameter combinations, taken from the system stability region, yielding qualitatively the same results, which are exhibited in Fig. 4.1 and 4.2. Fig. 4.1 indicates variation of absolute difference between exact and simulated values of P_0 and $P_{0,0}^{(k)}$ against the variation of $\log N$. It is seen that the differences fastly converge to zero for large values of $\log N$. Fig. 4.2 displays the differences for the probabilities $P_b^{(k)}$ and $P_{0,b}^{(k)}$. In all cases, simulation illustrates the validity of theoretical results and, moreover, allows us to select in the subsequent experiments a valid number of customers required to obtain performance measures with high accuracy.

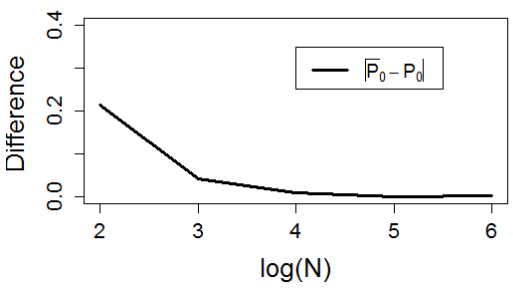

Figure 4.3: Estimated vs. exact values of P_0

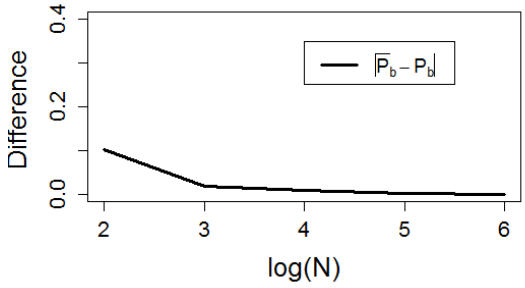

Figure 4.4: Estimated vs. exact values of P_b

In the second experiment, we establish the convergence of estimates of those performance measures for which exact values are available via (4.4.6) and (4.4.8). Single trace of simulation run is obtained for each value of number of events, $N = 10^i$, for $i = 2, \ldots, 6$. From each such trace, estimates of the probabilities P_0 and P_b are obtained. Absolute difference between these simulated estimates (denoted by 'bar') and exact values (obtained from (4.4.6) and (4.4.8)), for a system with $M = 1, L = 2$, is computed (see Fig. 4.3 and 4.4).

Now, in order to study the behaviour of Δ, given by (4.3.25), corresponding to variation of probabilities of different classes of incoming calls, we consider a model with 2 classes of incoming calls and no outgoing calls (the model discussed in section (4.3).

Referring to (4.3.25), we consider $\Delta = \Delta(p_1)$, as a function of probability p_1.

It is easy to deduce from (4.3.25) that $\Delta(p_1)$ attains its maximum at

$$p_1 = \frac{\lambda_1^{(r)}}{\lambda_1^{(r)} + \lambda_2^{(r)}}.$$

For better understanding, we depict the dependence of $\Delta(p_1)$ on $p_1 \in (0,\ 1)$ in Fig. 4.5 in the following cases: (i) $\lambda_1^{(r)} = \lambda_2^{(r)}$; (ii) $\lambda_1^{(r)} > \lambda_2^{(r)}$; and (iii) $\lambda_1^{(r)} < \lambda_2^{(r)}$. In the first case, from Fig. 4.5, it may be observed that $\Delta = 0$ if $p_1 \in \{0,\ 1\}$, which is consistent with the remark that $\Delta = 0$ if and only if $M = 1$ in the system with symmetrical orbits ($p_1 = p_2$). However, in asymmetric case ($p_1 \neq p_2$), $\Delta = 0$ only at one boundary point.

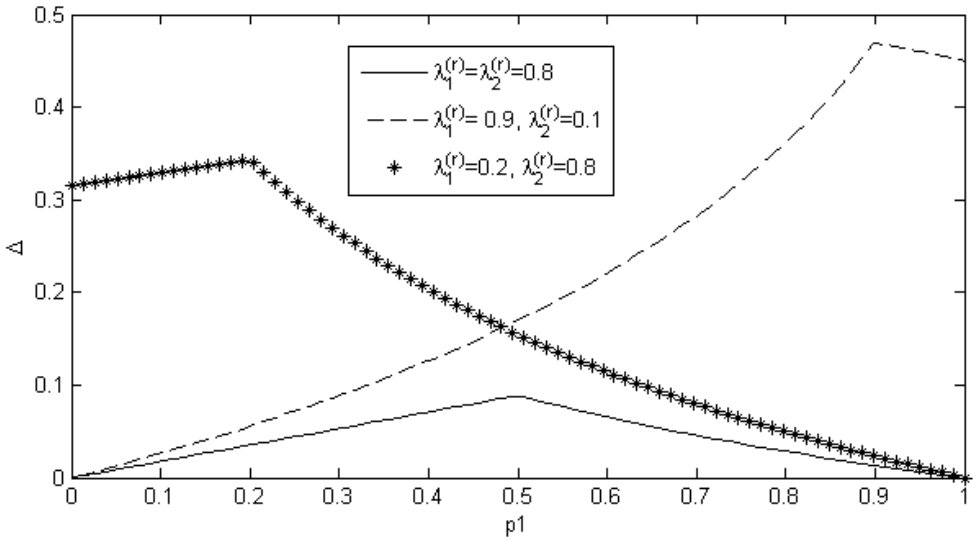

Figure 4.5: Dependence of Δ on the probability p_1, for symmetric and asymmetric orbit cases, for $M = 2$

Chapter 5

A Multi-class Orbit Queue with Constant Retrial Rates and Balking

5.1 Introduction

In the previous chapter, we have analysed various models involving multiple classes of incoming calls and with/without outgoing calls. From the server's perspective, it is immaterial whether the server itself initiates an outgoing call during idle time, or an arriving customer of some specific type finding an idle server starts getting service. In the latter case, a customer of this (outgoing) type finding the server busy is assumed to leave the system forever, or balks. At that, we may think of outgoing calls as being initiated from an independent Poisson stream of (outgoing class) customers meeting the server idle. We assume exponential times from beginning of idle period to outgoing call initiation.

At the same time, *incoming* customers finding the server busy join the orbit queue. A generalization of these two extremes (balk or join) is the probabilistic balking of arrivals facing a busy server. This motivates us to study the single-server multi-class retrial orbit queues with balking as a generalization of the two-way orbit queues studied in chapter 4. Regenerative analysis allows us to establish stability conditions and compute important performance measures for multi-class orbit queues with balking and general service time distribution.

Many variants of retrial models dealing with different types of customer behaviour like *impatience* and *balking* were studied by researchers. Falin and Artalejo [36] studied a multi-server model with classical retrial discipline and obtained approximations for performance measures of the system. Dudin and Klimenok [32] analysed a multi-server retrial model with balking, governed by batch Markovian

arrival process and phase-type service times, where the orbital customers were considered as impatient and non-persistent. With the same assumptions as above for arrival process and service times, Dudin and Klimenok [31] studied a multi-server model having Markov modulated retrials. We also mention the work by Artalejo et al. [5], where a constant retrial rate multi-server system with the *recovery factor* (probability to enter idle server) was analyzed by the matrix-analytic method.

The structure of the chapter is as follows. In section (5.2), we introduce the multi-class multi-orbit single-server model with balking and obtain basic stationary performance measures by regenerative approach. In section (5.3), we prove necessary and sufficient stability conditions of the model, where coupling method is used extensively. In section (5.4), we study some particular cases of our model and derive expression for mean stationary orbit size. Finally, in section (5.5), we perform extensive simulation to validate the performance, obtain the convergence rates and demonstrate some other basic performance measures of interest, which are having no explicit analytic expressions, to gain more insight into the model.

5.2 Description of the model and performance analysis

5.2.1 Description of the model

We consider a single-server retrial queue with M classes of *balking customers* who may or may not join the orbit queue if the server is busy upon their arrival. Class-k customers arrive at epochs $\{t_n^{(k)}, n \geqslant 1\}$ of a Poisson process with rate $\lambda_k \in (0, \infty)$ and have iid general service times $\{S_n^{(k)}, n \geq 1\}$ with rate μ_k and mean $E[S^{(k)}] = 1/\mu_k \in (0, \infty)$. Let

$$\lambda = \sum_k \lambda_k,\ p_k = \frac{\lambda_k}{\lambda},\ \rho_k = \frac{\lambda_k}{\mu_k} \text{ and } \rho = \sum_{k=1}^{M} \rho_k.$$

The class-k customer finding the server busy upon its arrival joins the k-orbit with probability b_k, or leaves the system forever. Thus, we consider a decision of the nth class-k customer to be a Bernoulli variable $\beta_n^{(k)}$ such that $E[\beta_n^{(k)}] = b_k$, and

$\{\beta_n^{(k)},\, n \geq 1\}$ are iid for each k. Time between retrial attempts from k-orbit follow exponential distribution with rate $\lambda_k^{(r)}$, independent of the orbit size.

It is worth mentioning that the present model incorporates multi-class orbit queues if $b_k = 1$, and two-way communication queueing system if $b_k \in \{0, 1\}$, $k = 1, \ldots, M$. Indeed, if $b_k = 0$ for some k, then an arriving class-k customer finding the server busy is lost, leaving the system state unaffected. In such a case, class-k customers may be considered as the outgoing calls initiated by an idle server.

Now, we describe the stochastic processes used hereafter to obtain the steady state results. Let $A_k(t)$ be the number of class-k arrivals in interval $[0,\, t)$ and $V_k(t) = \sum_{n=1}^{A_k(t)} S_n^{(k)}$ be the summary work of class-k customers *registered* in the system in interval $[0,\, t)$ (with obvious convention $V_k(t) = 0$ if $A_k(t) = 0$). Then, $V(t) = \sum_k V_k(t)$ represents the total workload of arrivals of all classes in $[0,\, t)$. In interval $[0, t)$, let $B(t)$ denote the busy time of the server, and $I(t)$ its idle time so that $B(t) + I(t) = t$. Further we split $B(t) = \sum_{k=1}^{M} B_k(t)$, where $B_k(t)$ is the time the server devotes to class-k customers in interval $[0,\, t)$. Also, let $N_k(t)$ denotes the number of orbital customers and $W_k(t)$ the workload (remaining work) in k-orbit queue, at instant t^-. Let $S(t)$ be the remaining service time at instant t^-, i.e. the time to departure of a customer being served. (By definition, $S(t) = 0$ if the server is free.) Similarly, we represent $S(t) = \sum_{k=1}^{M} S_k(t)$, where $S_k(t)$ is the remaining service time, provided class-k customer is at the server at t^-, and only one summand may be positive in this sum.

To define regeneration epochs, we consider the basic *non-Markovian* summary queue size process

$$X(t) = \sum_{k=1}^{M} N_k(t) + Q(t),\, t \geq 0,$$

where $Q(t) \in \{0,\, 1\}$ is the number of calls with the server at instant t^-. Let $\{t_n,\, n \geq 1\}$ be the instants of the superposed input (Poisson) process with rate λ. Let $X(t_n) = X_n$, $n \geq 1$, $T_0 = 0$ and $\{T_n : n \geq 1\}$ be defined recursively by

$$T_{n+1} = \inf\Big(t_k > T_n : X_k = 0\Big),\, n \geq 0.$$

We assume zero initial state, that is, the first customer arrives at instant $t = 0$ in the system, which is idle just before arrival, and with probability p_k, the arriving customer is of class k, $k = 1, \ldots, M$. It is easy to see that $\{T_n\}$ is a sequence of classical regeneration points of the basic process X, with the iid *regeneration periods*

$T_{n+1} - T_n$ (and with generic period T). If the mean period $E[T] < \infty$, then the process $\{X(t)\}$ (and the model) is *positive recurrent*, and there exists the weak limit $X(t) \Rightarrow X$, which is the stationary number of customers in the system (see, for instance, Asmussen [11]).

Now, let us introduce the following steady-state probabilities: P_0, the server is idle; $P_b = 1 - P_0$, the server is busy; $P_{0,0}^{(k)}$, the server is idle and k-orbit is empty; $P_{0,b}^{(k)}$, the server is idle and k-orbit is non-empty. Additionally, let us denote $P_b^{(k)}$ as the stationary probability that the server is busy by class-k customer, $k = 1, \ldots, M$. These performance measures have been obtained in Phung-duc et al. [65], Sakurai and Phung-duc [69] by the Kolmogorov's equations approach, for a restricted (Markovian two-way communication) model. Below, we obtain these probabilities for our (more general) model by the regenerative method.

5.2.2 Analysis of steady-state regime

For ease of notation, let us denote

$$C = 1 + \rho - \sum_{k=1}^{M} \rho_k b_k. \tag{5.2.1}$$

Theorem 5.2.1. *If M-class orbit queues are positive recurrent, then*

$$P_0 = 1 - \frac{\rho}{C} = 1 - P_b, \tag{5.2.2}$$

and, for each $k = 1, \ldots, M$, the following relations hold true:

$$P_b^{(k)} = \rho_k \left(1 - \frac{\rho}{C}(1 - b_k)\right); \tag{5.2.3}$$

$$P_{0,\,b}^{(k)} = \frac{\lambda_k b_k}{\lambda_k^{(r)}} \frac{\rho}{C}; \tag{5.2.4}$$

$$P_{0,\,0}^{(k)} = 1 - \left(1 + \frac{\lambda_k b_k}{\lambda_k^{(r)}}\right) \frac{\rho}{C}. \tag{5.2.5}$$

Proof. By denoting the embedded process $Q(t_n^{(k)})$ by $Q_n^{(k)}$, and considering the summary work $V_k(t)$ that class-k customers bring in the system in interval $[0,\, t)$,

we get

$$V_k(t) = \sum_{n=1}^{A_k(t)} S_n^{(k)} \Big(\mathbf{1}(Q_n^{(k)} = 0) + \mathbf{1}(Q_n^{(k)} = 1)\beta_n^{(k)} \Big), \tag{5.2.6}$$

where $\mathbf{1}$ denotes the indicator function. Then the equations

$$V_k(t) = S_k(t) + B_k(t) + W_k(t), \;\; k = 1, \ldots, M \tag{5.2.7}$$

hold good. By the positive recurrence, $S_k(t) = W_k(t) = o(t)$ with probability 1 (w.p.1) as $t \to \infty$, see Smith [67]. Moreover, w.p.1 the limits

$$P_b^{(k)} = \lim_{t\to\infty} B_k(t)/t, \;\; k = 1, \ldots, M, \tag{5.2.8}$$

exist. Note that, since the input is Poisson and the process X is positive recurrent, there exists the weak limit $Q_n^{(k)} \Rightarrow Q^{(k)}$, where $Q^{(k)}$ is the stationary number of customers in service observed by class-k customers. Moreover, since $A_k(t) \to \infty$ as $t \to \infty$, it follows by regenerative arguments that, w.p.1

$$\begin{aligned} &\lim_{t\to\infty} \frac{1}{A_k(t)} \sum_{n=1}^{A_k(t)} S_n^{(k)} \Big(\mathbf{1}(Q_n^{(k)} = 0) + \mathbf{1}(Q_n^{(k)} = 1)\beta_n^{(k)} \Big) \\ &= E\Big[S^{(k)} \Big(\mathbf{1}(Q^{(k)} = 0) + \mathbf{1}(Q^{(k)} = 1)\beta^{(k)} \Big) \Big] = E[S^{(k)}](P_0 + P_b b_k), \end{aligned} \tag{5.2.9}$$

where we apply independence between $S^{(k)}$, $\beta^{(k)}$ and indicators and use PASTA property. The probability $P_b\,(P_0)$ is also the *limiting fraction of the time when the server is busy (idle).* It is worth mentioning that these limits do not depend on customer class k. Since $A_k(t)/t \to \lambda_k$ as $t \to \infty$, it is easy to conclude that, as $t \to \infty$,

$$\frac{V_k(t)}{t} = \frac{A_k(t)}{t} \sum_{n=1}^{A_k(t)} \frac{S_n^{(k)} \Big(\mathbf{1}(Q_n^{(k)} = 0) + \mathbf{1}(Q_n^{(k)} = 1)\beta_n^{(k)} \Big)}{A_k(t)} \to \rho_k(P_0 + P_b b_k). \tag{5.2.10}$$

Finally, (5.2.8) and (5.2.10) give

$$P_b^{(k)} = \rho_k(P_0 + P_b b_k), \;\; k = 1, \ldots, M. \tag{5.2.11}$$

Now, it remains to define P_0. To this end, we sum up (5.2.7) for $k = 1, \ldots, M$, and

obtain the balance relation

$$V(t) = \sum_{k=1}^{M} V_k(t) = S(t) + \sum_{k=1}^{M} W_k(t) + t - I(t). \tag{5.2.12}$$

Since $S(t) = \sum_{k=1}^{M} S_k(t) = o(t)$ as $t \to \infty$, it follows from (5.2.8)–(5.2.12) that

$$\lim_{t\to\infty} \frac{V(t)}{t} = \sum_{k=1}^{M} \rho_k(P_0 + P_b b_k). \tag{5.2.13}$$

On the other hand, the r.h.s. of (5.2.12) gives in the limit

$$\lim_{t\to\infty} \frac{1}{t}\Big(S(t) + \sum_{k=1}^{M} W_k(t) + t - I(t)\Big) = 1 - P_0. \tag{5.2.14}$$

Note that, by the positive recurrence, the limit

$$\lim_{t\to\infty} \frac{I(t)}{t} = P_0 = P(Q = 0)$$

exists w.p.1. Here the weak limit $Q(t) \Rightarrow Q$ exists and represents the stationary number of customers in the server, see Smith [67]. We stress that, by PASTA, $P_0 = P(Q = 0) = P(Q^{(k)} = 0)$ (and $P_b = P(Q = 1) = P(Q^{(k)} = 1)$ as well) for each k, see (5.2.9). Now (5.2.2) follows from (5.2.12)-(5.2.14), and (5.2.3) follows from (5.2.2) and (5.2.11).

It is more challenging to find $P_{0,\,b}^{(k)}$. We use $D_k(t)$ to denote the number of retrial class-k customers which depart k-orbit in $[0,\, t)$. Also, let $\hat{D}_k(t)$ be the number of renewals in interval $[0,\, t)$ generated by the iid exponential variables with parameter $\lambda_k^{(r)}$ (with $\hat{D}_k(0) = 0$). That is, $\hat{D}_k$ is the Poisson process with rate $\lambda_k^{(r)}$. Denote $\{z_n^{(k)},\, n \geq 1\}$ the instants (renewal epochs) of the process $\hat{D}_k$. To connect the real number of departures D_k with the Poisson process $\hat{D}_k$, we use the following coupling. We sample the process $\hat{D}_k$ until the 1st class-k customer joins the orbit. At this instant, we resample the remaining renewal (exponential) time in the process $\hat{D}_k$ and then treat the subsequent (inter) renewal intervals in the process $\hat{D}_k$ as the process of the attempts from orbit k, until a class-k orbital customer leaves the orbit empty, if any. From this instant, we continue to sample the process $\hat{D}_k$ until the next customer joins orbit k. At this instant, we resample the remaining renewal time in the process $\hat{D}_k$ and, as above, interpret the subsequent renewals in the process $\hat{D}_k$ as the attempts

from orbit k, until the orbit becomes empty and so on. Thus, by construction, the instants of attempts of the (top) orbital customers constitute a subsequence of the renewal instants of the process $\hat{D}_k$ with resampling. Evidently, at each instant of this subsequence, orbit k is *not idle*. The modified renewal process (with resampling) is stochastically equivalent to originally defined process $\hat{D}_k$. In what follows, we keep the same notation $\hat{D}_k$ for the modified process, and $\{z_n^{(k)},\, n \geq 1\}$ for the instants of this process. For each k, let

$$Q(z_n^{(k)}) = Q_n^{(k)} \text{ and } \; N_k(z_n^{(k)}) = N_n^{(k)},\; n \geq 1.$$

(Recall that $Q(t)$ and $N(t)$ are left-continuous). It then follows from the above that $\mathbf{1}(Q_n^{(k)} = 0,\, N_n^{(k)} > 0) = 1$ means that the nth instant of the renewal process $\hat{D}_k$ is the successful attempt of a class-k customer to enter the server. Then, the number of customers which depart k-orbit in interval $[0,\, t)$ is defined as

$$D_k(t) = \sum_{n=1}^{\hat{D}_k(t)} \mathbf{1}(Q_n^{(k)} = 0,\, N_n^{(k)} > 0),\; k = 1, \ldots, M. \tag{5.2.15}$$

Since at the instant of each attempt, the orbit is not idle, a key observation is that D_k can be treated as the number of epochs, among Poisson epochs $\hat{D}_k$ (with rate $\lambda_k^{(r)}$), which "meet" *empty server and, simultaneously, non-empty orbit k*. (Note that the number of real attempts, in $[0,\, t)$, from orbit k is defined then as $\sum_{n=1}^{\hat{D}_k(t)} \mathbf{1}(N_n^{(k)} > 0)$.)

By the positive recurrence, there exists (w.p.1) the limit

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{\hat{D}_k(t)} \mathbf{1}(Q_n^{(k)} = 0,\, N_n^{(k)} > 0)}{\hat{D}_k(t)} = P_{0,b}^{(k)}. \tag{5.2.16}$$

We again note that, by PASTA property, $P_{0,b}^{(k)}$ is also the limiting fraction of the time when the server is free and the kth orbit is non-empty. Moreover, by the renewal theory, w.p.1

$$\lim_{t\to\infty} \frac{\hat{D}_k(t)}{t} = \lambda_k^{(r)}, \tag{5.2.17}$$

and thus (5.2.15)–(5.2.17) provide

$$\lim_{t\to\infty} \frac{D_k(t)}{t} = \lambda_k^{(r)} P_{0,b}^{(k)},\; k = 1, \ldots, M. \tag{5.2.18}$$

Now we introduce $J_k(t)$, the number of class-k customers that join the orbit in interval $[0,\, t)$. Let us denote $Q(t_n)$ by Q_n, the state of the server *just before* the nth arrival in the superposed input process. Then, we have

$$J_k(t) = \sum_{n=1}^{A_k(t)} \mathbf{1}(Q_n = 1)\beta_n^{(k)},\ k = 1, \ldots, M.$$

Again, by PASTA property, $Q_n \Rightarrow Q$, and moreover

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{A_k(t)} \mathbf{1}(Q_n = 1)\beta_n^{(k)}}{A_k(t)} = E[\mathbf{1}(Q = 1)\beta^{(k)}] = P_b\, b_k,$$

where we apply independence between $\mathbf{1}(Q = 1)$ and $\beta^{(k)}$. Now, it follows that

$$\lim_{t\to\infty} \frac{J_k(t)}{t} = \lambda_k b_k P_b, \quad k = 1, \ldots, M. \tag{5.2.19}$$

Obviously, for each k, we have the balance equation

$$J_k(t) = N_k(t) + D_k(t). \tag{5.2.20}$$

Since, by positive recurrence, $N_k(t) = o(t)$ as $t \to \infty$, (5.2.4) follows from (5.2.18)–(5.2.20) in the limit. Also, we have

$$P_0 = P_{0,b}^{(k)} + P_{0,0}^{(k)}\ k = 1, \ldots, M, \tag{5.2.21}$$

so that (5.2.5) follows from (5.2.2) and (5.2.4). □

Remark 5.2.1. *It is to be noted that when $b_k = 1$ for $k = 1, 2, ..., M$, then* (5.2.1) *gives $C = 1$. Hence, the measures obtained from equations* (5.2.2)–(5.2.5) *coincide with the results given by* (4.3.1) –(4.3.4).

5.3 Stability analysis

In this section, we establish necessary and sufficient stability conditions of the basic process X. Based on the explicit expression for the stationary probability $P_{0,\,b}^{(k)}$ derived above, we arrive at the following necessary stability conditions.

Theorem 5.3.1 (Necessary stability condition)**.** *If the system comprising M-class orbit queues with balking customers is stable, then*

$$\lambda_k b_k \frac{\rho}{C} < \lambda_k^{(r)}(1 - \frac{\rho}{C}),\ k = 1,\ \ldots, M. \tag{5.3.1}$$

Proof. The proof is straightforward. We write expression (5.2.4) as

$$\lambda_k^{(r)} P_{0,\,b}^{(k)} = \lambda_k b_k \frac{\rho}{C},\ \ k = 1, \ldots, M. \tag{5.3.2}$$

Since the input is Poisson, intervals between attempts are exponential and each orbit k (with $b_k > 0$) is visited by a blocked class-k customer with a positive probability within regeneration cycle. It is easy to construct an event, with a positive probability, when the server is idle and orbit k is nonempty. For example, consider a regeneration cycle containing exactly one class-j customer, $j \neq k$. Then, with the probability which is lower bounded by $P(S^{(j)} + \delta < \tau) \geq \varepsilon$ for some positive δ, ε, the idle time of *each orbit* within regeneration cycle is lower bounded by δ. (Here τ is the generic exponential interarrival time.) Hence, it follows that $P_{0,0}^{(k)} > 0$ and

$$1 - \frac{\rho}{C} = P_0 = P_{0,b}^{(k)} + P_{0,0}^{(k)} > P_{0,b}^{(k)}.$$

Now, (5.3.1) follows from (5.3.2). □

Note that $\lambda_k b_k$ is the *effective flow rate* to orbit k while $\rho/C = P_b$ is the traffic intensity.

Remark 5.3.1. *It may be noted that the necessary stability conditions, shown in* (5.3.1)*, of the M-class orbit queues with $b_k = 1$ (equivalently, $C = 1$) assume the form*

$$\lambda_k \rho = \lambda_k P_b < \lambda_k^{(r)}(1 - P_b),\ \ k = 1, \ldots, M, \tag{5.3.3}$$

which coincide with the stability condition given by (4.3.18)*.*

Remark 5.3.2. *It is worth mentioning that if all $b_k = 0$, then we obtain a loss system in which $C = 1 + \rho$ and hence it follows from* (5.2.2) *and* (5.2.3) *that*

$$P_b = \frac{\rho}{1+\rho} \ \textit{and} \quad P_b^{(k)} = \frac{\rho_k}{1+\rho},\quad k = 1, \ldots, M.$$

Now, we establish sufficient stability condition for the model by assuming that all customers persist, that is, $b_k = 1,\ k = 1, \ldots, M$. We emphasize that this condition, as we will see, differs from the necessary condition found above. The form of this condition is motivated by a negative drift of the workload process which in turn implies positive recurrence of the basic process X.

In the proof of Theorem (5.3.2) presented below, we make a comparison of the original retrial system with the classic multi-class system having infinite buffer for the awaiting customers. The basic difference between these two systems is that, in the original system, unlike classic system, there exists an idle time of the server after each service completion, and it makes service discipline *non work-conserving*. Note that it holds both for single class and multi-class systems. (For more about *work-conserving* discipline, see Asmussen [11].)

Recall that we consider zero initial state, that is, $X(0) = X_1 = 0$.

Theorem 5.3.2. *If* $b_k = 1,\ k = 1, \ldots, M$, *and the condition*

$$\rho + \max_{1 \le k \le M} \frac{\lambda}{\lambda + \lambda_k^{(r)}} < 1 \tag{5.3.4}$$

holds, then the basic process X *is positive recurrent.*

Proof. For each k, let $\mathcal{S}^{(k)} = \{S_n^{(k)},\ n \ge 1\}$, the sequence of the iid service times of class-k customers. Also, define indicator $I_n^{(k)} = 1$ if the nth arrival is class-k one, and $I_n^{(k)} = 0$, otherwise, implying $E[I_k] = p_k,\ k = 1, \ldots, M$. Then, the service time of the nth arrival, written in a more detailed form, is

$$S_n = \sum_{k=1}^{M} I_n^{(k)} S_n^{(k)},\ n \ge 1. \tag{5.3.5}$$

Note that, to compose S_n, we select from the (independent) sequences $\mathcal{S}^{(k)}$ only such elements $S_n^{(k)}$ which correspond to $I_n^{(k)} = 1$. Also, note that $I_n^{(k)}$ and $S_n^{(k)}$ are independent. In the retrial model, an idle time of the server occurs after each departure. At each such instant, the remaining time to next arrival epoch is exponential and distributed as interarrival time τ. We use $\{\xi_n^{(k)},\ n \ge 1\}$ to denote the iid exponential variables with parameter $\lambda_k^{(r)}$ between the attempts from the non-empty orbit k (with generic variable $\xi^{(k)}$). Then, provided orbit k is non-empty, the idle time of the server after departure is upper bounded by the variable $\min(\tau,\ \xi^{(k)})$ with the mean $\frac{1}{\lambda + \lambda_k^{(r)}}$. We denote ζ_n the idle time of server *after the nth departure.*

Note that, if at least one orbit is non-empty after the nth departure, then

$$E[\zeta_n] \leq \max_{1\leq k\leq M} \frac{1}{\lambda + \lambda_k^{(r)}}. \tag{5.3.6}$$

Now, we construct a *dominating buffered system*, under FIFO discipline, having same input and service times as in the original retrial model with the only exception that here each customer occupies the server, besides its service time, for an extra exponential time ξ^o with parameter $\min_{1\leq k\leq M}(\lambda + \lambda_k^{(r)})$ and mean

$$E[\xi^o] = \frac{1}{\min_{1\leq k\leq M}(\lambda + \lambda_k^{(r)})} = \max_{1\leq k\leq M} \frac{1}{\lambda + \lambda_k^{(r)}}. \tag{5.3.7}$$

In other words, ξ^o is the time between attempts from the nonempty orbit with the *slowest* retrial rate.

Now we describe this construction in more detail. Recall that we apply the following coupling: the identical interarrival times and service times in both systems, original and buffered. Moreover, we introduce the following iid vectors with the independent components

$$\xi(n) = (\xi_n^{(1)}, \ldots, \xi_n^{(M)}),\ n \geq 1,$$

describing the times between attempts from *non-empty* orbits. In particular, $\xi_n^{(l)}$ represents the time between $(n-1)$th and nth retrial attempt from l-orbit. By stochastic ordering and coupling, we can sample vector $\xi(n)$ in such a way that the component ξ_n^o (distributed as ξ^o) representing the nth time between attempts from (non-empty) "slowest" orbit, indeed is the *maximal component* of $\xi(n)$. Recall that, since we apply zero initial state, the 1st departure instant in the original systems is S_1, and that, at this instant, there are the same number of (the same types) customers in both systems as well. Denote the set of the non-empty orbits at the nth departure instant in the original system by $R(n)$, and let $\{\tau_n,\ n \geq 1\}$ be the interarrival times in the merged (Poisson) input. Assume first that the set $R(1)$ *is non empty*. In particular, since by assumption the 1st customer arrives at instant $t_1 = 0$, τ_1 is the instant of the 2nd arrival in both systems. Then we assign, using vector $\xi(1)$, the delay $\xi_1 = \min_{i\in R(1)} \xi_1^{(i)}$ in the original system, while assign the extra service time ξ_1^o in the buffered system. Since $\xi_1^o \geq \xi_1$, it follows that $\zeta_1^o = \min(\xi_1^o,\ \tau_1) \geq \min(\xi_1,\ \tau_1) = \zeta_1$

implying

$$\hat{S}_1 = S_1 + \zeta_1^o \geq S_1 + \zeta_1,$$

where, by construction, $\hat{S}_1$ is the 1st departure instant in the buffered system. Then, the 2nd service in the original system starts not later than that in the buffered system.

Now, we assign, for the 2nd service in the buffered system, the *same customer* as the 2nd customer entering server in the original system. (In this case we *may change FIFO order* in the buffered system but it does not change the distribution of the remaining workload.) Continuing in such a way and keeping the described synchronization between all variables, we see that, at each instant t, the number of customers $E(t)$ and $\hat{E}(t)$ entered the server in interval $[0, t)$ in the original and buffered system, respectively, are connected as $E(t) \geq \hat{E}(t)$.

It remains to consider the case when the original system *becomes empty* after a departure. Then the following two scenarios are possible. The buffered system is empty at the instant of the next arrival, say t_n. Then both systems start in zero state at instant t_n. Otherwise, in the buffered system, there exists at least one customer $i < n$ at instant t_n. Thus, in both cases, the inequality $E(t) \geq \hat{E}(t)$ remains true as well. Recall that $V(t)$ is the summary work arrived (in both systems) in interval $[0, t)$. Denote by $W(t)$, $\hat{W}(t)$ the remaining work in the original and buffered system, respectively, at instant t^-. Finally, let $B(t)$ and $\hat{B}(t)$ be the busy time of server in original and buffered system, respectively, in interval $[0, t]$. By synchronization of the interarrival times and service times, it follows that the inequality $B(t) \geq \hat{B}(t)$ and the equality

$$V(t) = W(t) + B(t) = \hat{W}(t) + \hat{B}(t), \ t \geq 0 \tag{5.3.8}$$

hold, implying $\hat{W}(t) \geq W(t)$. At the same time, the service time of the nth customer in the buffered system is distributed as $\hat{S}_n = S_n + \zeta_n^o$. (Recall that $\{\zeta_n^o\}$ are iid.) It remains to note that for positive recurrence of the buffered system, it is sufficient to satisfy the following well-known negative drift condition $E[\hat{S}] < E[\tau]$ (see Asmussen [11]). By (5.3.5) and (5.3.7), it is equivalent to

$$E[\hat{S}] = \sum_{k=1}^{M} \frac{p_k}{\mu_k} + \max_{1 \leq k \leq M} \frac{1}{\lambda + \lambda_k^{(r)}} < \frac{1}{\lambda}. \tag{5.3.9}$$

Since $\lambda p_k = \lambda_k$, it is easy to check that (5.3.9) coincides with (5.3.4). Since the

positive recurrent workload process $\hat{W}(t)$ dominates the process $W(t)$, the statement of Theorem (5.3.2) follows. □

Remark 5.3.3. *It is easy to see from the proof that the number of customers in the buffered system dominates the same measure in the original system as well.*

We write condition (5.3.4) as

$$\rho < \min_k \left(\frac{\lambda_k^{(r)}}{\lambda + \lambda_k^{(r)}} \right), \tag{5.3.10}$$

and compare with the necessary stability condition given in (5.3.3), which can also be written as

$$\rho < \min_k \left(\frac{\lambda_k^{(r)}}{\lambda_k + \lambda_k^{(r)}} \right). \tag{5.3.11}$$

It can be seen that the condition given in (5.3.10) and (5.3.11) are identical as those given in (4.3.22) and (4.3.23). This is due to the fact that both the conditions given in this section have been developed for the particular case that $b_k = 1, k = 1, 2, ...M$.

5.4 Some special cases

In this section, we discuss some known particular cases of our model.

Case 1: We first consider a single-server two-way communication system studied in [61, 69], with a *single class* of outgoing calls and a single class of incoming calls with general service time distribution, that is, $M = 2$, $b_1 = 1$ and $b_2 = 0$ (which means that $\lambda_2^{(r)} = 0$ as well). It then follows from (5.2.1) that $C = 1 + \rho_2$, so that equations (5.2.2)–(5.2.5) become

$$\begin{aligned}
P_0 &= \frac{1-\rho_1}{1+\rho_2} = 1 - P_b, \\
P_b^{(1)} &= \rho_1, \quad P_b^{(2)} = \rho_2 \frac{1-\rho_1}{1+\rho_2}, \\
P_{0,b}^{(1)} &= \frac{\lambda_1}{\lambda_1^{(r)}} \frac{\rho}{1+\rho_2}, \\
\text{and } P_{0,0}^{(1)} &= \frac{1-\rho_1-\lambda_1\rho/\lambda_1^{(r)}}{1+\rho_2},
\end{aligned}$$

which are the same as those given in section (4.2), ofcourse with slight changes in notation.

For such a single-class retrial system, the steady-state distribution and performance measures are available in an explicit form (for the specific case that service times are exponentially distributed, the same have been given in section (4.2)). In particular, the following expression for the mean stationary orbit queue, $E[N_1]$, is derived in Phung-Duc et al. [65]:

$$E[N_1] = P_{0,0}^{(1)}\, b\,(\lambda_1 + \lambda_1^{(r)})\frac{a - 2 + (\rho_1 + \theta)(1 - b) + \rho_1\theta(2b - a)}{\lambda_1(1 - a + b)^2} - \frac{\rho}{1 + \rho_2}, \quad (5.4.1)$$

where

$$\theta = \frac{\lambda_1}{\lambda_1 + \mu_2}, \quad a = \frac{(\lambda_1 + \lambda_1^{(r)})(\lambda_1 + \mu_2) + \lambda_2(\lambda_1 + \mu_1) + \lambda_1^{(r)}\mu_1}{\lambda_1(\lambda_1 + \lambda_2 + \lambda_1^{(r)})}, \quad b = \frac{\lambda_1^{(r)}\mu_1(\lambda_1 + \mu_2)}{\lambda_1^2(\lambda_1 + \lambda_2 + \lambda_1^{(r)})}.$$

In Section 5.5, we use (5.4.1) to verify simulation results.

Case 2: Here, we consider a more complicated model with $M = 2$, $b_1 = 1$ and $b_2 = 1$. That is, the model has two classes of incoming calls and no class of outgoing calls. This case has been treated in Avrachenkov et al. [15] by means of generating functions and Riemann–Hilbert boundary value technique. The authors obtained the mean stationary orbit size $E[N_i]$, $i = 1, 2$ explicitly in terms of complex variable integrals which, however, required tough numerical analysis. We also note that a dependence between orbits can be clearly observed from the expressions for $E[N_i]$ obtained in Avrachenkov et al. [15]. In section (5.5.2), we numerically study this dependence for the systems with $b_1 = 1$ and $b_2 \in [0, 1]$.

5.5 Simulation results

In this section, we present the results of some numerical experiments that are performed to gain deeper insight into the model stability and performance. In the first experiment, for model validation purpose, we study the convergence of sample estimates of some measures to their exact theoretical values for which analytical expressions are available. After validating the model, we perform a study of dependence of orbits of different classes by means of empirical correlation. Then, in subsequent experiments, we perform numerical study of some auxiliary performance measures for busy period, namely, the probability of non-empty/empty orbit together

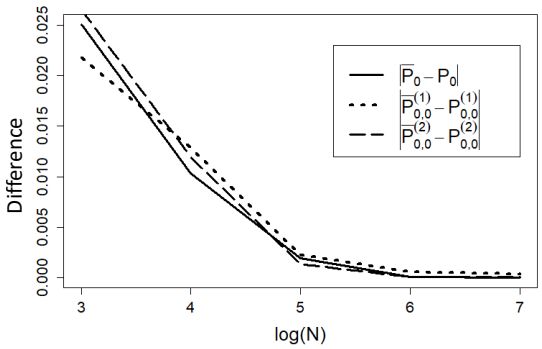

Figure 5.1: Estimated vs. exact values of P_0 and $P^{(k)}_{0,0}$

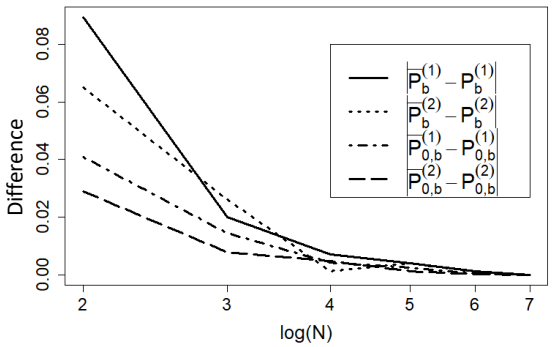

Figure 5.2: Estimated vs. exact values of $P^{(k)}_b$ and $P^{(k)}_{0,b}$

with busy server. Note that deriving theoretical expressions for these measures seem to be highly difficult or even impossible. The simulation is based on the discrete event procedure and is performed using R language [77].

5.5.1 Convergence of performance estimates

In this experiment, we establish the convergence of estimates of those performance measures for which exact values are available via (5.2.2)–(5.2.5). The single trace of simulation run is obtained for each value of number of events, $N = 10^i$, for $i = 3, \ldots, 7$. From each such a trace, estimates of the probabilities P_0, P_b, $P^{(k)}_b$, $P^{(k)}_{0,0}$, and $P^{(k)}_{0,b}$ are obtained. Absolute difference between these simulated estimates (denoted by 'bar') and exact values (obtained from (5.2.2)–(5.2.5)), for a system with $M = 2$ classes of customers, is computed, see Fig. 5.1 and 5.2. Values of p_i, b_i, $\lambda^{(r)}_i$ and μ_i, for $i = 1, 2$, are taken arbitrarily. Service times are assumed to be exponential. Note that we perform multiple experiments for various parameter combinations taken from the system stability region yielding qualitatively the same results, which are exhibited in Fig. 5.1 and 5.2. Fig. 5.1 indicates the variation of absolute difference between the exact and simulated values of P_0 and $P^{(k)}_{0,0}$ against the variation of $\log N$. It is seen that the differences fastly converge to zero for large values of $\log N$. Fig. 5.2 displays the differences for the probabilities $P^{(k)}_b$ and $P^{(k)}_{0,b}$. In all cases, simulation illustrates the validity of theoretical results and, moreover, allows us to select in the subsequent experiments a valid number of customers required to obtain performance measures with high accuracy.

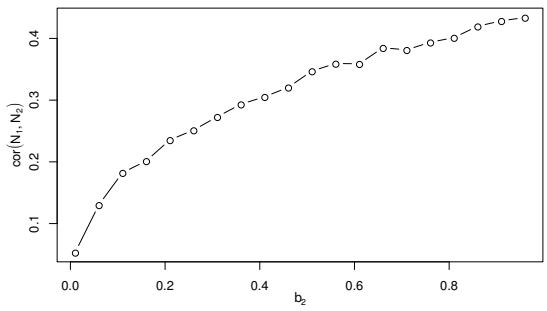

Figure 5.3: Estimate of the orbit sizes correlation $\mathrm{cor}(N_1, N_2)$ vs. the probability to join the 2nd orbit b_2.

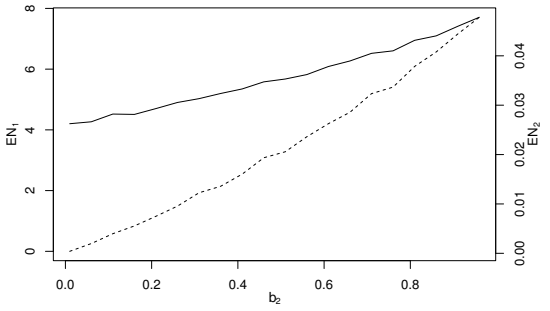

Figure 5.4: Estimates of the mean orbit sizes $E[N_1]$ (solid), $E[N_2]$ (dashed) vs. the probability to join the 2nd orbit b_2.

5.5.2 Inter-orbit correlation

To facilitate understanding of the interaction between orbits of different customer classes (and at the same time following the principle of parsimony), we perform a numerical study of a model with $M = 2$ classes of customers. We select a class of models in such a way that the models studied in Phung-Duc et al. [65] (the model with $b_1 = 1$ and $b_2 = 0$) and Avrachenkov et al. [15] (the model with $b_1 = 1$ and $b_2 = 1$) become borderline cases and may be used as additional accuracy checks.

At that, we fix the following characteristics of an M/M/1-type system with $M = 2$ orbit-queues:

- the arrival rates $\lambda_1 = 0.198, \lambda_2 = 0.02$;
- the service rates $\mu_1 = 20, \mu_2 = 0.01$;
- the retrial rates $\lambda_i^{(r)} = 1, i = 1, 2$;
- the probability to join orbit for class-1 customer $b_1 = 1$.

The values are selected in such a way to guarantee sufficient stability condition (5.3.4) holds true. Note that the model is highly asymmetric: most of arriving customers belong to class 1 and have relatively small service times, while class-2 customers, arriving rarely, bring a relatively large amount of work into the system.

In the experiment, we vary the probability b_2 in the range $(0, 1)$ with step size 0.05. In such a case, the leftmost point is related to the single-class two-way communication model studied in Phung-Duc et al. [65], while the rightmost point is a two-class model from Avrachenkov et al. [15]. Note that expression (5.4.1) is used as an accuracy

check. At each such step (with fixed b_2), we perform discrete-event simulation of the model. We use 2×10^7 arrivals and calculate the sample mean estimates of the orbit-queue sizes $EN_i, i = 1, 2$, and the empirical correlation coefficient of orbit queue size, $\text{cor}(N_1, N_2)$. The results of simulation are depicted at Figures 5.3 and 5.4. It is clearly seen that the correlation becomes larger when the probability b_2 increases. One of the possible intuitive explanations of this phenomenon is the appearance of *oscillations* of the orbit sizes. A large service time of class-2 customers forces the size of the 1st orbit to increase. At the same time, a large retrial rate of class-1 customers forces the 2nd orbit to increase. We also note that similar intermittent behavior of server state was studied in Phung-Duc et al. [65].

5.5.3 Performance in busy period

In Section (5.2), by regenerative approach, we could derive some basic performance measures (see (5.2.2)–(5.2.5)), of which the ones given by (5.2.4)–(5.2.5) are related to an idle server while the concerned orbit queue is non-empty or empty. However, it is relatively more difficult to obtain explicit expressions for the similar performance measures associated with a busy server. This difficulty arises from general service time distribution assumption, as well as complicated interaction of orbits through server occupation (as discussed in section (5.4)). So, in this experiment we attempt to perform such a study numerically (thus below we do not distinguish the notation used for theoretical values and numerical estimates). We are interested in the steady state probability that the server is busy and the kth orbit is non-empty (empty), denoted by $P_{b,b}^{(k)}$ ($P_{b,0}^{(k)}$). We note that the server busy probability, $P_b^{(k)} = P_{b,b}^{(k)} + P_{b,0}^{(k)}$ (for each k) is given by (5.2.3) (this analytical expression may be used as an additional accuracy check).

First, we discuss the effect of the input rate λ on the probabilities $P_{b,b}^{(k)}$ and $P_{b,0}^{(k)}$. In the absence of theoretical expression for these probabilities, here we present how the estimated values of these measures vary with increasing values of λ by keeping the probability p_i fixed, for $M = 2$. The simulation run was performed for $N = 10^6$ customers by assuming $p_1 = 0.9$, retrial rates $\lambda_1^{(r)} = 1$, $\lambda_2^{(r)} = 1.5$, and $b_i = 1$, $i = 1, 2$. Fig. 5.5 depicts the variation of $P_{b,b}^{(k)}$, $k = 1, 2$ and P_b for increasing values of λ for (standard) Pareto service time (with the density $f(x) = \alpha x^{-\alpha-1}$, $\alpha > 1$, $x \geq 1$, and the mean $\frac{\alpha}{\alpha-1}$), where (per-class) parameters α_i are selected in such a way to obtain

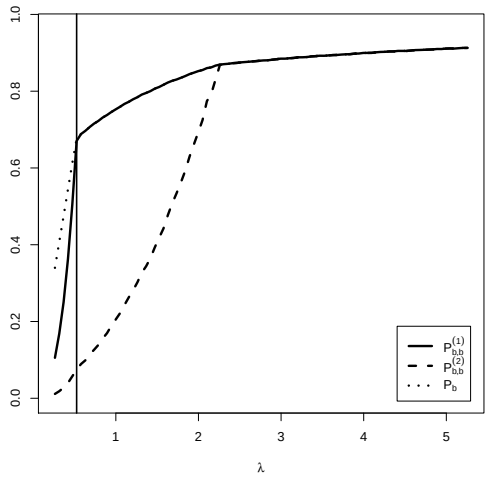

Figure 5.5: Estimated values of $P_{b,b}^{(k)}$ and P_b vs. λ (stability threshold indicated by vertical line at $\lambda = 1/2$).

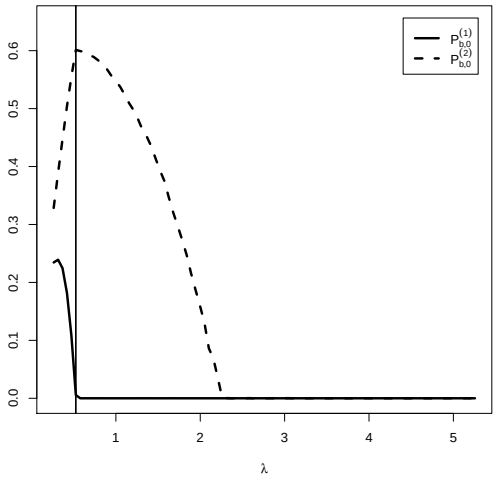

Figure 5.6: Estimated values of $P_{b,0}^{(k)}$ vs. λ (stability threshold indicated by vertical line at $\lambda = 1/2$).

the following service rates

$$\mu_1 = \frac{\alpha_1 - 1}{\alpha_1} = 0.8, \quad \mu_2 = \frac{\alpha_2 - 1}{\alpha_2} = 0.6.$$

It can be seen that, as λ increases beyond critical threshold (stability boundary) indicated by vertical line at point $1/2$ on Fig. 4, both $P_{b,b}^{(1)}$ and $P_{b,b}^{(2)}$ are increasing, as is the busy probability P_b. Note that the same experiment was conducted with exponential service times, *ceteris paribus*. Comparison of the performance measures (for each λ) to the values obtained for Pareto service times confirms insensitivity of the considered measures to the service time distribution. However, to save space, we omit the corresponding graphs.

Since the input rate λ_1 is assumed to be significantly larger than λ_2, and retrial as well as service rates for both classes are not significantly different, $P_{b,b}^{(1)}$ attains larger values at relatively smaller λ as compared to $P_{b,b}^{(2)}$, see Fig. 5.5. Under the same assumptions, a similar behaviour can be observed for $P_{b,0}^{(k)}$, $k = 1, 2$, in Fig. 5.6, where both measures approach zero for large λ corresponding to unstable system.

In Fig. 5.7, we visualize the dependence of $P_{b,0}^{(i)}$ and $P_{b,b}^{(i)}$ on $\lambda^{(r)}$, assuming $M = 2$ and $b_1 = b_2 = 1$. In particular, we configure the system in such a way to experience a transition between the stable and non-stable regimes by modifying $\lambda_1^{(r)}$, *ceteris*

paribus. At that, we study a system with arrival rates $\lambda_1 = 0.125$ and $\lambda_2 = 0.2$, (standard) Pareto service time with service rates $\mu_1 = 0.5$ and $\mu_2 = 0.8$. Then we obtain equal traffic intensities

$$\rho_1 = \rho_2 = 0.25$$

for both classes of customers. To satisfy (5.3.1) for class -2 customers, we set $\lambda_2^{(r)} = 0.4$. At the same time, we vary $\lambda_1^{(r)}$ in the interval $[0.05, 1.5]$ with step size 0.01, making a single simulation run for each such value, and performing estimation after $N = 10^7$ events. It is easy to calculate that $\rho = 0.5$ and see that (5.3.1) holds for $\lambda_1^{(r)} > 0.125$. Indeed, Fig. 5.7 shows that $P_{b,0}^{(1)} > 0$ only for $\lambda_1^{(r)} > 0.125$. In fact, $P_{b,0}^{(1)} = 0$ can be treated as instability of the 1st orbit. This again shows that in fact (5.3.1) must be *stability criterion* of the model.

We also note the nonlinear dependence of $P_{b,0}^{(2)}$ and $P_{b,b}^{(2)}$ on the 1st orbit retrial rate. Moreover, it is interesting to point out the steep increase in $P_{b,b}^{(i)}$, $i = 1, 2$, in the instability region of $\lambda_1^{(r)}$ (see Fig. 5.7, right). The reason for this shape is the increase in server load caused by increase in retrial rate from an overloaded orbit 1 (as long as $\lambda_1^{(r)} < \lambda_1 = 0.125$).

Moreover, we note that if the input rate approaches the stability boundary $\lambda_1^{(r)} = 0.125$, then the inequality $P_{b,0}^{(1)} < P_{b,b}^{(1)}$ holds good, whereas for larger values of $\lambda_1^{(r)}$ (say, $\lambda_1^{(r)} > 0.5$) the inequality is reversed, $P_{b,0}^{(1)} > P_{b,b}^{(1)}$.

Also, at $\lambda_1^{(r)} = 0.25$, in addition to the assumption $\rho_1 = \rho_2$, the following symmetry holds

$$\frac{\lambda_1}{\lambda_1^{(r)}} = \frac{\lambda_2}{\lambda_2^{(r)}}, \quad \frac{\mu_1}{\lambda_1^{(r)}} = \frac{\mu_2}{\lambda_2^{(r)}}.$$

One can expect that in this case the corresponding performance measures for both orbits must be equal, and indeed, Fig. 5.7 shows that in this case $P_{b,0}^{(1)} = P_{b,0}^{(2)}$ and $P_{b,b}^{(1)} = P_{b,b}^{(2)}$.

To illustrate our results, we simulate two-orbit system. The main motivation for our simulation study is to learn the behaviour of the performance measures $P_{b,0}^{(k)}$ and $P_{b,b}^{(k)}$, $k = 1, 2$, which are associated with busy server states, against the variation of some system parameters. As we mentioned above, deriving the analytic expressions for these measures is almost impossible unlike in the case of similar measures associated with idle server states. From the simulation study (see Figure 5.5 and 5.6), we could see that when the total input rate λ increases, both $P_{b,b}^{(1)}$ and $P_{b,b}^{(2)}$ also display the same trend and once the input rate crosses the critical threshold for maintaining the stability of the system, both these measures move to 1 rapidly irrespective of the

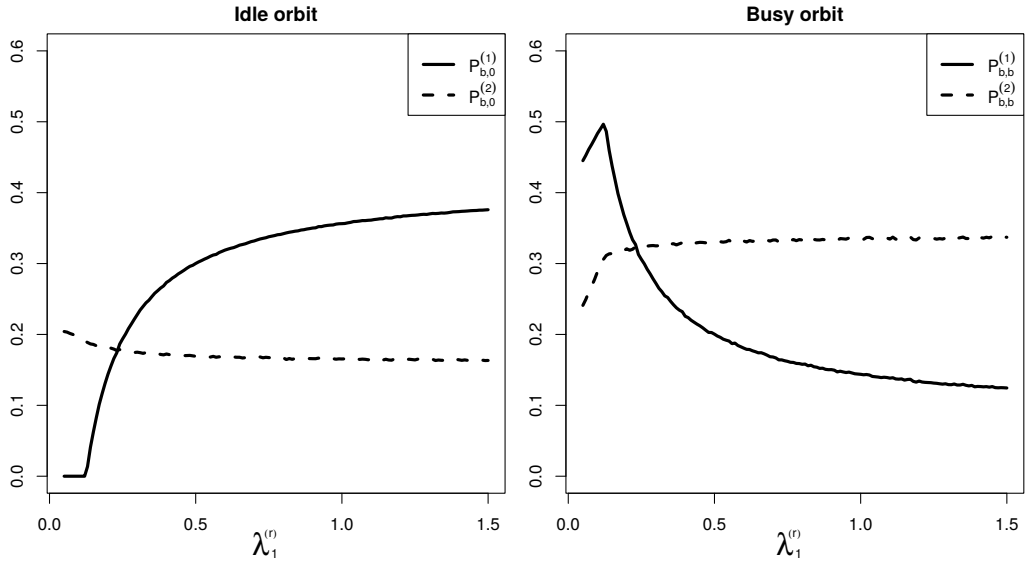

Figure 5.7: Variation of estimated values of $P_{b,0}^{(i)}$, $P_{b,b}^{(i)}$, $i = 1, 2$, for two-class model vs. retrial rate $\lambda_1^{(r)}$.

kind of the service time distribution. As a result, both $P_{b,0}^{(1)}$ and $P_{b,0}^{(2)}$ move to 0 under the unstable regime. From the same situation, it also could be observed that when the class-1 input rate λ_1 is significantly larger than λ_2, *ceteris paribus*, $P_{b,b}^{(1)}$ attains larger value at relatively smaller λ as compared to $P_{b,b}^{(2)}$. In another simulation experiment, as seen from Figure 5.7, behaviour of the same performance measures have been studied against the variation of the class-1 retrial rate $\lambda_1^{(r)}$. Selection of the range of variation of $\lambda_1^{(r)}$ has been made in such a way that the system transits from unstable to stable regime while $\lambda_1^{(r)}$ moves over this range. It can be seen that $P_{b,0}^{(1)} > 0$ only in the stable region. Also, the non-linear dependence of $P_{b,0}^{(2)}$ and $P_{b,b}^{(2)}$ on $\lambda_1^{(r)}$ has been observed. Moreover, when the input rate approaches the least value of $\lambda_1^{(r)}$ that assures the system stability, the inequality $P_{b,0}^{(1)} < P_{b,b}^{(1)}$ appears to hold good.

Chapter 6

Two-way Communication Orbit Queues with Server Vacation

6.1 Introduction

Queuing models with server vacations have been introduced to model situations where the server may not be available for serving customers over some intervals of time. Server vacations may happen due to server failure, lack of work, or another task being assigned to the server. In these systems, the server is not always available to serve its customers. Many researchers like Doshi [30] ,Takagi [71], Tian and Zhang [72] conducted excellent surveys on server vacation models in the queueing literature. An M/M/1 retrial queue with working vacations has been studied by Do [29]. Li et al. [51] considered a discrete time Geo/Geo/1 retrial queue with working vacations and vacation interruption. Liu and Song [52] analysed a system by incorporating non-persistent customers into the Geo/Geo/1 retrial queue with working vacations.

In this chapter, we consider a two-way communication retrial model with multiple classes of incoming and outgoing calls. The server goes for vacation if no request from outgoing or incoming call is received during a random amount of time. These kind of situations arise in many practical cases where, due to excess work load, systems struggle to find any time to spend for their periodical maintenance. The novelty of this work is the inclusion of multiple classes of calls in each stream (outgoing/incoming) requiring class dependent service times. Even though many of the underlying random variates are assumed as exponential, this model can be considered as a general one that can be used for performance analysis of several systems dealing with various classes of customers of heterogeneous nature where the server can opt for vacation if no request from any of this class is pending. Moreover, unlike in many other models,

here we could use a blend of regenerative and matrix analytic approaches to derive the steady state distribution and some system performance measures explicitly.

The structure of this chapter is as follows: In section (6.2), first we describe a model by considering a common orbit and exponential distributional assumptions for service times for both types of calls, and derive the steady state distribution by using matrix analytic approach. After this, regenerative approach is used to find the boundary state probabilities even for the case with general service time assumptions and class dependent orbits for calls, and then we combine both matrix analytic and regenerative approaches for the computation of steady state distribution. Section (6.3) provides various interesting measures which are useful in the system performance analysis. Theoretical results are numerically illustrated in section (6.4).

6.2 Mathematical model

6.2.1 Matrix analytic approach

We consider a single server system in which incoming customers (calls) arrive according to a Poisson rule with rate λ. Upon seeing an idle server, an arriving customer defines its class as i with probability p_i, where $i = 1, 2, ..., M$ and occupies the server for an exponential amount of time with rate μ_i. On the other hand, if the customer meets a busy server, it joins an orbit. The customer at the head of the orbit queue retries to get the server in a gap of exponential amount of time with rate $\lambda^{(r)}$ and upon successful attempt, the customer defines its class as i with probability $p_i^{(r)}$, where $i = 1, 2, ..., M$. Also, there are L classes of internal calls (outgoing calls) -which are assumed to station at a virtual pool and are originating from an infinite source- and the one at the head of the pool is assumed to make attempts to capture the server in a gap of exponential amount of time with rate γ. The successful outgoing call will be of jth type with probability q_j, where $\sum_j^L q_j = 1$ and its service time is exponential with rate η_j, $j = 1, 2, ..., L$. If the server is idle for an exponential amount of time having mean $\frac{1}{\alpha}$ and there is no attempt either from incoming or outgoing customers to get the server during this period, then the server decides to go for vacation and the vacation period is exponentially distributed with parameter β. After vacation, the server will again come back to the idle state.

In relation with the events stated above, let the state variables at time t be defined as follows:

- $N(t)$, the number of customers in the orbit.
- $J(t)$, the state of the server.

More specifically,

$$J(t) = \begin{cases} 0 & \text{if the server is idle} \\ 1, ..., M & \text{if busy with serving M type of incoming calls} \\ M+1, M+2, .., M+L & \text{if busy with serving L type of outgoing calls} \\ M+L+1 & \text{if the server is in vacation mode} \end{cases}$$

Then the process $\{X(t) = (N(t), J(t)); t \geq 0\}$ is a continuous time Markov chain with state space $E = \cup_i E_i$, where $E_i = \{0, 1, 2, 3, ...\} \times \{i\}$, for $i = 0, 1, 2, \ldots, M + L + 1$.
Since $N(t)$ may jump only one step in an infinitesimal time interval, the process $\{X(t), t \geq 0\}$ is a QBD (Quasi bith-death process) with the generator

$$Q = \begin{bmatrix} A^{00} & A_0 & 0 & 0 & 0 & \cdots & 0 \\ A_2 & A_1 & A_0 & 0 & 0 & \cdot & 0 \\ 0 & A_2 & A_1 & A_0 & 0 & \cdot & 0 \\ 0 & 0 & A_2 & A_1 & A_0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

The matrices appeared in Q are explicitly expressed as follows:

$$A_0 = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & \lambda & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda \end{pmatrix};$$

$$A_2 = \begin{pmatrix} 0 & \lambda^{(r)} p_1^{(r)} & \ldots & \lambda^{(r)} p_M^{(r)} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix};$$

$$A_1 = \begin{pmatrix} -\lambda-\lambda^{(r)}-\gamma-\alpha & \lambda p_1 & \lambda p_2 & \dots & \lambda p_M & \gamma q_1 & \dots & \gamma q_L & \alpha \\ \mu_1 & -\lambda-\mu_1 & 0 & \dots & \dots & \dots & \dots & 0 & 0 \\ \mu_2 & 0 & -\lambda-\mu_2 & \dots & \dots & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ \mu_M & 0 & 0 & \dots & -\lambda-\mu_M & \dots & \dots & 0 & 0 \\ \eta_1 & 0 & 0 & \dots & 0 & -\lambda-\eta_1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & 0 \\ \eta_L & 0 & 0 & \dots & 0 & 0 & \dots & -\lambda-\eta_L & 0 \\ \beta & 0 & 0 & \dots & 0 & 0 & \dots & 0 & -\lambda-\beta \end{pmatrix};$$

$$A^{00} = A_1 + \begin{pmatrix} \lambda^{(r)} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

For representational convenience, we make the following notations:

$$\rho_i = \lambda\frac{p_i}{\mu_i},\ i = 1,\dots,M;\ \rho = \sum_{i=1}^{M}\rho_i;$$
$$\rho_i^{(r)} = \lambda^{(r)}\frac{p_i^{(r)}}{\mu_i},\ i = 1,\dots,M;\ \rho^{(r)} = \sum_{i=1}^{M}\rho_i^{(r)};$$
$$\rho_j^{(o)} = \gamma\frac{q_j}{\eta_j},\ j = 1,\dots,L;\ \rho^{(o)} = \sum_{j=1}^{L}\rho_j^{(o)};$$
$$\rho^{(v)} = \frac{\alpha}{\beta}.$$

A necessary and sufficient condition for ergodicity of the above QBD process is (Neuts [63])

$$\alpha A_0\mathbf{e} < \alpha A_2\mathbf{e}, \tag{6.2.1}$$

where α is the solution of the system

$$\begin{cases} \alpha A &= 0 \\ \alpha \mathbf{e} &= 1 \end{cases}$$

and $A = A_0 + A_1 + A_2$.

By solving the above system, we obtain
$\alpha = \alpha_0(1, \rho_1+\rho_1^{(r)}, ..., \rho_M+\rho_M^{(r)}, \rho_1^{(o)}, ..., \rho_L^{(o)}, \rho^{(v)})$, where $\alpha_0 = (1+\rho+\rho^{(r)}+\rho^{(o)}+\rho^{(v)})^{-1}$.
So, the stability condition (6.2.1) yields

$$\lambda(\rho + \rho^{(r)} + \rho^{(o)} + \rho^{(v)}) < \lambda^{(r)}. \tag{6.2.2}$$

This gives a second degree polynomial in λ, which has two distinct roots. The rightmost value of λ that offers the stability of the system is

$$\lambda^* = \frac{-\rho^{(r)} - \rho^{(o)} - \rho^{(v)} + \sqrt{(\rho^{(r)} + \rho^{(o)} + \rho^{(v)})^2 + 4\lambda^{(r)} \sum_{i=1}^{M} \frac{p_i}{\mu_i}}}{2\sum_{i=1}^{M} \frac{p_i}{\mu_i}} \tag{6.2.3}$$

and hence the system is stable for $\lambda < \lambda^*$ with the other parameters fixed.
Here, the matrix A_2 can be written in the form $A_2 = cr$, where $c = (\lambda^{(r)}, 0, 0, ..., 0)$ is a column vector and $r = (0, p_1^{(r)}, .., p_M^{(r)}, 0, 0, .., 0)$ is a row vector. Then the first passage time matrix G and the rate matrix R are obtained explicitly as

$$G = \mathbf{e}r \tag{6.2.4}$$

and

$$R = -A_0(A_1 + A_0G)^{-1}. \tag{6.2.5}$$

For more details on G and R, refer Latouche and Ramaswami [50]. Then the steady state vector $\pi = (\pi_0, \pi_1, ..)$, where $\pi_i = (\pi_{i,0}, \pi_{i,1}, ..., \pi_{i,M+L+1})$ represents the probability that the orbit is having i customers with the server being any of the said states, assumes the matrix geometric form

$$\pi_i = \pi_0 R^i, i \geq 1 \tag{6.2.6}$$

and π_0 is obtained from the system

$$\begin{cases} \pi_0(A^{00} + RA_2) &= \mathbf{0} \\ \pi_0(I-R)^{-1}\mathbf{e} &= 1 \end{cases}$$

However, computation of π_0 demands matrix inversion which, in general, is complicated. To avoid this situation, we consider the following equivalent system

$$\begin{cases} \xi_0(I-R)(A^{00} + RA_2) = \mathbf{0} \\ \xi_0\mathbf{e} = 1 \end{cases}$$

by taking $\xi_0 = \pi_0(I-R)^{-1}$. Once ξ_0 is computed, $\pi_0 = \xi_0(I-R)$.
It is to be noted that $\xi_0 = \sum_i \pi_i$ so that the jth component of ξ_0 represents the steady state probability that the server is in state j, for $j = 0, 1, 2, \ldots, M+L+1$. Let us write $\xi_0 = (P_0, P_1^{(r)}, P_2^{(r)}, \ldots, P_M^{(r)}, P_{M+1}^{(o)}, P_{M+2}^{(o)}, \ldots, P_{M+L}^{(o)}, P_{M+L+1}^{(v)})$. In the next section, we compute the probability measures appeared in ξ_0 by using regenerative approach, which in turn can be used for computing ξ_0 explicitly for the particular case that $M = 1$.

6.2.2 Regenerative approach

In this section, instead of exponential distribution assumptions taken for service times of incoming and outgoing calls, we go with general service time assumptions. Also, we assume that after finding a busy server, a class k incoming call joins a k-orbit (unlike the assumption in section (6.2.1), where we use a common orbit for all classes of incoming calls). Regenerative approach helps us to compute the components of ξ_0 (they themselves serve as important measures to assess the system performance) even in the said non-Markovian set up. More precisely, in connection with service times, we make the following assumptions (note that the other assumptions remain as such):

- Class i incoming calls possess independent and identically distributed (iid) service times $\{S_n^{(i)}, n \geq 1\}$ following a general distribution with mean $E[S^{(i)}] = \frac{1}{\mu_i} < \infty$. Upon meeting a busy server, each of the class k customer joins a k-orbit.

- Class j outgoing calls have generally distributed iid service times $\{Z_n^{(j)} : n \geq 1\}$ with rate $\eta_j = 1/E[Z^{(j)}]$.

In connection with the interval $[0, t)$, denote by

- $H'(t)$, the summary work arrived in the system.That is, the total service time required for all incoming and outgoing customers.
- $H(t)$, the summary work of all incoming arrivals.
- $H_k(t)$, the summary work of class-k incoming customers so that

$$H(t) = \sum_k H_k(t).$$

- $B(t)$, the busy time of the server (including the service of outgoing calls).
- $V(t)$, the vacation time of the server.
- $I(t)$, the idle time of the server.
  Then

$$B(t) + V(t) + I(t) = t.$$

  At time instant t^-, let

- $N_k(t)$ be the number of orbital customers of class k and $W_k(t)$ be the remaining workload for the entire k class customers present in the system. Since we are talking about a stable system, as $t \to \infty$, $\sum_{k=1}^{M} W_k(t) = o(t)$ with probability 1.

- $S(t)$ be the remaining service time of the customer who is undergoing service. Therefore, $S(t) = 0$ if the server is free at t^-. Also, as $t \to \infty$, $S(t) = o(t)$ with probability 1.

Now, we consider the basic summary queue size process defined as follows:

$$Y(t) = \sum_k N_k(t) + C(t),\ t \geq 0,$$

where $C(t) \in \{0,\ 1\}$ is defined as the number of customers attended by the server at instant t^-.

Let $\{t_n,\, n \geq 1\}$ be the instants of the superposed arrival process, which is Poisson with rate λ. At instant t_n, denote $Y(t_n) = Y_n$, $n \geq 1$.
We assume that at instant t=0, the first customer breaks the idle state of the system. So, put $T_0 = 0$ and define

$$T_{n+1} = \inf\Big(t_k > T_n : Y_k = 0\Big),\ n \geq 0.$$

$\{T_n\}$ are classical regeneration points of the basic process Y, which is having $T_{n+1} - T_n$ as regeneration periods. We know that the regenerative process $\{Y(t)\}$ is positive recurrent if the mean generic period $E[T]$ (where T is the generic period)is finite and there exists the weak limit $Y(t) \Rightarrow Y$, which denotes the number of incoming customers in the system in the stationary scenario.

Denote $\hat{G}_j(t)$ as the renewal process generated by the attempts of class-j outgoing calls in interval $[0, t)$. Let $\{v_i^{(j)},\, i \geq 1\}$ be the renewal points of the process $\hat{G}_j$ and let $\hat{C}_i^{(j)} = C(v_i^{(j)} - 0)$ be the state of server when the ith j-outgoing call (in the renewal process $\hat{G}_j$) appears, $j = 1, \ldots, L$. Therefore, in the interval $[0, t)$, the number of successful attempts made by class-j customers is

$$G_j(t) = \sum_{i=1}^{\hat{G}_j(t)} \mathbf{1}(\hat{C}_i^{(j)} = 0),$$

where $\mathbf{1}$ stands for the indicator function. Let $U_j(t)$ be the summary work generated by class-j outgoing calls in interval $[0, t)$. Then

$$U_j(t) = \sum_{i=1}^{\hat{G}_j(t)} \mathbf{1}(\hat{C}_i^{(j)} = 0) Z_i^{(j)},\ j = 1, \ldots, L.$$

Denote by $A_k(t)$ the number of class-k arrivals in interval $[0, t)$.

Then, we have the balance equation

$$H'(t) \;=\; \sum_{k=1}^{M} \sum_{i=1}^{A_k(t)} S_i^{(k)} + \sum_{j=1}^{L} U_j(t) = S(t) + \sum_{k=1}^{M} W_k(t) + t - I(t) - V(t) \quad (6.2.7)$$

As $t \to \infty$, for each k, with probability 1

$$\frac{1}{t}\sum_{i=1}^{A_k(t)} S_i^{(k)} = \frac{1}{A_k(t)}\sum_{i=1}^{A_k(t)} S_i^{(k)} \cdot \frac{A_k(t)}{t} \to \rho_k. \tag{6.2.8}$$

It is to be noted that $\rho_j^{(o)} = \gamma q_j/\eta_j$, $j = 1, \ldots, L$ and $\rho^{(v)} = \alpha/\beta$. Then, by the strong law of large numbers and PASTA property

$$\frac{U_j(t)}{t} \to P_0\rho_j^{(o)}, \frac{V(t)}{t} \to P_0\rho^{(v)} \tag{6.2.9}$$

so that

$$\lim_{t\to\infty} \frac{H'(t)}{t} = \rho + \sum_{j=1}^{L} \rho_j^{(o)} P_0 = \rho + \rho^{(o)} P_0$$

and

$$\lim_{t\to\infty} \frac{1}{t}\Big(S(t) + \sum_{k=1}^{M} W_k(t) + t - I(t) - V(t)\Big) = 1 - P_0 - P_0\rho^{(v)}.$$

Hence from the above two expressions and eqn (6.2.7), we get

$$P_0 = \frac{1-\rho}{1+\rho^{(o)}+\rho^{(v)}}. \tag{6.2.10}$$

Now, by the definition of $U_j(t)$ and by eqns (6.2.9) and (6.2.10), we have

$$P_{M+j}^{(o)} = \lim_{t\to\infty} \frac{U_j(t)}{t} = \frac{1-\rho}{1+\rho^{(o)}+\rho^{(v)}}\rho_j^{(o)} \quad \text{for} \quad j = 1, 2, \ldots, L \tag{6.2.11}$$

and

$$P_{M+L+1}^{(v)} = \lim_{t\to\infty} \frac{V(t)}{t} = \frac{1-\rho}{1+\rho^{(o)}+\rho^{(v)}}\rho^{(v)}. \tag{6.2.12}$$

Also, related to $[0, t)$, define

- $B_k(t)$, the busy time of the server caused by k-type incoming customers and
- $S_k(t)$, the remaining service time of the kth type incoming customer who is undergoing service. Therefore, $S_k(t) = 0$ if the server is not busy with the said

type of customer at t^-. Also, as $t \to \infty$, $S_k(t) = o(t)$ with probability 1.

Then

$$H_k(t) \quad = \quad \sum_{i=1}^{A_k(t)} S_i^{(k)} = S_k(t) + W_k(t) + B_k(t).$$

Now, with probability 1

$$\lim_{t\to\infty} \frac{B_k(t)}{t} = P_k^{(r)}$$

so that

$$P_k^{(r)} = \rho_k \quad \text{for} \quad k = 1, 2, .., M. \tag{6.2.13}$$

Thus, eqns (6.2.10)- (6.2.13) help us to compute ξ_0 and hence π_0 and the stationary vector π explicitly in the particular case that $M = 1$.
In the next section, we are computing some important measures that are useful for system performance analysis.

6.3 Some important system performance measures

Here, we discuss some important performance measures that are associated with the Markovian model considered in section (6.2.1).

6.3.1 First passage time and busy period analysis of the orbit

In our model, we define the busy period of the orbit as the time between the epoch at which an incoming unit comes to the system with busy server and empty orbit and the first epoch thereafter at which the orbit becomes empty again.To study this busy period, we first introduce the matrix $G(k, x)$. The entry $G_{jj'}(k, x)$ is the conditional probability that the Markov process starts from the state (i, j) at time $t = 0$ and reaches the level $i - 1$ by entering the state $(i - 1, j')$ for the first time no later than time x after exactly k transitions to the left. This is defined for $x \geq 0, k \geq 1, 0 \leq j, j' \leq M + L + 1$. Let

$$G^*(z, s) = \sum_{k=1}^{\infty} z^k \int_0^{\infty} e^{-sx} dG(k, x).$$

Then, $G^*(z,s)$ is the minimal non-negative solution to the equation

$$zA_2 - (sI - A_1)G^*(z,s) + A_0 G^{*^2}(z,s) = 0$$

for $s > 0$ and $0 < z < 1$.
Let τ be the first passage time from level i to level $i-1$. It is well known that (see Neuts [63]) $\lim_{z\to 1, s\to 0} G^*(z,s) = G = (G_{jj'})$ where

$$G_{jj'} = P\{\tau < \infty \text{ and } (N(t), J(t) = (i-1, j') | (N(0), J(0) = (i,j)\}.$$

Define $\tilde{m}$ as a row vector of dimension $M + L + 2$ with elements m_{1j}, where m_{1j} is the mean first passage time from level i (given that the first passage time started in (i,j)) to the level $i-1$, $i \geq 1$.
Then

$$\begin{aligned} \tilde{m} &= -\frac{\partial}{\partial s} G^*(z,s)|_{s=0,z=1} \\ &= -(A_1 + A_0(I+G))^{-1}\mathbf{e}. \end{aligned}$$

Similarly, define matrices $G^{*(0,0)}(z,s)$, $\tilde{m_1}^{(0,0)}$ for the first passage time from the level zero to itself. Then

$$G^{*(0,0)}(z,s) = (sI - A^{00})^{-1} A_0 G^*(z,s)$$

and

$$\tilde{m}^{*(0,0)} = -A^{00^{-1}}(A_0\tilde{m} + \mathbf{e}).$$

A detailed analysis of the similar kind can also be seen in Neuts [63] and Deepak et al. [25]. Busy period of the orbit is basically the first passage time from the level 1 to the level 0. So, if we know the G matrix, which is given by eqn (6.2.4), all these measures can be computed directly.

6.3.2 Distribution of the number of orbital incoming calls served during system busy period

We define the system busy period as the duration of time between the epoch at which the server becomes busy (either by incoming or outgoing call), after it becomes idle

and orbit is empty, and the epoch at which the orbit becomes empty again. In order to distinguish this from orbit busy period, we call this as *system busy period* even though it is not the system busy period in strict sense.

Here, we are using the method presented in Amador and Artalejo [1] and Deepak [26] to derive the distributions of some important random variates. Let t_n be the nth service completion epoch (including both incoming as well as outgoing calls) and i_n be the number of orbital customers at $t_n + 0$. Then, $\{i_n : n \geq 1\}$ is a Markov chain having non-negative integers as states. Let P_{il} be the one-step transition probability of the above Markov chain. That is,

$$P_{il} = Pr\{i_{n+1} = l | i_n = i\}.$$

Then

$$\begin{aligned} P_{0l} &= \sum_{i=1}^{M} \frac{\lambda p_i}{\lambda+\gamma+\alpha} \int_0^\infty \frac{e^{-\lambda u}(\lambda u)^l}{l!} \mu_i e^{-\mu_i u} du + \sum_{j=1}^{L} \frac{\gamma q_j}{\lambda+\gamma+\alpha} \int_0^\infty \frac{e^{-\lambda u}(\lambda u)^l}{l!} \eta_j e^{-\eta_j u} du \\ &= \frac{\lambda}{\lambda+\gamma+\alpha} \sum_{i=1}^{M} \left(\frac{\lambda}{\lambda+\mu_i}\right)^{l+1} p_i \mu_i + \frac{\gamma}{\lambda(\lambda+\gamma+\alpha)} \sum_{j=1}^{L} \left(\frac{\lambda}{\lambda+\eta_j}\right)^{l+1} q_j \eta_j. \qquad (6.3.1) \end{aligned}$$

Let the random variable B denote the number of orbital units that are taken into service by retrial during a system busy period and $x_i(b)$ be the probability that exactly $b \geq 0$ orbital customers are getting service during the remaining busy period, given that a service has just been completed leaving behind i customers in the orbit, where $0 \leq i \leq b$. Clearly, $x_0(b) = \delta_{b0},\ b \geq 0$ and for $i \geq 1$,

$$\begin{aligned} x_i(b) &= \sum_{k=1}^{M} \frac{\lambda p_k}{\lambda+\gamma+\alpha+\lambda^{(r)}} \int_0^\infty \sum_{l=i}^{b} \frac{e^{-\lambda u}(\lambda u)^{l-i}}{(l-i)!} \mu_k e^{-\mu_k u} du\ x_l(b) \\ &+ \sum_{k=1}^{M} \frac{\lambda^{(r)} p_k^{(r)}}{\lambda+\gamma+\alpha+\lambda^{(r)}} \int_0^\infty \sum_{l=i-1}^{b-1} \frac{e^{-\lambda u}(\lambda u)^{l-i+1}}{(l-i+1)!} \mu_k e^{-\mu_k u} du\ x_l(b-1) \\ &+ \sum_{j=1}^{L} \frac{\gamma q_j}{\lambda+\gamma+\alpha+\lambda^{(r)}} \int_0^\infty \sum_{l=i}^{b} \frac{e^{-\lambda u}(\lambda u)^{l-i}}{(l-i)!} \eta_j e^{-\eta_j u} du\ x_l(b). \end{aligned}$$

Therefore

$$x_i(b) = \frac{1}{\lambda+\lambda^{(r)}+\gamma+\alpha}\Big[\sum_{k=1}^{M}\sum_{l=i}^{b}\left(\frac{\lambda}{\lambda+\mu_k}\right)^{l-i+1} p_k\mu_k x_l(b) + \sum_{k=1}^{M}\sum_{l=i-1}^{b-1}\frac{\lambda^{l-i+1}}{(\lambda+\mu_k)^{l-i+2}}\lambda^{(r)}p_k^{(r)}\mu_k x_l(b-1) + \sum_{j=1}^{L}\sum_{l=1}^{b}\frac{\lambda^{l-i}}{(\lambda+\eta_j)^{l-i+1}}\gamma q_j\eta_j x_l(b)\Big].$$

To derive $x_i(b)$, we consider the time between two successive service completion epochs where after a service completion, the idle time of the server will be broken either by primary arrivals or retrials or outgoing customers. In all cases, we keep count of the number of incoming arrivals occurring during the subsequent service time, so that the total number of orbital customers never reaches b + 1 at this service completion epoch.

The above system of equations can be written in the following matrix form

$$M_b x_b = B_b \tilde{x}_{b-1},$$

where $x_b = (x_1(b), x_2(b), ..., x_b(b))'$, $\tilde{x}_{b-i} = (\delta_{bi}, x_{b-i})'$ $M_b = (m_{il}), B_b = (b_{il})$ are square matrices of order b defined by

$$m_{il} = \begin{cases} 0 & if \quad 1 \le l < i \le b \\ (\lambda+\gamma+\alpha+\lambda^{(r)}) - \lambda\sum_{k=1}^{M}\frac{p_k\mu_k}{\lambda+\mu_k} - \gamma\sum_{j=1}^{L}\frac{q_j\eta_j}{\lambda+\eta_j} & if \quad l = i \\ -\sum_{k=1}^{M}\frac{\lambda^{l-i+1}p_k\mu_k}{(\lambda+\mu_k)^{l-i+1}} - \sum_{j=1}^{L}\frac{\lambda^{l-i}\gamma q_j\eta_j}{(\lambda+\eta_j)^{l-i+1}} & if \quad 1 \le i \le l \le b \end{cases}$$

$$b_{il} = \begin{cases} 0 & if \quad 1 \le l < i \le b \\ \sum_{k=1}^{M}\frac{\lambda^{l-i}\lambda^{(r)}p_k^{(r)}\mu_k}{(\lambda+\mu_k)^{l-i+1}} & if \quad 1 \le i \le l \le b \end{cases}$$

This equation can be solved recursively to get x_b, which eventually can be used to get the distribution of B by the formula

$$P\{B = b\} = \sum_{i=0}^{b} P_{0i}x_i(b).$$

6.3.3 Distribution of the number of incoming calls that are taken into service upon their arrival during system busy period

Let A be the number of incoming calls that are directly taken into service upon their arrival during a system busy period. To derive A, we need to approximate our model with the one having finite orbit capacity say, W. Let $y_i(a)$ be the probability that a number of incoming customers get the service, upon their arrival, during the remaining busy period, given that a service has just been completed leaving behind i customers in the orbit, for $0 \leq i \leq W$. Then

$$P\{A = a\} = \sum_{i=0}^{W} P_{0i} y_i(a) = \sum_{i=0}^{W-1} P_{0i} y_i(a) + (1 - \sum_{i=0}^{W-1} P_{0i}) y_W(a).$$

Now, for $1 \leq i \leq W$,

$$\begin{aligned}
y_i(0) &= \sum_{k=1}^{M} \frac{\lambda^{(r)} p_k^{(r)}}{\lambda + \gamma + \alpha + \lambda^{(r)}} \int_0^\infty \sum_{l=i-1}^{W} \frac{e^{-\lambda u}(\lambda u)^{l-i+1}}{(l-i+1)!} \mu_k e^{-\mu_k u} y_l(0) du \\
&+ \sum_{j=1}^{L} \frac{\gamma q_j}{\lambda + \gamma + \alpha + \lambda^{(r)}} \int_0^\infty \sum_{l=i}^{W} \frac{e^{-\lambda u}(\lambda u)^{l-i}}{(l-i)!} \eta_j e^{-\eta_j u} y_l(0) du \\
&= \frac{1}{\lambda + \lambda^{(r)} + \gamma + \alpha} \Bigg[\sum_{k=1}^{M} \sum_{l=i-1}^{W} \frac{\lambda^{l-i+1}}{(\lambda + \mu_k)^{l-i+2}} \lambda^{(r)} p_k^{(r)} \mu_k y_l(0) \\
&+ \sum_{j=1}^{L} \sum_{l=i}^{W} \frac{\lambda^{l-i}}{(\lambda + \eta_j)^{l-i+1}} \gamma q_j \eta_j y_l(0) \Bigg].
\end{aligned}$$

For $a \geq 1$,

$$\begin{aligned}
y_i(a) &= \sum_{k=1}^{M} \frac{\lambda p_k}{\lambda + \gamma + \alpha + \lambda^{(r)}} \int_0^\infty \sum_{l=i}^{W} \frac{e^{-\lambda u}(\lambda u)^{l-i}}{(l-i)!} \mu_k e^{-\mu_k u} y_l(a-1) du \\
&+ \sum_{k=1}^{M} \frac{\lambda^{(r)} p_k^{(r)}}{\lambda + \gamma + \alpha + \lambda^{(r)}} \int_0^\infty \sum_{l=i-1}^{W} \frac{e^{-\lambda u}(\lambda u)^{l-i+1}}{(l-i+1)!} \mu_k e^{-\mu_k u} y_l(a) du \\
&+ \sum_{j=1}^{L} \frac{\gamma q_j}{\lambda + \gamma + \alpha + \lambda^{(r)}} \int_0^\infty \sum_{l=i}^{W} \frac{e^{-\lambda u}(\lambda u)^{l-i}}{(l-i)!} \eta_j e^{-\eta_j u} y_l(a) du
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda+\lambda^{(r)}+\gamma+\alpha}\Bigg[\sum_{k=1}^{M}\sum_{l=i}^{W}\left(\frac{\lambda}{\lambda+\mu_k}\right)^{l-i+1} p_k\mu_k y_l(a-1)\\
&+ \sum_{k=1}^{M}\sum_{l=i-1}^{W}\frac{\lambda^{l-i+1}}{(\lambda+\mu_k)^{l-i+2}}\lambda^{(r)}p_k^{(r)}\mu_k y_l(a)\\
&+ \sum_{j=1}^{L}\sum_{l=1}^{W}\frac{\lambda^{l-i}}{(\lambda+\eta_j)^{l-i+1}}\gamma q_j\eta_j y_l(a)\Bigg]
\end{aligned}
$$

and $y_0(a) = \delta_{a0}$.
The above system of equations can be written in the matrix form

$$
\begin{aligned}
M'y(0) &= b(0)\\
M'y(a) &= B'y(a-1), \quad a \geq 1
\end{aligned}
$$

where $y(a) = (y_1(a), y_2(a), ..., y_W(a))'$. $M' = (m'_{il})$, and $B' = (b'_{il})$ are square matrices of order W defined by

$$
m'_{il} = \begin{cases}
0 & if \;\; 1 \leq l < i-1 \;\; \& \;\; i \leq W\\
-\sum_{k=1}^{M}\frac{\lambda^{(r)}p_k^{(r)}\mu_k}{\lambda+\mu_k} & if \;\; l = i-1\\
(\lambda+\gamma+\alpha+\lambda^{(r)}) - \lambda\lambda^{(r)}\sum_{k=1}^{M}\frac{p_k^{(r)}\mu_k}{(\lambda+\mu_k)^2} - \gamma\sum_{j=1}^{L}\frac{q_j\eta_j}{\lambda+\eta_j} & if \;\; l = i\\
-\lambda^{l-i+1}\lambda^{(r)}\sum_{k=1}^{M}\frac{p_k^{(r)}\mu_k}{(\lambda+\mu_k)^{l-i+2}} - \gamma\lambda^{l-i}\sum_{j=1}^{L}\frac{q_j\eta_j}{(\lambda+\eta_j)^{l-i+1}} & if \;\; 1 \leq i \leq l \leq W
\end{cases}
$$

$$
b'_{il} = \begin{cases}
0 & if \;\; 1 \leq l < i \leq W\\
\sum_{k=1}^{M}\frac{\lambda^{l-i+1}p_k\mu_k}{(\lambda+\mu_k)^{l-i+1}} & if \;\; 1 \leq i \leq l \leq W
\end{cases}
$$

Also,

$$
b(0) = \begin{pmatrix} \sum_{k=1}^{M}\frac{\lambda^{(r)}p_k^{(r)}\mu_k}{\lambda+\mu_k} & 0 & \cdots & 0 \end{pmatrix}'.
$$

To derive $y_i(a)$, we use a similar argument that we have used for deriving $x_i(b)$. Depending on the type of the customer which breaks the idle period of the server, the index a will fall to different levels and by keeping track of the number of primary customers arriving in the concerned service period, it is possible to write the system as above. From the above system of equations, it can be seen that the value of $y(a)$ is completely determined by the value of $y(a-j)$ for $j = 1, 2, ...a$. We will start with $y_0(a) = \delta_{a0}$ and then we can solve for $y(0)$, which can be used later for the computation of $y(1)$ and so on. In this manner, for any a, $y(a)$ can be computed recursively.

6.3.4 Distribution of number of retrials made by an orbital customer

Here, we use the method proposed in Artalejo and Lopez-Herrero [6], which deals with the distribution of number of retrials in M/G/1 and M/M/C set up, to compute the same distribution in our model. Let $\hat{R}$ denotes the number of retrials made by a tagged incoming customer in a system having a single class of incoming and a single class of outgoing calls. Let us define $z^r_{l,(n,i)}$ as the probability that the tagged customer produces exactly r retrials before getting the server successfully, given that it has accumulated l retrials and the present state of the system is (n, i).
Then

$$P\{\hat{R}=0\} = \sum_{n=0}^{\infty}\pi_{n,0}, \tag{6.3.2}$$

$$\text{and} \quad P\{\hat{R}=r\} = \sum_{n=0}^{\infty}\sum_{i=1}^{3}\pi_{n,0}z^r_{l,(i,n)}, \quad r\geq 1. \tag{6.3.3}$$

Now,

$$\begin{aligned} z^r_{r-1,(n,0)} &= \frac{\lambda}{\lambda+\lambda^{(r)}+\lambda^{(o)}+\alpha}z^r_{r-1,(n,1)} + \frac{\lambda^{(o)}}{\lambda+\lambda^{(r)}+\lambda^{(o)}+\alpha}z^r_{r-1,(n,2)} \\ &+ \frac{\alpha}{\lambda+\lambda^{(r)}+\lambda^{(o)}+\alpha}z^r_{r-1,(n,3)} + \frac{\lambda^{(r)}}{\lambda+\lambda^{(r)}+\lambda^{(o)}+\alpha} \end{aligned} \tag{6.3.4}$$

and

$$z^r_{r-1,(n,i)} = \frac{g_i}{\lambda+\lambda^{(r)}+g_i}z^r_{r-1,(n,0)} + \frac{\lambda}{\lambda+\lambda^{(r)}+g_i}z^r_{r-1,(n+1,i)}. \tag{6.3.5}$$

For $l < r-1$,

$$\begin{aligned} z^r_{l,(n,0)} &= \frac{\lambda}{\lambda+\lambda^{(r)}+\lambda^{(o)}+\alpha}z^r_{l,(n,1)} + \frac{\lambda^{(o)}}{\lambda+\lambda^{(r)}+\lambda^{(o)}+\alpha}z^r_{l,(n,2)} \\ &+ \frac{\alpha}{\lambda+\lambda^{(r)}+\lambda^{(o)}+\alpha}z^r_{l,(n,3)} \end{aligned} \tag{6.3.6}$$

and

$$\begin{aligned} z^r_{l,(n,i)} &= \frac{g_i}{\lambda+\lambda^{(r)}+g_i} z^r_{l,(n,0)} + \frac{\lambda}{\lambda+\lambda^{(r)}+g_i} z^r_{l,(n+1,i)} \\ &+ \frac{\lambda^{(r)}}{\lambda+\lambda^{(r)}+g_i} z^r_{l+1,(n,i)}. \end{aligned} \tag{6.3.7}$$

Note that in the above $i = 1, 2, 3$ and $g_1 = \mu, g_2 = \gamma, g_3 = \beta$.

To derive formulas (6.3.4)– (6.3.7), we don't consider all vain retrials because they neither affect the event under the study nor modify the current system state. We develop a method to find an approximate solution based on the truncation of the orbit capacity as say, W.
Then, equation (6.3.4) and (6.3.6) remain valid for $1 \leq n \leq W$ for the truncated model, but (6.3.2),(6.3.3) , (6.3.5),(6.3.7) assume the following:

$$P\{\hat{R_W} = 0\} = \sum_{n=0}^{W-1} \pi_{n,0}, \tag{6.3.8}$$

$$P\{\hat{R_W} = r\} = \sum_{n=0}^{W-1} \sum_{i=1}^{3} \pi_{n,0} z^r_{l,(i,n)}, \quad r \geq 1 \tag{6.3.9}$$

$$z^r_{r-1,(n,i)} = \frac{g_i}{\lambda(1-\delta_{nW})+\lambda^{(r)}+g_i} z^r_{r-1,(n,0)} + \frac{\lambda(1-\delta_{nW})}{\lambda(1-\delta_{nW})+\lambda^{(r)}+g_i} z^r_{r-1,(n+1,i)} \tag{6.3.10}$$

$$\begin{aligned} z^r_{l,(n,i)} &= \frac{g_i}{\lambda(1-\delta_{nW})+\lambda^{(r)}+g_i} z^r_{l,(n,0)} + \frac{\lambda(1-\delta_{nW})}{\lambda(1-\delta_{nW})+\lambda^{(r)}+g_i} z^r_{l,(n+1,i)} \\ &+ \frac{\lambda^{(r)}}{\lambda(1-\delta_{nW})+\lambda^{(r)}+g_i} z^r_{l+1,(n,i)}. \end{aligned} \tag{6.3.11}$$

The finite system (6.3.4),(6.3.6), (6.3.10) and (6.3.11)can be solved numerically. For a detailed discussion on the computation, refer Artalejo and Lopez-Herrero [6].

6.3.5 Waiting time distribution of an orbital call

In this section, we derive the distribution of the time since an incoming customer joins the orbit till it is served. Let us assume that the customer joins the orbit say, as the rth unit, $r > 0$. Now, we consider the Markov process $\{Z(t) = (R(t), J(t)) : t \geq 0\}$, where $R(t)$ is the rank of the said customer at time t and $J(t)$ is the same as defined earlier. The rank R(t) of the call is assumed to be i if it is the ith unit in the orbital queue at time t.

The infinitesimal generator $\bar{Q}$ of $\{Z(t) : t \geq 0\}$ assumes the form

$$\bar{Q} = \left[\begin{array}{c|cccccc} & \overline{r} & \overline{r-1} & \overline{r-2} & \cdots & \overline{1} & \overline{0} \\ \hline \overline{r} & \overline{A_1} & A_2 & 0 & \cdots & 0 & 0 \\ \overline{r-1} & 0 & \overline{A_1} & A_2 & \cdots & 0 & 0 \\ \overline{r-2} & 0 & 0 & \overline{A_1} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \overline{2} & 0 & 0 & 0 & \cdots & A_2 & 0 \\ \overline{1} & 0 & 0 & 0 & \cdots & \overline{A_1} & A_2 \\ \overline{0} & 0 & 0 & 0 & \cdots & 0 & 0 \end{array}\right]$$

where

$$\overline{A_1} = A_1 + \begin{bmatrix} 0 & & & & \\ & \lambda & & & \\ & & \lambda & & \\ & & & \ddots & \\ & & & & \lambda \end{bmatrix}.$$

That is,

$$\bar{Q} = \begin{bmatrix} \overline{T} & \overline{T}^0 \\ 0 & 0 \end{bmatrix},$$

where $\overline{T}$ is the part of the generator corresponding to the transient states $\overline{r}, \overline{r-1}, ..., \overline{1}$.

Hence, the waiting time W of a customer that joins the orbit as the rth unit is a PH

$(\gamma', \overline{T})$ variate with $\gamma' = (\overline{\pi_r}, 0, 0, ..., 0)$, where $\overline{\pi_r} = \pi_r/\pi_r\mathbf{e}$ and π_r is given by eqn (6.2.6). Using the uniformization approach, the distribution function of the waiting time of such a customer can be computed as

$$W(t) = 1 - \sum_{k=0}^{\infty} e^{-ct}\frac{(ct)^k}{k!}\gamma\bar{P}e,$$

where

$$\bar{P} = \frac{1}{c}\overline{T} + I,$$

and c is the maximum of the negative of the diagonal elements of $\overline{T}$.
Also, the average waiting time of such a customer

$$E[W] = \gamma(-\overline{T})^{-1}e = -\overline{\pi_r}(\overline{A_1})^{-1}[I + \sum_{i=1}^{r-1}(-A_2(\overline{A_1})^{-1})^i].$$

6.4 Numerical illustration

In order to illustrate the performance of the system, we are presenting some numerical results based on our theoretical findings. In table 6.1, we exhibit the variation of different probability measures in connection with various server status, while λ varies over the stability region. For the illustration, we take $M = 1$, $\mu = 1.5$, $\lambda^{(r)} = 3.5$, $L = 2$, $\gamma = 8$, $q = [3/4, 1/4]$, $\eta = [1, 0.7]$, $\alpha = 0.6$ and $\beta = 1$. Then the system will be stable for $\lambda < 0.2920$. From table 6.1, it is clear that the server idle probability P_0

λ	0.1	0.15	0.2	0.25	0.29
P_0	0.0893	0.0861	0.0829	0.0797	0.0711
$P_1^{(r)}$	0.0667	0.1000	0.1333	0.1667	0.1933
$P_2^{(o)}$	0.5355	0.5164	0.4973	0.4781	0.4628
$P_3^{(o)}$	0.2550	0.2459	0.2368	0.2277	0.2204
$P_4^{(v)}$	0.0535	0.0516	0.0497	0.0478	0.0463

Table 6.1: Various probability measures

decreases while the inflow rate increases whereas $P_1^{(r)}$, the probability that the server is busy with an incoming call increases as λ increases. Also, it can be seen that the proportion of time the server is busy with either class of outgoing calls as well as on

vacation decrease against the increase in the inflow rate λ.

Table 6.2 displays how mean first passage times vary against the variation of λ over the stable region. Here, we take $M = 2$, $p = [0.25, 0.75]$, $\mu = [1.5, 1.2]$, $\lambda^{(r)} = 3.5$, $p^{(r)} = [0.5, 0.5]$, $L = 2$, $\gamma = 4$, $q = [1/3, 2/3]$, $\eta = [1, 0.7]$, $\alpha = 0.6$ and $\beta = 0.8$. This system will be stable for $\lambda < 0.3963$.

$\lambda = 0.05$	$\tilde{m}$	2.2444	3.0147	3.2073	3.3999	3.8951	3.6888
	$\tilde{m}^{*(0,0)}$	27.3422	27.2026	27.1768	27.1545	27.1124	27.1273
$\lambda = 0.1$	$\tilde{m}$	2.6122	3.5212	3.7485	3.9757	4.5600	4.3166
	$\tilde{m}^{*(0,0)}$	16.2754	16.1033	16.0810	16.0663	13.0616	13.0577
$\lambda = 0.15$	$\tilde{m}$	3.125	4.2333	4.5093	4.7852	5.4948	5.1992
	$\tilde{m}^{*(0,0)}$	13.2632	13.0484	13.0313	13.0270	13.0688	13.0426
$\lambda = 0.20$	$\tilde{m}$	3.9106	5.3083	5.6577	6.0072	6.9057	6.5313
	$\tilde{m}^{*(0,0)}$	12.6533	12.3774	12.3682	12.3789	12.4871	12.4289
$\lambda = 0.25$	$\tilde{m}$	5.2256	7.1180	7.5911	8.0642	9.2808	8.7739
	$\tilde{m}^{*(0,0)}$	13.7400	13.3654	13.3695	13.4048	13.6191	13.5099
$\lambda = 0.3$	$\tilde{m}$	7.9064	10.8069	11.5321	12.2572	14.1219	13.3449
	$\tilde{m}^{*(0,0)}$	17.5614	16.9912	17.0222	17.1066	17.5295	17.3205
$\lambda = 0.35$	$\tilde{m}$	16.3772	22.4627	23.9841	25.5054	29.4175	27.7875
	$\tilde{m}^{*(0,0)}$	31.5465	30.3684	30.4840	30.7210	31.7892	31.2720

Table 6.2: First passage time

It can be seen that the mean first passage time from a level i to $i - 1$ increases significantly while λ increases. This is due to the fact that an increase in λ results in an increase in the probability that the server engages primary arrival so that the mean time taken for reduction in orbit size will be more. Also, it can be seen that for a particular value of λ, the mean first passage time by starting from server idle state is less compared to starting from other server states.

Probability mass function of the number of orbital calls served during system busy period is shown in table 6.3. The same system parameters specified in connection with table 6.2 are considered here. It can be seen that with increase in values of λ, $P\{B = b\}$ increases and for a specific value of λ, probability of serving more and more number of customer decreases.

By taking the same system parameters that have been considered in connection with the previous two tables, along with $W = 20$, we compute the probability distribution of the number of direct primary calls served during system busy period and the results are displayed in table 6.4. Corresponding to increase in values of λ,

	λ=0.1	0.15	0.2	0.25	0.3	0.35
B=1	0.0383	0.0494	0.0571	0.0624	0.0660	0.0681
2	0.0036	0.0063	0.0090	0.0113	0.0132	0.0148
3	$4.01 * 10^{-4}$	$9.67 * 10^{-4}$	0.0017	0.0024	0.0031	0.0038
4	$4.8 * 10^{-5}$	$1.5 * 10^{-4}$	$3.34 * 10^{-4}$	$5.5 * 10^{-4}$	$8.002 * 10^{-4}$	0.0010

Table 6.3: Distribution of the number of orbital calls served during busy period

proportion of the time the busy period expires without serving a direct primary call decreases whereas, probability of serving increasing number of such calls decreases for a particular value of λ.

	λ=0.1	0.15	0.2	0.25	0.3	0.35
A=0	0.7563	0.7107	0.6708	0.6358	0.6051	0.5781
1	0.0400	0.0526	0.0618	0.0686	0.0731	0.0769
2	$5.44 * 10^{-4}$	0.0011	0.0017	0.0023	0.0029	0.0035
3	$7.52 * 10^{-6}$	$2.21 * 10^{-5}$	$4.59 * 10^{-5}$	$7.89 * 10^{-5}$	$1.202 * 10^{-4}$	$1.689 * 10^{-4}$
4	$1.0529 * 10^{-7}$	$4.6 * 10^{-7}$	$1.29 * 10^{-6}$	$2.78 * 10^{-6}$	$5.0759 * 10^{-6}$	$8.28 * 10^{-6}$

Table 6.4: Distribution of the number of incoming calls that are taken into service upon arrival

Table 6.5 exhibits the probability mass function of the number of retrials made by an orbital customer during its stay in the system. Here, we take $M = 1$, $\mu = 1.5$, $\lambda^{(r)} = 3.5$, $L = 1$, $\gamma = 3$, $\eta = 0.8$, $\alpha = 0.7$, $\beta = 1$ and $W = 50$ so that the system will be stable for $\lambda < 0.4922$.

	λ=0.1	0.2	0.3	0.4
r=0	0.1713	0.1590	0.1468	0.1345
1	0.0887	0.0928	0.0968	0.1008
2	0.0215	0.0283	0.0350	0.0416
3	0.0110	0.0163	0.0216	0.0267
4	0.0069	0.0108	0.0147	0.0185

Table 6.5: Distribution of the number of retrials

It is evident from table 6.5 that the proportion of time an incoming call leaves the system, after service, without making a single retrial, decreases with the increase

in values of λ. Also, the proportion of time it goes for at least one retrial increases corresponding to increasing values of λ.

	r=2	3	4
t=1	0.0715	0.0087	0.0006
2	0.2215	0.0612	0.0122
3	0.3750	0.1518	0.0475
4	0.5104	0.2607	0.1073
5	0.6229	0.3729	0.1854
6	0.7132	0.4794	0.2741
7	0.7841	0.5752	0.3663
E[W]	4.8059	7.1110	9.4167

Table 6.6: Waiting time distribution of tagged calls

Finally, the probabilities that the waiting times of incoming calls, that join the orbit at various positions, do not exceed some pre-determined values are computed and shown in table 6.6. Here, we take $\lambda = 0.3924$, $M = 1$, $\mu = 1.5$, $\lambda^{(r)} = 3.5$, $L = 2$, $\gamma = 3.5$,$q = [3/4, 1/4]$, $\eta = [1, 0.7]$, $\alpha = 0.6$, $\beta = 1$. Note that the last row of the table represents average waiting times of such tagged calls.

Summary of the Thesis and Future Work

Here, we discuss the summary of each chapter of this thesis. In chapter 2, we analysed a queuing model to study the characteristics of nodes in a wireless network under the standard BEB contention resolution scheme. Modelled as a QBD process, we used the standard matrix analytic approach to study some of the important characteristics of the model, which can be used for performance evaluation of the actual system. We have used these results to compute the joint system size distribution at nodes in some multi-hop wireless network. Extensive simulation analysis was performed to establish the validity of our theoretical results and some real life data was used for numerical illustration.

In chapter 3, we considered a wireless sensor network model that handles emergency packets. As in chapter 2, we used standard BEB scheme for collision avoidance, but took exponential distribution assumptions for channel busy and idle periods. We derived distribution of time since a packet is ready for transmission till it is successfully transmitted/ timed out, and the probability mass function of the number of collisions experienced by packet.

In chapter 4, we developed a single server multi-class orbit queue with Poisson inputs, constant retrial rates and general service times. Both service times and (exponential) retrial times were assumed to be class-dependent. Different variants of the model by considering presence/absence of outgoing calls were analysed by using regenerative approach and coupling method. Besides, we demonstrated how a combination of the matrix-analytic and regenerative methods simplifies the analysis of a Markovian model of the considered system.

In chapter 5, we carried out performance analysis of the multi-class orbit queue model with Poisson inputs, constant retrial rates, general service times and balking customers. This model could be viewed as a generalisation of the one studied in the previous chapter. We derived necessary stability condition and for the variant with persistent customers, derived sufficient stability condition also. We applied the discrete-event simulation method to validate our analytical results.

In chapter 6, we considered a single server retrial model with two streams of incoming and outgoing calls. Once the server becomes idle, if neither an incoming nor

an outgoing call is being turned up for exponential amount of time, the server goes for vacation and vacation time was assumed to be exponential. Arrivals of incoming calls obey Poisson law. Service times and retrial times were class-dependent exponential variates. Matrix analytic method and regenerative approach were used to analyse the system.

As future work, we can extend the two way communication models and their general version studied here to those having additional features such as service interruption, ambiguity in class determination for arriving customers etc. In this thesis, all the models, in the context of two-way communication systems, are involving single server only. So, these models could be extended to multi-server set up by assuming more general processes such as Markovian Arrival processes (MAP) to catch the correlation between inter-arrival times. Similarly, analysis of a multi-server model with more general phase type distribution assumptions, instead of exponential distribution assumptions, is another potential problem that we can explore. In the case of wireless network systems, faster changes, regarding nature of packets, traffic, routing etc, have been happening on a regular basis. Modelling such kind of systems by incorporating these changes in a realistic manner, and analysing them to get a clear picture of system performance once they would be practically implemented, is another future challenge.

List of Publications

1. S. Dey and T. G. Deepak, "A matrix analytic approach to study the queuing characteristics of nodes in a wireless network", *OPSEARCH*, vol. 56, no. 2, pp. 477-496, June 2019.

2. E. Morozov, A. Rumyantsev, S. Dey and T. G. Deepak, " Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking ", *Performance Evaluation*, vol. 134, Oct. 2019.

3. S. Dey and T. G. Deepak , "An approximation to joint system size distribution at nodes in some multi-hop wireless networks", *Reliability: Theory & Applications*, vol. 14, no. 4, pp. 37–45, Dec. 2019.

4. S. Dey and T.G. Deepak, "Two way communication orbit queues with server vacation", *International Journal of Applied and Computational Mathematics*, vol. 6, no. 4, June 2020.

List of Papers Presented

1. "Regenerative and matrix-analytical performance analysis of multiclass constant retrial rate queues," *European Conference on Queuing Theory* 2018, Jerusalem .

2. "A multi-class multi-orbit retrial model experiencing customer collision," *International Conference on Advances in Applied Probability and Stochastic Processes*, Jan. 2019, Kottayam.

Bibliography

[1] J. Amador and J. R. Artalejo, " The M/G/1 retrial queue: New descriptors of the customer's behaviours ", *Journal of Computational and Applied Mathematics* , vol. 223, no. 1, pp. 15–26, Jan. 2009.

[2] M. A. Ameen and C. S. Hong, " An on-demand emergency packet transmission scheme for wireless body area networks ", *Sensor's (Bessel)*, vol. 15, no. 12, pp. 30584—30616, Dec. 2015.

[3] J. R. Artalejo, " A classified bibliography of research on retrial queues: Progress in 1990-1999 ", *TOP.*, vol. 7, no. 2, pp. 187–211, Dec. 1999.

[4] J. R. Artalejo, " Accessible bibliography on retrial queues ", *Mathematical and Computer Modelling*, vol. 30, pp. 1–6, Aug. 1999.

[5] J. R. Artalejo, A. Gómez-Corral and M. F. Neuts, " Analysis of multiserver queues with constant retrial rate ", *European Journal of Operational Research*, vol. 135, no. 3, pp. 569–581, Dec. 2001.

[6] J. R. Artalejo and M. J. Lopez-Herrero, " On the distribution of the number of retrials ", *Applied Mathematical Modelling*, vol. 31, no. 3, pp. 478-489, Mar. 2007.

[7] J. R. Artalejo and J. A. Resing, " Mean value analysis of single server retrial queues ", *Asia-Pacific Journal of Operational Research*, vol. 27, no. 3, pp. 335–345, 2010.

[8] J. R. Artalejo and T. Phung-Duc, " Markovian retrial queues with two way communication ", *Journal of Industrial & Management Optimization*, vol. 8, no. 4, pp. 781–806, Nov. 2012.

[9] J. R. Artalejo and T. Phung-Duc, " Single server retrial queues with two way communication ", *Applied Mathematical Modelling*, vol. 37, no. 4, pp. 1811–1822, Feb. 2013.

[10] S. Asmussen, O. Nerman and M. Olsson, " Fitting Phase-type distributions via the EM algorithm ", *Scandinavian Journal of Statistics*, vol. 23, no. 4, pp. 419–441, Dec. 1996.

[11] S. Asmussen, *Applied probability and queues* , Springer, New York, 2003.

[12] A. Aissani and T. Phung-Duc, " Profiting the idleness in single server system with orbit-queue ", *Proceedings of the 11th EAI International Conference on Performance Evaluation Methodologies and Tools*, New York, USA, 2017.

[13] K. Avrachenkov, A. Dudin and V. Klimenok, " Retrial queueing model MMAP/M_2/1 with two orbits ", *Multiple Access Communications*, pp. 107–118, 2010.

[14] K. Avrachenkov, E. Morozov, R. Nekrasova and B. Steyaert, " Stability analysis and simulation of N-class retrial system with constant retrial rates and Poisson inputs ", *Asia-Pacific Journal of Operational Research*, vol. 31, no. 2, pp. 149–171, Apr. 2014.

[15] K. Avrachenkov, P. Nain and U. Yechiali, " A retrial system with two input streams and two orbit queues ", *Queueing Systems*, vol. 77, no. 1, pp. 1–31, May. 2014.

[16] K. Avrachenkov, E. Morozov and B. Steyaert, " Sufficient stability conditions for multi-class constant retrial rate systems ", *Queueing Systems*, vol. 82, no. 1-2, pp. 149–171, Feb. 2016.

[17] A. D. Barbour, " Networks of queues and the method of stages ", *Advances in Applied Probability*, vol. 8, no. 3, pp. 584–591, Sep. 1976.

[18] S. Bhulai and G. Koole, " A queueing model for call blending in call centers ", *IEEE Transactions on Automatic Control*, vol. 48, no. 8, pp. 1434-1438, Aug. 2003.

[19] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function ", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, March 2000.

[20] N. Bisnik and A. Abouzeid, " Queuing network models for delay analysis of multihop wireless ad hoc networks ", *Ad Hoc Networks*, vol. 7, no. 1, pp. 79–97, Jan. 2009.

[21] L. Boroumand, R. H. Khokhar, L. A. Bakhtiar and M. Pourvahab, " A review of techniques to resolve the hidden node problem in wireless networks ", *Smart Computing Review*, vol. 2, no. 2, pp. 95–110, April 2012.

[22] B.D. Choi, K.B. Choi and Y.W. Lee, " M/G/1 retrial queueing systems with two types of calls and finite capacity ", *Queueing Systems*, vol. 19, no. 1, pp. 215–229, Mar. 1995.

[23] E. Cinlar, *Introduction to stochatic processes*, Dovar Publication, Mineola, New York, 1975.

[24] D. R. Cox, " Some statistical methods connected with series of events ", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 17, no. 2, pp. 129–157, Jul. 1955.

[25] T. G. Deepak, V. C. Joshua and A. Krishnamoorthy, " Queus with postponed work ", *TOP*, vol. 12, no. 2, pp. 375–398, Dec. 2004.

[26] T. G. Deepak, " On a retrial queueing model with single/batch service and search of customers from the orbit ", *TOP*, vol. 23, no. 2, pp. 493–520, July 2015.

[27] T. G. Deepak, " A queueing network model for delay and throughput analysis in multi-hop Wireless ad hoc networks ", *RT&A*, vol. 12, no. 2, pp. 68–81, June 2017.

[28] A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson and A. .N. Avramidis, " Markov chain models of a telephone call center with call blending ", *Computers & Operations Research*, vol. 34, no. 6, pp. 1616–1645, Jun. 2007.

[29] T. V. Do, " M/M/1 retrial queue with working vacations ", *Acta Informatica*, vol. 47, no. 1, pp. 67-–75, Feb. 2010.

[30] B. Doshi, " Queueing systems with vacations-A survey ", *Queueing Systems*, vol. 1, no. 1, pp. 29–66, June 1986.

[31] A. Dudin and V. Klimenok, " A retrial BMAP/PH/N queueing system with Markov modulated retrials ", *2012 2nd Baltic Congress on Future Internet Communications*, pp. 246–251, April 2012.

[32] A. Dudin and V. Klimenok, " Retrial queue of BMAP/PH/N type with customers balking, impatience and non-persistence ", *2013 Conference on Future Internet Communications (CFIC)*, pp. 1–6, 2013.

[33] T. D. Duy and M. A. Vázquez-Castro, " Efficient communication over cellular networks with network coding in emergency scenarios ", *2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)* , Feb. 2016.

[34] A. K. Erlang, " The theory of probabilities and telephone conversations ", *Nyt Tidsskrift for Matematik B*, vol. 20, Jul. 1909.

[35] A. K. Erlang, " Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges ", *Post Office Electrical Engineer's Journal*, vol. 10, pp. 189–197, 1917.

[36] G. I. Falin, and J. R. Artalejo, ″ Approximations for multiserver queues with balking/retrial discipline ", *Operations-Research-Spektrum*, vol. 17, no. 4, pp. 239–244, Dec. 1995.

[37] G. I. Falin, " Model of coupled switching in presence of recurrent calls ", *Engineering Cybernetics Review*, vol. 17, no. ,pp. 53–59, 1979.

[38] G. I. Falin, " A survey of retrial queues A survey of retrial queues ", *Queueing Systems*, vol. 7, no. 2, pp. 127–167, Jun. 1990.

[39] J. Falin and J. G. C.Templeton, *Retrial queues*, Chapman and Hall, London, 1997.

[40] D. Gross, J. Shortle, J. Thompson and C. Harris, *Fundamentals of queuing theory*, Wiley Series in Probability and Statistics, 2008.

[41] Q. He, *Fundamentals of matrix-analytic methods*, Philadelphia, 2014.

[42] F. P. Kelly, " Networks of queues ", *Advances in Applied Probability*, vol. 8, no. 2, pp. 416–432, June. 1976.

[43] D. G. Kendall, " Some problems in the theory of queues", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 13, no. 2, pp. 151–173, July 1951.

[44] D. G. Kendall, " Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain ", *The Annals of Mathematical Statistics*, vol. 24, no. 3, pp. 338–354, 1953.

[45] A. Y. Khintchine, " The mathematical theory of stationary queues ", *Matematicheskii Sbornik* , vol. 39, no. 4, pp. 73–84, 1932.

[46] A. Y. Khintchine, " Mathematical methods in the theory of queueing. ", New York, Hafner Pub. Co., 1960.

[47] J. Kim and B. Kim, " A survey of retrial queueing systems ", *Annals of Operations Research*, vol. 247, no. 1, pp. 3–36, Dec. 2016.

[48] H. Kobayashi, " Application of the diffusion approximation to queuing networks: Equilibrium queue distributions ", *Journal of the ACM*, vol. 21, no. 2, pp. 316–328, April 1974.

[49] A. N. Kolmogorov, " On the problems of expectation ", *Math. Soc.*, vol. 38, no. 1-2, pp. 101–106, 1931.

[50] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modelling*, Philadelphia, 1999.

[51] T. Li, Z. Wang and Z. Liu, " Geo/Geo/1 retrial queue with working vacations and vacation interruption ", *Journal of Applied Mathematics and Computing* , vol. 39, no. 1-2, pp. 131—143, June. 2012.

[52] Z. Liu and Y. Song, " Geo/Geo/1 retrial queue with non-persistent customers and working vacations ", *Journal of Applied Mathematics and Computing* , vol. 42, no. 1-2, pp. 103—115, Jul. 2013.

[53] M. Li, J. Liu, Q. Shen and B. Yuan, " QS-PS: A new approach for emergency packet delivery in WBAN ", *International Conference on Wireless Algorithms, Systems, and Applications*, vol. 8491, pp. 589-600, 2014.

[54] D. M. Lucantoni, " The BMAP/G/1 queue: A tutorial ", *In: Donatiello L., Nelson R. (eds) Performance Evaluation of Computer and Communication Systems. Performance 1993, SIGMETRICS 1993. Lecture Notes in Computer Science*, vol. 729. Springer, Berlin, Heidelberg, 1993.

[55] E. C. Molina, " Application of the Theory of probability to telephone trunking problems ", *Bell System Technical Journal*, vol. 6, no. 3, pp. 461–494, Jul. 1927.

[56] E. Morozov, " The tightness in the ergodic analysis of regenerative queuing processes ", *Queueing Systems*, vol. 27, pp. 179–203, 1997.

[57] E. Morozov, " A multiserver retrial queue: regenerative stability analysis ", *Queueing Systems*, vol. 56, no. 3 pp. 157–168, Aug. 2007.

[58] E. Morozov and I. Dimitriou, " Stability analysis of a multiclass retrial system with coupled orbit queues ", *Computer Performance Engineering: 14th European Workshop, EPEW 2017*, Springer International Publishing, pp. 85–98, Sep. 2017.

[59] E. Morozov and R. Delgado, " Stability analysis of regenerative queueing systems ", *Automation and Remote Control*, vol. 70, no. 12, pp. 1977–1991, Dec. 2009.

[60] E. Morozov and T. Phung-Duc, " Stability analysis of a multiclass retrial system with classical retrial policy ", *Performance Evaluation*, vol. 112, pp. 15–26, Jun. 2017.

[61] E. Morozov and T. Phung-Duc, " Regenerative analysis of two-way communication orbit-queue with general service time ", *Queueing Theory and Network Applications*, Springer International Publishing, pp. 22–32, 2018.

[62] M. F. Neuts, *Probability distribitions of Phase type.*, In Liber Amicorum Prof. Emeritus H. Florin; Department of Mathematics. University of Louvain, Belgium, pp. 173-206, 1975.

[63] M. F. Neuts, *Matrix-geometric solutions in stochastic models: An algorithmic approach*, Baltimore: The Johns Hopkins University Press, 1981 (Reprinted by Dover Publications, 1994).

[64] F. Pollaczek, " Über eine Aufgabe der Wahrscheinlichkeitstheorie. ", *Mathematische Zeitschrift* , vol. 32, pp. 729–750, Dec. 1930.

[65] T. Phung-Duc, W. Rogiest, Y. Takahashi and H. Bruneel, " Retrial queues with balanced call blending: analysis of single-server and multiserver model ", *Annals of Operations Research*, vol. 239, no. 2, pp. 429–449, Apr. 2016.

[66] Y. Shin and D. Moon, " M/M/c retrial queue with multiclass of customers ", *Queueing Systems*, vol. 16, no. 4, pp. 931–949, Dec. 2014.

[67] W. L. Smith, " Regenerative stochastic processes ", *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 232, no. 1188, pp. 6, Oct. 1955.

[68] Y. Song, Z. Liu and Y. Zhao, " Exact tail asymptotics: revisit of a retrial queue with two input streams and two orbits ", *Annals of Operations Research*, vol. 247, no. 1, pp. 97–120, Dec. 2016.

[69] H. Sakurai and T. Phung-Duc, " Two-way communication retrial queues with multiple types of outgoing calls ", *TOP*, vol. 23, no. 2, pp. 466–492, July. 2015.

[70] F. Tobagi and L. Kleinrock, " Packet switching in radio channels: part 2 - The hidden terminal problem in carrier sense multiple-access and the busy - tone solution ", *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1417–1433, Dec. 1975.

[71] H. Takagi, *Queueing Analysis, A foundation of performance evaluation, volume 1: vacation and priority systems, part 1*, North-Holland, Amsterdam , 1991.

[72] N. Tian and Z. G. Zhang, *Vacation queueing models: theory and applications*, Springer, New York, 2006.

[73] Y. Wang, M. C. Vuran and S. Goddard, " Cross-layer analysis of the end-to-end delay distribution in wireless sensor networks ", *IEEE/ ACM Transactions on Networking*, vol. 20, no. 1, pp. 305–318, Feb. 2012.

[74] *Wireless LAN Medium Access Control (MAC) and Physical Layer(PHY) Specifications*, IEEE standards 802.11, 1997.

[75] T. Yang and J. G. C. Templeton, " A survey on retrial queues ", *Queuing Systems*, vol. 2, no. 3, pp. 201–233, Nov. 1987.

[76] J. Zhou and K. Mitchell, " A scalable delay based analytical framework for CSMA/CA wireless mesh networks ", *Computer Networks*, vol. 54, no. 2, pp. 304–318, Feb. 2010.

[77] R Core Team, *R: A language and environment for statistical computing* , R Foundation for Statistical Computing, Vienna, Austria , 2018.