Generalized Pollaczek-khinchin Formula for Queueing Systems with Markov Modulated Services Rates

HUANG, Liang

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Information Engineering

The Chinese University of Hong Kong
August 2013

Abstract

Abstract of thesis entitled:

Generalized Pollaczek-khinchin Formula for Queueing Systems with Markov Modulated Services Rates

Submitted by HUANG, Liang

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in August 2013

This dissertation is aimed to study the queueing behavior of communication systems with time-varying service rates. The system may change its service rate several times while serving a customer subject to the condition of external environment. The time-varying server is modeled as a Markov modulated service process (MMSP) and the communication system is modeled as an $M/MMSP/1$ queue. The existing performance analyses of Markov modulated service process are almost all based on the on the matrix-geometric method, which provides little physical insights for system design. By contrast, we focus on deriving closed-form analytic expressions with physical interpretations in terms of system parameters of interest. Our main contribution is to derive the generalized Pollaczek-Khinchin (P-K) formula of the

$M/MMSP/1$ queue from the start service probability to explore the impact of channel state transitions on the queueing behavior of the system. This generalized P-K formula reveals that the performance of $M/MMSP/1$ queue can be fully characterized by a newly defined system parameter, called state transition factor β, which clearly explains the reason that the system with slow state transition rate owns a larger delay for the same system/channel capacity. In the extreme case when the state transition factor β approaches 0, we show that the system under consideration can be approximately modeled as an $M/G/1$ queue. Both the two-state and the finite-state $M/MMSP/1$ queues are studied in details.

The wireless fading channels with finite input buffer, Poisson arrivals, and two-state Markov modulated service processes (MMSP) are modeled as $M/MMSP/1/K$ queues. We first obtain the buffer overflow probability and its large-deviation approximation with an exact asymptotic constant from generating functions. The state transition factor β is given in a simple expression and the start-service probabilities are obtained in closed-form expressions. The generalized P-K formula is derived based on finite buffer capacity which gives the exact value of the mean waiting time. We use a Type I Hybrid ARQ system with a fixed data-rate as a running example to illustrate our results with two-state MMSP.

We then extend our generalized P-K formula for Markov channels with two states to general Markov modulated service process with finite states. For a special three-state Markov channel with no service rate in one state, we show that a simple closed-form expression of the state transition factor is available. For a N-state MMSP, it is impossible to obtain closed-from expression for the start-service probability and we propose two approximations: Linear Approximation for small N and CDF Approximation for large N. The approximate generalized P-K formula can well predict the mean queue length as verified by simulation results through a general queueing model for peer-to-peer file-sharing systems.

摘要

本文旨在研究通信系統在服務率隨時間變化下的排隊行為。在服務一個顧客的過程中，系統的服務率可能會隨著外部環境的變化而改變。我們建立一個馬爾科夫調製服務過程（MMSP）模型來表示這種隨著時間變化速率的服務台，同時用一個$M/MMSP/1$排隊模型描述整個通信系統。現存的馬爾科夫調製服務過程的性能分析都是基於矩陣幾何方法，不能清晰地解釋系統設計時中各個參數的物理意義。相比之下，我們專注於推導能體現各個系統參數物理意義的閉式表達式。我們的主要貢獻是基於起始服務概率（start-service probability），針對$M/MMSP/1$排隊模型推導出廣義的 Pollaczek-Khinchin (P-K) 公式，從而探究服務速率變化對排隊系統性能的影響。根據廣義的 P-K 公式可知，$M/MMSP/1$排隊模型的性能完全由本文中新定義的狀態轉換因子（state transition factor）β所決定，並且解釋了爲什麼在相同的系統或者信道容量下服務狀態轉換慢的系統的延遲更大。我們證明當狀態轉換因子β極限趨近於零時，$M/MMSP/1$排隊模型近似等價於$M/G/1$排隊模型。我們通過分析具有兩個服務狀態的無線混合自動重傳請求協議（Hybrid ARQ）和具有多個服務狀態的點對點傳輸系統（P2P）作為示例以展示我們的研究成果。

我們用$M/MMSP/1/K$排隊模型來描述這樣的無線衰落信道：具有有限輸入緩存區，顧客到達過程服從泊松分佈，具有兩個服務且屬於馬爾科夫調製

服務過程（MMSP）。我們首先通過矩母函數方法得到緩存溢出概率和其相應的帶確切漸進常數的大偏差近似。在當前系統下，狀態轉換因子β是一個簡單的表達式，同時可以求得起始服務概率的閉式表達式。在有限緩存的條件下，我們直接推導出廣義的P-K公式，準確地描述顧客的平均等待時間。我們列舉了一個具有固定傳輸速率的第一類混合自動重傳請求(Type I Hybrid ARQ)系統，以闡釋上述關於兩個服務狀態的結果。

我們將上述在兩個服務狀態的馬爾科夫調製服務過程的條件下獲得的廣義P-K公式推廣到了多個服務狀態。如果一個三個服務狀態的馬爾科夫信道中的某一個服務率是零，那麼其相應的狀態轉換因子β可以寫成一個簡單的表達式。在多個服務狀態下，不可能獲得起始服務概率的閉式表達式。因此，我們提供了兩種近似方法分別針對服務狀態較少和較多的情況：線性近似法和累積分佈函數法。通過分析一個點對點文件共享系統的一般排隊模型，仿真結果顯示這種近似的廣義P-K公式可以很好的預測平均隊列長度。

Acknowledgement

“Ah, ’All things come to those who wait.’
They come, but often come too late.”
From Lady Mary M. Currie: *Tout Vient à Qui Sait Attendre* (1872)
Quoted by L. Kleinrock: *Queueing Systems* (1975)

I would like to express my deepest gratitude towards my doctoral advisor, Professor Tony T. Lee, for his guidance, support as well as the invaluable advice in the past four years. I do feel what I have learned from him is not only the methodical approach to problems, but also his insights and attitude of research and life. I would also like to acknowledge his contributions to the technical content in this research work.

I would like to thank all of the members in Broadband Communications Laboratory, including Angel Wong, Dongjie Yin, Yujie Wan, Hao Guan, Dr. Fangfang Yan and Qingqi Shi. Without them these graduate school life would lack of much wonderful and enjoyable memory.

Additionally, I also sincerely express my thanks to my best friends: Suzhi Bi, Xiaokun Li, Lei Jiang, Mingzhen Mo, Yi Xu, Mingdong Zhou, Yulin Deng, Ke Liu, and Chunfeng Yang. They gave me pleasant surroundings outside the academic life.

Finally, I would like to mention that all this would not have been possible without the unconditional love and continuous support from my family.

To *mom*.

Contents

List of Figures

List of Abbreviations

α	Asymptotic constant of large-deviation approximation
β	State transition factor
$\hat{\boldsymbol{Q}}$	State transition matrix
$\boldsymbol{Q}$	Infinitesimal generator matrix
η_j	The extreme of $\hat{\pi}_j$ when the arrival rate goes to infinity
γ	Ratio of the cumulative distribution function $F(x)$ when $x = \lambda_c$ and $x = \rho_s$
$\hat{\mu}$	Channel capacity
$\hat{\pi}_j$	Start-service probability that a packet's service starts in channel state j
$\hat{\pi}_j(i, m)$	The probability that the m^{th} packet starts its service in channel state j, given that an arrival seeing i packets in the system
λ	Arrival rate of customers
λ'	Net arrival rate without those blocked customers due to buffer overflow
λ_c	Arrival rate of jobs

λ_s	Arrival rate of servers
μ_j	Service rate in channel state j
μ_s	Departure rate of servers
π_j	Steady-state probability that the channel is in state j
ρ	Server utilization
ρ_c	Ratio of customer arrival rate λ_c and service rate μ_c
ρ_s	Ratio of service arrival rate λ_s and departure rate μ_s
ξ	Eigenvalues of state transition matrix $\hat{\boldsymbol{Q}}$
$E[T]$	Mean service time
$E[T^2]$	Second moment of service time
$E[T_j]$	First conditional moment of service time
$E[T_j^2]$	Second conditional moment of service time
$F(x)$	Cumulative distribution function of the steady-state probability π_j
f_j	Total transition rate from channel state j to other states
$f_{j,j'}$	Transition rate from channel state j to j'
$G(z)$	System generating function
$G_j(z)$	Conditional generating function
L	Mean queue length
$n_c(t)$	Number of jobs in the queue at time t
$n_s(t)$	Number of customers in the queue at time t

P_i	Steady-state probability that there are i packets in the system
P_K	Buffer overflow probability
$p_{i,j}$	Steady-state probability that there are i packets in the system when the channel is in state j
Q	Number of packets in the system seen by an arrival
Q_c	Number of customers in the system seen by an arrival
R	Mean residual service time
R_i	Residual service time of an arrival seeing i packets in the system
$V(t)$	State of channel at time t
W	Mean waiting time
W_i	Waiting time of an arrival seeing i packets in the system
$X(t)$	Number of customers in the system at time t
X_m	Service time of the m^{th} packet in the queue

Chapter 1

Introduction

1.1 Statement of the Problem

With the development of communication networks, systems with time-varying service rates have received more and more attention in the field of wireless communication systems and sharing systems with limited resource.

1.1.1 Wireless Channels

Most newly emerging real-time telecommunication services are deployed over wireless networks. The challenge of providing stringent quality of service (QoS) guarantees is how to cope with the time-varying service rate that is subject to the radio propagation characteristics of wireless channels [23]. The impact of variations of channel service rate on the queueing behavior of input buffer is vital to the design of wireless system for delay-sensitive traffic [4, 54].

The variations of the fading channels are commonly estimated by the normalized Doppler frequency, which is the product of the maximum Doppler frequency f_D and the symbol period. It is widely known that slower fading channels own larger loss probabilities or longer delay, which can only be intuitively explained by the normalized Doppler frequency [4,29,46,64]. There is no rigorous analysis in current literature that explores the effect of slow/fast fading on the delay performance of Markov channels.

The performance of wireless fading channels is widely studied by using the Markov chain method invented and pioneered by Wang and Moayeri in [50], which explicitly established the link between the physical parameters of wireless channels and the states of the finite-state Markov channel (FSMC). In particular, each channel state corresponds to a range of the received signal-to-noise ratio (SNR), which, in return, determined a constant error probability in that state. The state transition rates are calculated from the level-crossing rate at the physical layer, which are linear functions of the Doppler frequency f_D. With the help of FSMC modeling, system performance metrics such as packet error probabilities and throughput of the system can be analyzed and improved. For delay-sensitive systems, the analyses of packet queueing delay and buffer overflow probabilities in the literature are almost all based on the matrix-geometric method [35], which was developed in the 1980s. However, those results expressed as functions of matrices provide little physical insights for system design, and we do not know the range that they are bounded.

1.1.2 Markov modulated service processes in different fields

Besides in the wireless communications systems, systems with time-varying service rates are studied in different fields in the literature, by modeling the service process as a finite-state Markov chain:

A general P2P queueing model is proposed by Li et al. [27], where the system has independent Poisson job and server arrivals (with rates λ_c and λ_s respectively), and independent exponentially-distributed job service time and server life time (with rates μ_c per server and μ_s respectively). Li et al. [27] prove the P2P system stability conditional is $\lambda_c/\mu_c < \lambda_s/\mu_s$, and observe that systems with higher server dynamics lead to lower waiting time.

Perel and Yechiali [37] analyze a similar queueing systems of Markov models with applications in computer networks, where the system is comprised of two connected $M/M/-/-$ type queues with customers of one queue, Q_s, act as servers for the other queue, Q_c. Q_s operates as a finite buffer $M/M/1/N$ system and Q_c operates as an infinite-buffer, single-server $M/M/1/\infty$ queue with Poisson arrival rate λ_c and dynamically changing service rate $\mu_c L_s$, where L_s denotes the number of customers in Q_s. They derive the generating functions of the systems's steady state probabilities and calculate numerically the mean total number of customers in the second queue.

Another example is the parallel processes [32], where a Central Processor Unit (CPU) runs multiple processors in parallel. A software agent submits tasks to the CPU continuously throughout the day according to a Poisson process and each task has $\exp(\mu)$ work in it. If there is only one process running on the CPU, it receives all the CPU speed. However if there are few other processes running at the same time, each of those processes shares a fraction of the CPU speed. Then the service rates vary according to an external environment process which is modeled as a Continuous Time Markov Chain (CTMC). The author analyzed the conditional moments of service time and obtained the queue length and delay using Matrix Geometric Method (MGM).

The Markov modulated service rates model is also used to study the service systems with human servers [61], where employee learning and turnover cause the sequence of service-time distribution to exhibit systematic non-stationary. A new employee may lean and advance to a higher skill level with a pre-specified probability; a skilled employee may also turn over with another probability and the position is usually replaced by another new person with lower skill level. Such a system is modeled as a $M/MMPP/1$ process and obtained a complete characterization of the system's behavior using a matrix difference equation approach when considering only two states.

In the field of transportation, If an incident occurs on a road segment all the vehicles on the road have to lower their speed. Consider a section of a road subject to incidents. The space occupied by an individual vehicle on the road segment represents one queueing server, which starts its service as soon as a vehicle joins the link and carries the service (the act of traveling) until the end of the link is reached. A two-mile roadway section contains hundreds or thousands of such servers, and Baykal-Gursoy and Xiao [3] consider an $M/M/\infty$ queueing system subject to random interruptions of exponentially distributed durations.

A system with Markov modulated service rates is also a Quasi-birth-death (QBD) process. All the previous system performance analysis concerning queueing length and delay is carried out by applying MGM. MGM is so powerful to solve all those models. However, the computing complexity considerably increases when the number of service states is large. What's more, only numerical results can be obtained which implies little physical insights on the relationships between the system performance metrics and the system parameters.

1.1.3 Purpose of the Study

In this dissertation, we study the queueing performance of communication systems with Markov modulated service rates. We focus on deriving analytic expressions with clear physical interpretations in terms of system parameters of interest, that is, the closed-form physical laws that govern the system behavior. The results we obtained based on Markov channels could also be applied to other related applications with Markov modulated service time in different fields.

1.2 Our Methodologies

In this dissertation, we study on the performance analysis of Markov channels by modeling the channel as a Markov modulated service process (MMSP). We consider Poisson arrivals, and the channel systems are modeled as $M/MMSP/1$ queues. We introduce the sate transition factor β, which indicates how fast the channel state changes in comparison with service rates, to characterize the performance of the *M/MMSP/1* queue. From the generalized P-K formula for the *M/MMSP/1* queue derived by using the start-service probability, we show that the queueing delay is very sensitive to this state transition factor β. When this factor is close to 1, the queueing delay of the wireless channel becomes extremely large. On the other hand, the performance of the *M/MMSP/1* queueing system can be approximated by an *M/G/1* queue when this factor β is close to 0.

The state transition factor β is derived by studying the channel state transition progress during one service time of a random packet. Purdue [39] analyzed the similar process by studying the busy period of an $M/M/1$ queue embedded in an N-state irreducible continuous-time Markov chain. He obtained a so called *busy-period matrix*, corresponding to the state transition matrix $\hat{\boldsymbol{Q}}$ in this dissertation, from which he derived the stability condition of the queueing system by obtaining the extreme value of the start-service probability when the arrival rate goes to infinity. We define the state transition factor β from this matrix $\hat{\boldsymbol{Q}}$ and show it is essential to the queue length of the *M/MMSP/1* queue by deriving its generalized P-K formula.

The derivation of the generalized P-K formula is based on the residual service time similar to the method described in pages 141-144 of [5] for proving the traditional P-K formula for the $M/G/1$ queue. We obtain the mean waiting time from the start-service probabilities and the state transition factor, by invoking to the conditional moments of service time [32]. Mahabhashyam and Gautam [32] derived the expressions of conditional moments of service time of queueing systems with Markov modulated service process to analyze the first and second moments of service time. He also studied the start-service probability under two extreme cases when the arrival rate goes to 0 and infinity.

We first analyze a Markov channel with two-state. The wireless fading channels with finite input buffer, Poisson arrivals and two-state Markov modulated service processes (MMSP) are modeled as $M/MMSP/1/K$ queues. We derive the closed-from expressions of buffer overflow probability and queueing delay from conditional generating functions. A simple expression of state transition factor β is derived and closed-form expressions of start-service probabilities are provided. The closed-form expression of the generalized P-K formula is derived based on finite buffer capacity. We use the Type I Hybrid ARQ system with a fixed data-rate as an example to illustrate our results for two-state channels.

We then extend our generalized P-K formula for Markov channels with two states to general Markov modulated service process with finite states. For a special three-state Markov channel with no service rate in one state, we show that a simple closed-form expression of the state transition factor is available. We provide a Linear Approximation method to approximately calculate the start service probability. For Markov channels with large number of states, we propose a CDF Approximation method to approximately calculate the start service probability. We take a P2P system with the $M/M/(M/M)$ model defined in [27] as an example to illustrate our results for Markov channels with large states.

1.3 Contributions

In this dissertation, we consider the wireless channel as a Markov modulated service process (MMSP). Assuming Poisson arrivals and exponential service time in each channel state, we obtain the following results:

1. We derive a closed-form expression of buffer overflow probability of the *M/MMSP/1/K* queue with two states from conditional generating functions. We also provide the exact expression of the asymptotic constant for the large-deviation approximation of buffer overflow probability.

2. We define a key parameter, called *state transition factor*, to completely determine the start-service probability first introduced in Mahabhashyam and Gautam's approach [32]. The state transition factor indicates how fast the channel state changes with respect to service rate. Based on this generalized start-service probability, we obtain the closed-form expressions of the first and second moments of service time and mean delay. For a Markov channel with N states, we provide an Linear Approximation method for small N and a CDF Approximation method for large N to approximately calculate the start service probability.

3. We derive a generalized P-K formula for $M/MMSP/1$ from conditional moments of service time and start-service probability which reveals that the queueing delay is highly related to the state transition factor. For wireless channels with the same channel capacity, we show that the mean delay of a channel with a slow state transition rate is longer than the one with fast channel state transition rate, simply because the former owns a larger state transition factor, say close to 1, than the latter. Thus, this state transition factor should be reduced as much as possible for delay-sensitive wireless systems. Furthermore, we show that the $M/MMSP/1$ can be approximated by an $M/G/1$ queue with the same first and second moments of service time when this factor approaches 0.

1.4 Dissertation Overview

The dissertation is organized as follows:

Chapter 2 provides an extensive review of the related research works in the literature, including *a*) Markov channels, *b*) Queues with Markov modulated service rates, *c*) Delay analysis on Hybrid ARQ systems, and *d*) Delay analysis on Peer-to-peer systems.

In Chapter 3, we study the queueing performance of Markov channels with two states. We first derive the closed-from expressions of buffer overflow probability and queueing delay from conditional generating functions. A simple expression of state transition factor β is derived and closed-form expressions of start-service probabilities are provided. The closed-form expression of the generalized P-K formula method is derived based on finite buffer capacity and the impact of the state transition factor on queue length is discussed.

In Chapter 4, we extend our generalized P-K formula method derived in Chapter 3 to Markov channel finite states. We define state transition factor for finite-state Markov channel from state transition matrix $\hat{\boldsymbol{Q}}$. The results are illustrated in details through a three-state Markov channel model and a finite-state P2P model. We provide two approximation methods, Linear Approximation and CDF Approximation, to approximately calculate the start service probability for small and large number of channel states respectively. The impact of the state transition factor on queue length of $M/MMSP/1$ with finite states is discussed and verified by simulations.

Chapter 5 will give the summary of this dissertation, together with several further research directions.

Chapter 2

Review of Related Literature

In this Chapter, we provide an extensive review of the literature and research related to the queueing analysis over Markov channels. The chapter will be divided into four sections that include *a*) Markov channels, *b*) Queues with Markov modulated service rates, *c*) Delay analysis on Hybrid ARQ systems, and *d*) Delay analysis on Peer-to-peer systems.

2.1 Markov Channels

The study of communication channels dates back to the work by Shannon in [47], where the channel is defined as "merely the medium used to transmit the signal from transmitter to receiver" and the capacity of a channel is defined relating to information entropy. In 1960, Gilbert [17] introduced a Markov chain with *Good* and *Bad* states to model a burst-noise channel, where each channel state is associated with a discrete memoryless channel and is statistically independent of the channel inputs. In 1963, Elliott studied the error rate performance of error correcting and error detecting codes over Gilbert's model in [11]. In the original version of Gilbert's model, the transmission over channel with *Good* state is error-free; Elliott later modified the model to allow error probabilities in both states and the error probability in the *Good* state is smaller than that in the *Bad* state [44]. The Gilbert-Elliott channel, which models the channel by a two-state Markov chain, is widely used for modeling wireless fading channels over the past 60 years for its simplicity.

Meanwhile, researchers are trying to extend the Gilbert-Elliott model from two states to finite states since 1960s, so as to model much complicated channels. Fritchman [14] studied a N-state Markov channel, which is partitioned into a group of N_G *Good* states and $N - N_G$ *Bad* states in 1967. For the special Markov channel with only one *Bad* state, Fritchman well derived the channel error statistics and showed the Gilbert-Elliott as an extreme case. Gallager [16] further explored the information theoretical aspect of finite-state Markov channels (FSMCs) in 1968, by providing standard definitions, coding theorems and error exponents. The applications of these Markov channels were limited to model error bursts in digital wireline circuits or wireless links between fixed stations until the commercial success of digital cellular networks in the 1990s, after which the demand for modeling of fading channels was arising. Wang and Moayeri [50] explicitly established the fading channel models by using finite-state Markov chains in 1995. They explicitly established the link between the physical parameters of wireless channels and the states of the FSMC. In particular, each channel state corresponds to a range of the received signal-to-noise ratio (SNR), which, in return, determined a constant error probability in that state. The state transition rates are calculated from the level-crossing rate at the physical layer, which are linear functions of the Doppler frequency f_D. An improved method to partition the received SNR into finite states for Rayleigh fading channels was proposed in [60], and the level-crossing rate for general Nakagami fading channels was analyzed in [22]. Higher order Markov modeling of Rician fading channels was investigated in [38]. The packet transmission process over Rayleigh fading channel is adequately modeled three-state Markov model with one *Good* state and two *Bad* states [64]. Due to their simplicity, the proposed Markov chain modeling [50] and its variations [22, 60] are still used by researchers today to determine model parameters and to analyze system performance.

2.2 Queues with Markov Modulated Service Rates

The queueing analysis of Markov modulated service rates traces back to the early work of queueing problems where the service of a customer may breakdown and resume according to different rules [51]. By considering Poisson arrivals and a single server with service rate varying between μ and 0, the moments of the queue length distribution are found by a generating function approach in [51]. A multi-server queue where each server may be down independently of the others for an exponential amount of time is analyzed in [34], where the explicit from of the moment generating function is obtained when there are two servers. Eisen and Tainiter [10] studied a single-server queue with the arrival and service rates alternate according to two external environment states, and each state corresponding to an arrival and service rate in 1963. Analytic expressions are obtained for the generating functions, the mean queue length, and the mean waiting time. Apparently unaware of the work done by Eisen and Tainiter, Yechiali and Naor [57] also obtained similar steady-state results for the same model in 1971. The work is generalized to a queue whose arrival and service rates are changing according to a continuous-time N-state Markov chain in [56]. Purdue [39] studied the busy period of this queue by defining $busy-periodmatrix$ and obtained its generalized equilibrium conditions with Yechiali's result [56] as a special case.

Neuts [36] generalized the service distribution of queueing model in [57] from exponential distribution to general distribution, where he assumes that the service state only changes only at the beginning of the service. There are followed analysis on queues with general distributed service time which depends on the underlining finite-state Markov chain [6, 41, 49].

In general, the queueing models with Markov modulated service rates can be represented as quasi-birth-and-death (QBD) processes [15] and their analysis can be solved by the matrix-geometric method [35]. However, the matrix-form solution requires complicated computations but provides little physical insights for practical system operation. A novel approach based on conditional moments of service time is proposed by Mahabhashyam and Gautam in [32], in which the moments of service time are evaluated by conditioning on the start-service state of the server. However, the analysis in [32] is incomplete because the start-service probability is only available for two extreme cases, packet arrival rate approaches zero or infinity, which are not in the region of interest in the practical operation of wireless channels.

2.3 Delay Analysis on Hybrid ARQ Systems

Automatic Repeat reQuest (ARQ) and Forward Error Correction (FEC) are commonly used to improve the quality of digital data delivery over wireless channels. ARQ detects error bits and requests a retransmission of the current packet, while FEC corrects error bits at the receiver with additional redundant bits. The former guarantees transmission reliability and the latter is more suitable for delay-sensitive applications. However, ARQ technology may cost large delay for multiple retransmissions for the same packet. On the other hand, FEC consumes excessive bandwidth, especially under good channel condition, in order to improve transmission reliability over time-varying channel.

The Hybrid ARQ scheme designed for delay-sensitive wireless systems comprises a stereotypical class of Markov fading channels [2, 12, 19, 24–26, 43, 46]. It combines conventional ARQ with Forward Error Correction (FEC) to reduce the number of retransmissions and improve the throughput by correcting some error bits at the receiver [28]. The receiver first corrects received packets with FEC bits and requests a retransmission if there are remaining errors. There are two types of Hybrid ARQ systems. For the Type I Hybrid ARQ scheme, the transmitter retransmits the same packet. For the Type II Hybrid ARQ scheme, the transmitter sends additional redundancy bits, along with the erroneous packets previously received, to help decode the original message [31].

The packet delay analyses of different Hybrid ARQ schemes have been well studied in the literature. Packet transmission delay of SR ARQ systems with Markov channels was derived based on heavy traffic assumption in [43]. For a Type I Hybrid ARQ wireless system with infinite buffer capacity, the packet delay and buffer overflow probability were derived from tail distribution analysis by Kim and Krunz [24, 25]. The queueing behavior of a Type II Hybrid ARQ system over a Markov channel was analyzed in [46] with a resort to matrix-geometric method. Assuming that packets are discarded after limited timeslots instead of buffer overflow, the packet loss probability of a Type I Hybrid ARQ wireless system with batch arrivals was obtained in [26]. A wireless channel with Type II Hybrid ARQ was modeled as a three-state continuous-time Markov process in [19], which provides average delay analysis from the tail asymptote buffer overflow probabilities. Various approximations of buffer overflow probability were obtained in previous work based on the theory of large deviations. Due to the infinite buffer assumption, however, the asymptotic constant in these approximate overflow probabilities has never been exactly determined. Under the finite input buffer assumption, an approximation of the overflow probability of a Hybrid ARQ system was obtained from the first and second moments of service time of an *M/G/1/K* queue in [12], while we directly calculated this probability from the generating functions of $M/MMSP/1/K$ queue in [21]. The analyses in [2, 24–26, 43] assume that the channel condition does not change during the transmission of a packet. However, for fading channels, channel variations are independent of the packet transmission. In this dissertation, our analysis allows channel variations during the packet's transmission and reveals how the channel variations (slow or fast) affecting the delay performance.

2.4 Delay Analysis on Peer-to-peer Systems

Typical file sharing systems, peer-to-peer (P2P), are widely used to for distribution of resources over Internet, including video downloads, online storage and media streaming. In a P2P network, a connected customer operates as a user downloading a file and as a server uploading the file at the same time.

Stochastic models have been used to analyze P2P file sharing systems where the number of servers is greatly correlated to the number of jobs. In [55], the service capacity of a P2P system is modeled in two regimes, the transient phase by a brunching process and the steady state by a Markov model. A fluid model is used in [40] to characterize the performance and efficiency of BitTorrent like networks in terms of average downloading time. In [13], the fluid model is supplemented obtain the analytical expressions of system performance with higher accuracy. [48] studied the population dynamics of system by a deterministic fluid model and adopted a more detailed Markov chain to estimate the life time of a P2P file sharing system. The Markov model approximates to $M/M/\infty$ when the mean service times are very small.

With the development of P2P technology, such as streaming, the server arriving process tends to a random process and less correlated to the job dynamics. Idle Internet resources are leveraged to act as additional servers provide a scalable solution to P2P Video-on-Demand (VOD) systems [58]. A new View-Upload Decoupling

(VUD) design for multi-channel P2P streaming systems is proposed in [53], where a user might be assigned to one or more unwatched channels to contribute the upload bandwidth. [27] developed queueing models for P2P service systems where the server dynamics may or may not correlate to the job dynamics.

Chapter 3

Generalized Pollaczek-Khinchin Formula for the *M/MMSP/1/K* Queue — Two-state

Although the two-state Markov model has been widely used in [8, 21, 24–26, 33, 63], the analysis of wireless channels is still incomprehensible because of the dependent service time distribution. In this chapter, we use generating function methods to compute the buffer overflow probability and introduce the generalized P-K formula method to complete the queueing analysis on two-state Markov channel.

3.1 Two-state Channel Model

The aim of this paper is to analyze the performance of Markov fading channels without resorting to the matrix-geometric method. We are interested in the physical laws that govern system behavior, and parameters that characterize the impact of wireless channel variations on the queueing delay. There is an inherent trade-off between the accuracy of modeling and the complexity of analysis. Our analysis focuses on

Markov channels with two service states, which are mathematically tractable, and enables us to capture the essential characteristics of wireless fading channels. The parameters and assumptions of the system are introduced in this section to facilitate our discussions. We expect that the results obtained by the two-state model will shed some light on the analysis of general Markov channels in the future research.

3.1.1 Markov model of fading channels

The two-state Markov chain of a wireless fading channel is shown in Fig. 3.1, which is called Markov modulated service process (MMSP) in this paper. The state $j = 0$ ($j = 1$) represents the channel is in *Good* (*Bad*) state, respectively. The transition rate from state j to $\bar{j}$ is denoted as f_j, where $\bar{j}$ is the complement of j. The state transition rate f_j is determined from level-crossing rate at the physical layer, which is a linear function of the maximum Doppler frequency, i.e. $f_j \propto f_D$, as shown in [50, 60] for Rayleigh fading channels and in [22] for general Nakagami fading channels. Thus, the steady-state probability that the channel is in state j is given by

$$\pi_j = \frac{f_{\bar{j}}}{f_0 + f_1}. \tag{3.1}$$

We assume that the channel service rate is a constant during one symbol interval, but it may vary during the transmission period of a packet. The channel is in state $j = 0$ ($j = 1$) if the received signal-to-ratio (SNR) is above (below) some predetermined threshold. Each channel state is associated with a constant error probability of received symbol, which is averaged over all probabilities in that channel state [50, 60]. For a specific FEC coding scheme, the probability of having one or more uncorrectable error symbols in the received bits in each state is reported in [26, 64], from which the net data-rate (symbols of the original message transmitted per unit time) in each state j of the channel can be determined along with the specific coding rate.

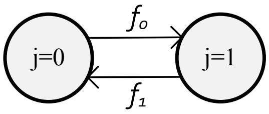

Figure 3.1: Two-state Markov model.

3.1.2 Parameters of $M/MMSP/1/K$ queueing model

We assume that the channel under consideration is a non-adaptive wireless system that employs an ideal Type I Hybrid ARQ scheme with a fixed data-rate, and there is an error-free feedback channel [62], through which an ACK/NACK will be sent back immediately after each transmission. If a packet is transmitted successfully, the next packet in the queue will be served; otherwise, the same head-of-line (HOL) packet will be retransmitted until it is successfully received. For the tractability of our model, we further assume that the stream of packets input to a finite first-in-first-out (FIFO) buffer is a Poisson process with mean rate λ packets per unit time, and the packet lengths are independent and identically distributed exponential random variables. The packet length, along with the net data-rate, determines the service rate μ_j in each channel state $j \in \{0, 1\}$. That is, packets are successfully transmitted with mean service rate μ_j (packets per unit time), and the service rate μ_0 in *Good* state is larger than μ_1 in *Bad* state. In this paper, we show that the queueing behavior of the system can be described by closed-form expressions under these assumptions.

At time t, the state of the system $X(t)$ is defined as the number of packets in the system, including the one in service, while the channel state is denoted by $V(t)$. The process $\{(X(t), V(t)), t \geq 0\}$ is a Continuous Time Markov Chain (CTMC) with a finite state space $\{(i, j), i = 0, 1, 2, ..., K, j = 0, 1\}$. We denote this queuing model as $M/MMSP/1/K$ with the state transition diagram shown in Fig. 3.2.

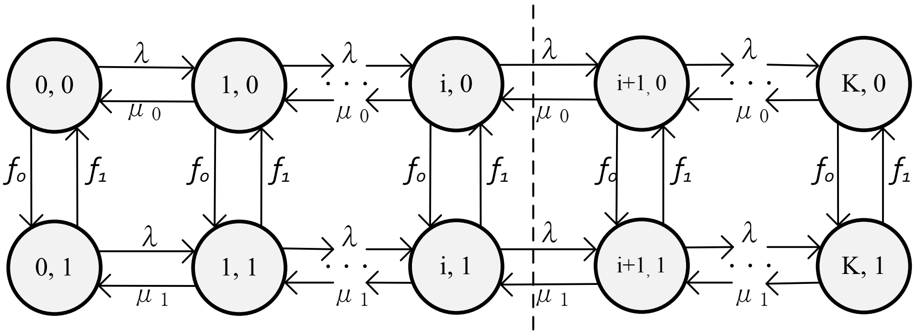

Figure 3.2: Rate transition diagram with finite buffer K.

The steady state probabilities of the $M/MMSP/1/K$ are defined by:

$$p_{i,j} = \lim_{t\to\infty} Pr\{X(t) = i, V(t) = j\}, \tag{3.2}$$

where $i = 0, 1, 2, ..., K$ and $j = 0, 1$. Thus, the probability that there are i packets in the system in steady state is given by:

$$P_i = p_{i,0} + p_{i,1}. \tag{3.3}$$

The following set of Kolmogorov forward equations, or so called balance equations, can be directly derived from the state transition diagram shown in Fig. 3.2:

$$(\lambda + f_j)p_{0,j} = f_{\bar{j}}p_{0,\bar{j}} + \mu_j p_{1,j} \tag{3.4a}$$

$$(\lambda + f_j + \mu_j)p_{i,j} = f_{\bar{j}}p_{i,\bar{j}} + \lambda p_{i-1,j} + \mu_j p_{i+1,j} \tag{3.4b}$$

$$(f_j + \mu_j)p_{K,j} = f_{\bar{j}}p_{K,\bar{j}} + \lambda p_{K-1,j} \tag{3.4c}$$

for all $i = 0, 1, 2, ..., K$ and $j = 0, 1$. The balance equation with respect to the dashed line of the state transition diagram is given by

$$\lambda(p_{i,0} + p_{i,1}) = \mu_0 p_{i+1,0} + \mu_1 p_{i+1,1}. \tag{3.5}$$

Summing equation (3.5) over index i , we obtain

$$\lambda(1 - P_K) = \mu_0\pi_0 + \mu_1\pi_1 - \mu_0 p_{0,0} - \mu_1 p_{0,1}. \tag{3.6}$$

The left-hand side of (3.6) is the net packets input rate, denoted as $\lambda' = \lambda(1 - P_K)$, while the right-hand side is the system throughput, or the mean number of packets transmitted by the system per unit time. Thus, the capacity of this channel can be defined as follows [18]:

$$\hat{\mu} = \pi_0\mu_0 + \pi_1\mu_1, \tag{3.7}$$

which is the maximum of the right-hand side of (3.6) when $p_{0,j}$ approaches 0, meaning that the system is busy with probability 1. Detailed discussions on channel capacity $\hat{\mu}$ are presented in subsection A of Section 3.4.

3.2 $M/MMSP/1/K$ Queue

In this section, we derive the buffer overflow probability and queueing delay of the $M/MMSP/1/K$ queue from the conditional generating function $G_j(z)$, which is defined as

$$G_j(z) = \sum_{i=0}^{i=K} z^i P_{i,j}, \qquad |z| \leq 1, \qquad j = 0, 1.$$

Multiply $z^i (i = 0, 1, 2, ..., K)$ on both sides of (3.4) appropriately, and sum over all i, we arrive at

$$(\lambda + f_j + \mu_j)G_j(z) = f_{\bar{j}}G_{\bar{j}}(z) + \lambda z G_j(z) + \frac{\mu_j}{z}G_j(z) + (\mu_j P_{0,j} - \lambda P_{K,j} z^{K+1})(1 - \frac{1}{z}) \tag{3.8}$$

Solving the equation set (3.8) for $j = 0, 1$, we get

$$\begin{aligned} G_j(z) = &\frac{1}{g(z)}\Big(f_{\bar{j}}(\mu_{\bar{j}} p_{0,\bar{j}} - \lambda z^{K+1} p_{K,\bar{j}})z + \\ &(\mu_j p_{0,j} - \lambda z^{K+1} p_{K,j})\left(-\lambda z^2 + (\lambda + \mu_{\bar{j}} + f_{\bar{j}})z - \mu_{\bar{j}}\right)\Big) \end{aligned} \tag{3.9}$$

where

$$g(z) = (z - 1)(\lambda z - \mu_0)(\lambda z - \mu_1) - z(\lambda z - \hat{\mu})(f_0 + f_1) \tag{3.10}$$

Define the system generating function $G(z) = \sum_{i=0}^{i=K} z^i P_i$, whose closed-from expression can be obtained from conditional generating functions $G_j(z)$, as

$$\begin{aligned} G(z) &= G_0(z) + G_1(z) \\ =&\frac{1}{g(z)}\Big(- \lambda(\hat{\mu} - \lambda(1-P_K))z^2 - \mu_0\mu_1 P_0 + ((\lambda + f_0 + f_1)(\hat{\mu} - \lambda(1-P_K)) + \mu_0\mu_1 P_0)\, z \\ &- \lambda z^{K+1}\big(- \lambda P_K z^2 - \mu_0 p_{K,1} - \mu_1 p_{K,0} + ((\lambda + f_0 + f_1)P_K + \mu_0 p_{K,1} + \mu_1 p_{K,0})\, z\big)\Big), \end{aligned} \tag{3.11}$$

The function $g(z)$ in the denominator processes three roots z_0, z_1 and z_2 ($z_0 < z_1 < z_2$ by default) which locate at the three points of intersections of the following two curves [10], as plotted in Fig. 3.3:

$$y_1(z) = (z-1)(\lambda z - \mu_0)(\lambda z - \mu_1)$$

and

$$y_2(z) = z(\lambda z - \hat{\mu})(f_0 + f_1).$$

For positive λ, μ_0, μ_1, f_0 and f_1, it has been proved in [10] that the following relationship always holds:

$$0 < z_0 < min(1, \tfrac{\mu_0}{\lambda}) \;\text{ and }\; z_2 > max(1, \tfrac{\mu_1}{\lambda}). \tag{3.12}$$

3.2.1 Buffer overflow probability

Since z_0, z_1 and z_2 are the three roots of $g(z)$, (3.11) can be expressed as

$$G(z) = A\frac{1-(z_1^{-1}z)^{K+1}}{1-z_1^{-1}z} + B\frac{1-(z_2^{-1}z)^{K+1}}{1-z_2^{-1}z} + C\frac{z_0^K - z_0^{-1}z^{K+1}}{1-z_0^{-1}z}. \tag{3.13}$$

The probability that there are i packets in the system

$$P_i = Az_1^{-i} + Bz_2^{-i} + Cz_0^{K-i}. \tag{3.14}$$

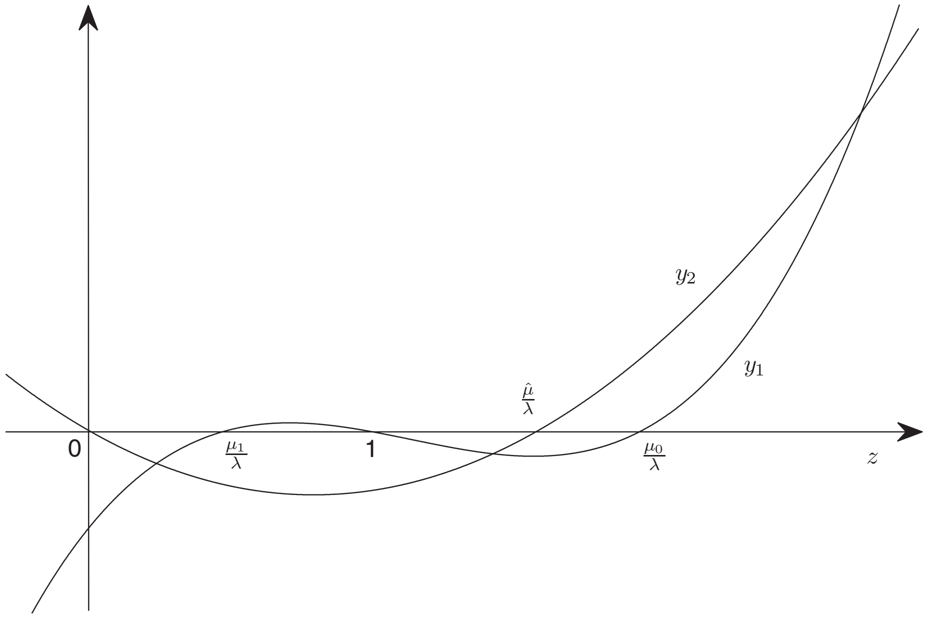

Figure 3.3: z_0, z_1 and z_2 locate at the three points of intersections of y_1 and y_2.

can be obtained by taking the inverse z-transform of $G(z)$. The probability $p_{i,j}$ that there are i packets in the system while the wireless channel is in state j can be derived from P_i. Due to the assumption $\mu_0 \neq \mu_1$, from (3.3) and (3.5) we have

$$p_{i,j} = \frac{\lambda P_{i-1} - \mu_{\bar{j}} P_i}{\mu_j - \mu_{\bar{j}}} \tag{3.15}$$

for $i = 1, 2, 3..., K$. Substituting (3.14) into (3.15), it yields

$$p_{i,j} = \frac{\lambda z_1 - \mu_{\bar{j}}}{\mu_j - \mu_{\bar{j}}} A z_1^{-i} + \frac{\lambda z_2 - \mu_{\bar{j}}}{\mu_j - \mu_{\bar{j}}} B z_2^{-i} + \frac{\lambda z_0 - \mu_{\bar{j}}}{\mu_j - \mu_{\bar{j}}} C z_0^{K-i}. \tag{3.16}$$

It can be verified that (3.16) is also satisfied under the condition $i = 0$.

Since $z_0 \in (0, 1)$, the last component of (3.14) is given as $C z_0^{K-i}$ instead of $C z_0^{-i}$. A general form for the stationary probability distribution of $M/G/1$-type Markov chains is presented in [1], where the authors decompose the generalized system into $forward$ and $backward$ subsystems using Matrix-geometric method. Although our system model is different from $M/G/1$, the component $C z_0^{K-i}$ in our expression corresponds to the *backward* subsystem defined in [1]. For the special case when $\mu_1 = 0$, there does not exist z_0. However, our analysis are still valid if we take $z_0 = 0$ under this environment and the discussion is detailed in Appendix A.

Parameters A, B and C can be derived from the initial conditions (3.4a), the end boundary conditions (3.4c), and the property that the summation of all probabilities goes to unity, specifically,

$$\begin{aligned} A =& \tfrac{1}{R}\tfrac{1-z_1^{-1}}{\lambda z_1-\hat{\mu}}\left(1-\left(\tfrac{z_0}{z_2}\right)^{K+1}\right), \\ B =& \tfrac{1}{R}\tfrac{1-z_2^{-1}}{\lambda z_2-\hat{\mu}}\left(\left(\tfrac{z_0}{z_1}\right)^{K+1}-1\right), \\ C =& \tfrac{1}{R}\tfrac{1-z_0}{\lambda z_0-\hat{\mu}}\left(z_1^{-(K+1)}-z_2^{-(K+1)}\right) \end{aligned} \tag{3.17}$$

where R is a normalization coefficient

$$R = \tfrac{1-z_1^{-(K+1)}}{\lambda z_1-\hat{\mu}}\left(1-\left(\tfrac{z_0}{z_2}\right)^{K+1}\right)+\tfrac{1-z_2^{-(K+1)}}{\lambda z_2-\hat{\mu}}\left(\left(\tfrac{z_0}{z_1}\right)^{K+1}-1\right)+\tfrac{1-z_0^{K+1}}{\lambda z_0-\hat{\mu}}\left(z_1^{-(K+1)}-z_2^{-(K+1)}\right).$$

From the relationships of z_0, z_1 and z_2 shown in (3.12), we conclude that coefficients A, B and C are all positive values. In this derivation, we use (3.15) which are based on $\mu_0 \neq \mu_1$. However, the expressions of A, B and C we obtained are still valid under the special case $\mu_0 = \mu_1$.

Now that we have the stationary probability distribution of the $M/MMSP/1/K$ with two states, its buffer overflow probability is given by

$$P_K = p_{K,0} + p_{K,1} = Az_1^{-K} + Bz_2^{-K} + C. \tag{3.18}$$

Here the buffer overflow probability is derived based on that a new arrival is blocked when there are already K packets in the system. For some real systems, the buffer is estimated by memory capacity instead of number of number of packets. With the random packet length assumption, the buffer may not be full if the sizes of all those K packets in the system are small and additional new packets can be stored in the buffer. The analysis of buffer overflow probability by tracking the sizes of packets in the buffer is much complicated and is beyond our interest. However, our P_K still provides a good estimation with respect to average packet length, especially when K is large, if those packet lengths are independent and identically distributed.

3.2.2 Large-deviation approximation

Due to (3.18), expressing K in terms of P_K as a simple expression does not seem to be possible. It is also common to estimate the buffer overflow probability from the tail distribution of infinite queue by using the theory of large deviations. Let $X(\infty)$ denote the number of packets in the system when the system is stable. According to the theory of large deviations, for large values of K, we have

$$P\{X(\infty) > K\} \sim \alpha e^{-\theta K}, \tag{3.19}$$

where $\theta = \log z_1$ is the asymptotic decay rate and α is the asymptotic constant [9,19,32]. When K is not too large, an approximation by setting $\alpha \approx P\{X(\infty) > 0\}$ is commonly adopted to analyze the wireless system performance, and is given in [24, 54] as follows:

$$P\{X(\infty) > K\} \approx P\{X(\infty) > 0\}e^{-\theta K}, \tag{3.20}$$

with parameters defined in [54] as follows:

$$P\{X(\infty) > 0\} = 1 - \lim_{K - \infty} P_0.$$

A comparison between the exact buffer overflow probability (3.18) and the approximation (3.20) obtained from the large deviation method with $\alpha \approx P\{X(\infty) > 0\}$ is shown in Fig. 3.4. These curves show that the exact buffer overflow probabilities agree with the simulation results, which are upper bounded by the approximation (3.20) when $K > 1$. Thus, the approximation obtained from the large deviation method can only serve as a conservative estimate of the overflow probability, which agrees with the observation reported in [9]. An alternate approach was proposed

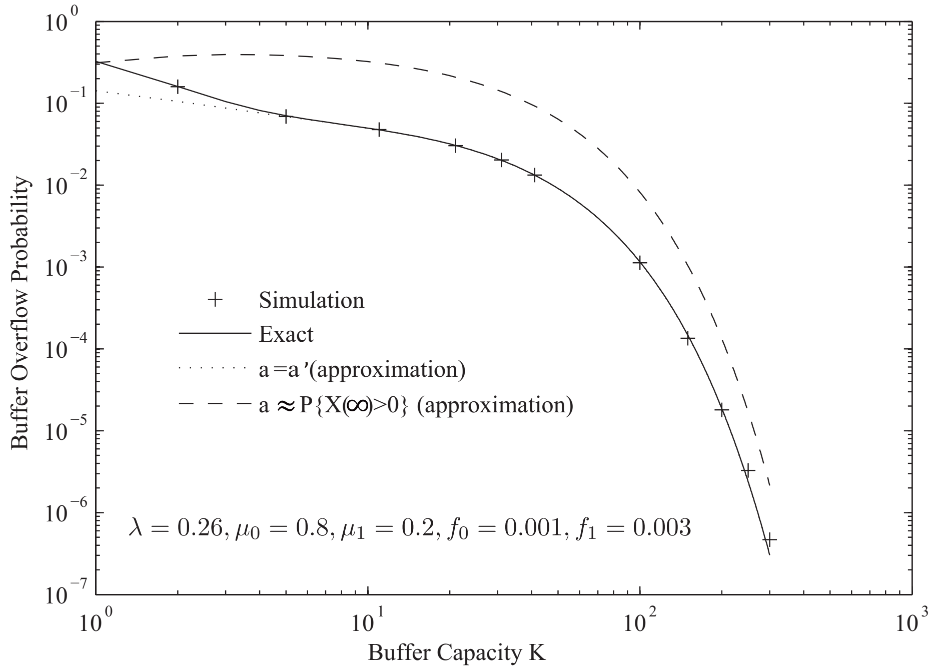

Figure 3.4: A comparison between approximations and exact values of buffer overflow probability.

in [30] to improve the large-deviation bound given in (3.20). Actually, a much more improved asymptotic constant α in expression (3.19) can be simply determined from our exact expression of buffer overflow probability (3.18). Substituting the constant C given in (3.17) into (3.18), we have

$$P_K = \left(A + \tfrac{1}{Rz_1}\tfrac{1-z_0}{\lambda z_0 - \hat{\mu}}\right) z_1^{-K} + \left(B + \tfrac{1}{Rz_2}\tfrac{1-z_0}{\lambda z_0 - \hat{\mu}}\right) z_2^{-K} \approx \left(A + \tfrac{1}{Rz_1}\tfrac{1-z_0}{\lambda z_0 - \hat{\mu}}\right) z_1^{-K}. \tag{3.21}$$

We know from (3.12) that $0 < z_1 < z_2$. It follows that a theoretically sound estimation of the constant α in (3.19) for large value of K should be given by

$$\alpha' = A + \tfrac{1}{Rz_1}\tfrac{1-z_0}{\lambda z_0 - \hat{\mu}}. \tag{3.22}$$

The approximation curve with α' is also plotted in Fig. 3.4. It fits well with the exact overflow probability (3.18) when $K > 10$. Due to the limitation of the theory of large deviations, however, both approximations failed to predict the buffer overflow probability for small values of K ($K < 10$), as shown in Fig. 3.4.

3.2.3 Queueing delay

The mean queue length $L = \sum_{i=0}^{i=K} iP_i$ can be directly derived from the generating function (3.11) and is given as follows:

$$L = \frac{\lambda(1-(K+1)P_K)}{\hat{\mu}-\lambda} + \frac{(1-p_{K,1})\lambda\mu_0+(1-p_{K,0})\lambda\mu_1-\lambda\hat{\mu}-\mu_0\mu_1\rho}{(f_0+f_1)(\hat{\mu}-\lambda)}, \tag{3.23}$$

where $\rho = 1 - P_0$ is the server utilization. If we take the limit $K \to \infty$, then $\lim_{K\to\infty} P_K = 0$, $\lim_{K\to\infty} p_{K,j} = 0$, and $\lim_{K\to\infty} \lambda' = \lambda$. From (3.23), the mean queue length of the system with infinite buffer capacity becomes:

$$\lim_{K\to\infty} L = \frac{\lambda}{\hat{\mu}-\lambda} + \frac{\lambda(\mu_0+\mu_1-\hat{\mu})-\mu_0\mu_1\rho}{(f_0+f_1)(\hat{\mu}-\lambda)} \tag{3.24}$$

which agrees with the result presented by Eisen and Tainiter [10] and Yechiali and Naor [57]. The mean waiting time is obtained from Little's Law as follows:

$$W = (L - \rho)/\lambda'. \tag{3.25}$$

With respect to different channel varying rates, the trade-off between buffer overflow probability and delay is shown in Fig. 3.5. The wireless channel alternates between *Good* and *Bad* states with frequencies $f_0 = 0.1f$ and $f_1 = 0.3f$ respectively, where f is a parameter representing different state-varying speeds, or different Doppler frequencies f_D, of the wireless channel. We consider a fixed channel capacity $\hat{\mu} = 0.65$ with three different values of f, 0.0001, 0.01 and 1, corresponding to three different channel state-varying speeds. Our results demonstrate that the buffer overflow probabilities and delays of *M/MMSP/1/K* queue are very sensitive to the value of f. As shown in Fig. 3.5, the buffer overflow probability gradually decreases as the mean waiting time increases in all three cases. With the same channel capacity and buffer size, a slow varying channel (small f) generates a much larger delay and

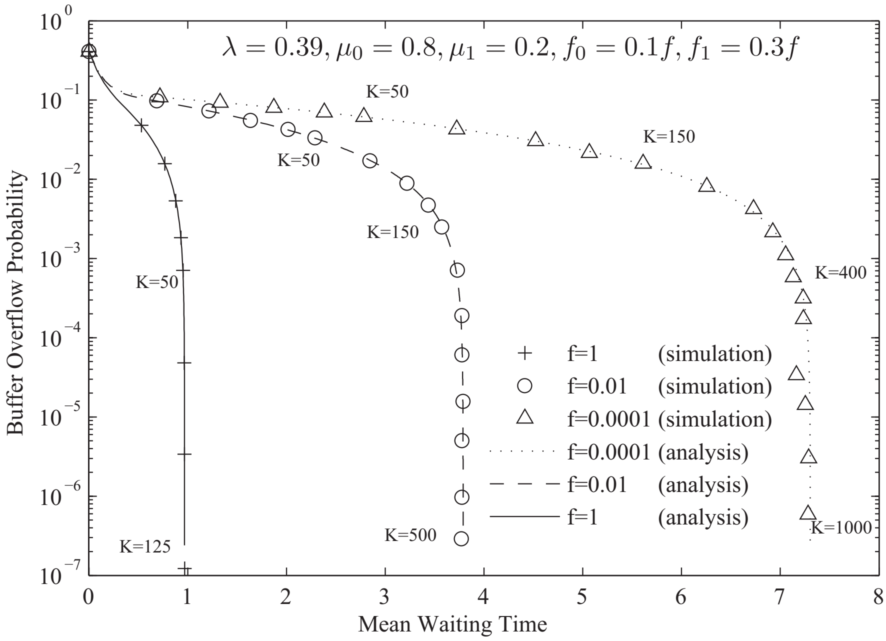

Figure 3.5: Buffer overflow probability versus mean waiting time for different channel varying rates.

a larger buffer overflow probability. In the following sections, we will concentrate our discussion on this phenomena based on the start-service probability and the generalized P-K formula of the *M/MMSP/1/K* queue.

3.3 Start-service Probability

The channel state at the beginning of the service of a head-of-line (HOL) packet is called the *start-service state* of this packet. The start-service probability $\hat{\pi}_j$ of an HOL packet is defined as the probability that the start-service state of this packet is j. Note that the probability $\hat{\pi}_j$ is different from π_j. The latter is the probability that the channel state is in j and it is also the probability that a newly arrived packet sees the channel in state j due to PASTA [52]. That is, the probability $\hat{\pi}_j$ is averaged over all HOL packets whereas π_j is averaged over time. It should be noted that the difference between $\hat{\pi}_j$ and π_j was first studied in [43]. It is shown in [32] that the probability $\hat{\pi}_j$ is not only a function of π_j but also dependent on the packet arrival rate λ. However, only the following two extreme cases

$$\hat{\pi}_j = \begin{cases} \pi_j & \text{when } \lambda \to 0 \\ \eta_j = \frac{\mu_j f_{\bar{j}}}{\mu_0 f_1 + \mu_1 f_0} & \text{when } \lambda \to \infty \end{cases} \tag{3.26}$$

were known. The impact of the arrival process on the service time was first investigated via start-service probability in [2], where the Markov channel is bound to change only after each packet transmission. Unfortunately, the authors concluded from numerical results that the service time is insensitive to the packet arrival rate. In the next theorem, we show that the state transition factor $\beta \in (0, 1)$ defined as follows:

$$\beta = \frac{\mu_0 \mu_1}{\mu_0 \mu_1 + \mu_0 f_1 + \mu_1 f_0} \tag{3.27}$$

is the key to derive the start-service probability $\hat{\pi}_j$. Furthermore, we prove that the service time is insensitive to packet arrival rate only for small β, as depicted in Fig. 3.11.

Suppose that a new arrival sees i packets in the system. If $i = K$, then this newly arrived packet will be blocked. If $i \leq K - 1$, then we assume that these packets in the FIFO buffer are sequentially numbered by $m = 0, 1, \cdots, i$, the one in service is numbered 0 and the new arrival is numbered i, as shown in Fig. 3.6. Let $\hat{\pi}_j(i, m)$ be the conditional start-service probability of m^{th} packet defined as follows:

$$\begin{aligned} \hat{\pi}_j(i, m) = P\{&\text{start-service state of } m^{th} \text{ packet is } j \\ &\mid \text{an arrival sees } i \text{ packets in the system}\} \end{aligned}$$

for $0 \leq m \leq i$, and $j = 0, 1$. The start-service probability $\hat{\pi}_j$ of an HOL packet expressed in terms of state transition factor β is derived from the conditional start-service probability $\hat{\pi}_j(i, m)$ of m^{th} packet in the following theorem.

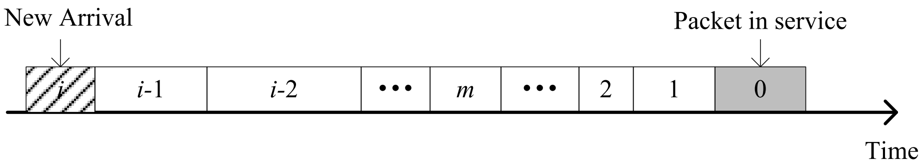

Figure 3.6: A new arrival sees i packets in the system.

Theorem 3.1. *The probability that the service of an HOL packet starts with channel state j is given by:*

$$\hat{\pi}_j = \eta_j + \tfrac{1}{1-P_K}\left(\eta_{\bar{j}}(G_j(\beta) - \beta^K p_{K,j}) - \eta_j(G_{\bar{j}}(\beta) - \beta^K p_{K,\bar{j}})\right), \tag{3.28}$$

for $j = 0, 1$.

Proof. The channel may change its state during the service of a packet. We first consider the transition probability of two consecutive start-service states. Let $P\{j'|j\}$ be the probability that a packet starts the service in channel state j and finishes the service in channel state j'. Define the following conditional state transition probabilities during the service of a packet:

$$\begin{aligned} q_j =& P\{\text{next state transition occurs before service} \\ & \text{completion} \mid \text{channel is in state } j\} = \tfrac{f_j}{\mu_j+f_j}, \end{aligned}$$

for $j = 0, 1$. Then we have

$$P\{j'=0|j=0\} = \sum_{n=0}^{\infty}(q_0q_1)^n(1-q_0) = \sum_{n=0}^{\infty}\left(\tfrac{f_0f_1}{(\mu_0+f_0)(\mu_1+f_1)}\right)^n \tfrac{\mu_0}{\mu_0+f_0} = \beta\left(1+\tfrac{f_1}{\mu_1}\right).$$

Similarly, we can obtain:

$$\begin{aligned} P\{j'=1|j=1\} =& \sum_{n=0}^{\infty}(q_0q_1)^n(1-q_1) = \beta\left(1+\tfrac{f_0}{\mu_0}\right), \\ P\{j'=1|j=0\} =& 1 - P\{j'=0|j=0\} = \beta\tfrac{f_0}{\mu_0}, \\ P\{j'=0|j=1\} =& 1 - P\{j'=1|j=1\} = \beta\tfrac{f_1}{\mu_1}. \end{aligned}$$

We then establish the following set of equations from the definition of $\hat{\pi}_j(i,m)$:

$$\begin{aligned}\hat{\pi}_0(i,m+1) =&\hat{\pi}_0(i,m)P\{j'=0|j=0\}+\hat{\pi}_1(i,m)P\{j'=0|j=1\},\\ \hat{\pi}_1(i,m+1) =&\hat{\pi}_0(i,m)P\{j'=1|j=0\}+\hat{\pi}_1(i,m)P\{j'=1|j=1\}.\end{aligned}$$

Solving the above set of difference equations together with the condition $\hat{\pi}_0(i,m)+\hat{\pi}_1(i,m)=1$, for all $0\leq m\leq i$, we get

$$\hat{\pi}_j(i,m)=\beta^m\left(\hat{\pi}_j(i,0)-\eta_j\right)+\eta_j. \tag{3.29}$$

From the probability P_i that a newly arrived packet sees i packets in the system and $p_{i,j}$ that a newly arrived packet sees i packets in the system while the current channel is in state j, we obtain the following initial state probability:

$$\hat{\pi}_j(i,0)=p_{i,j}/P_i. \tag{3.30}$$

Since an arrival that sees i packets in the system starts the service in state j with probability $\hat{\pi}_j(i,i)$, we then obtain the following start-service probability by combining (3.29) and (3.30):

$$\hat{\pi}_j=\tfrac{1}{1-P_K}\sum_{i=0}^{K-1}P_i\hat{\pi}_j(i,i)=\eta_j+\tfrac{1}{1-P_K}\left(\eta_{\bar{j}}(G_j(\beta)-\beta^Kp_{K,j})-\eta_j(G_{\bar{j}}(\beta)-\beta^Kp_{K,\bar{j}})\right). \tag{3.31}$$

□

As we mentioned before, the probability $\hat{\pi}_j$ is averaged over all HOL packets whereas π_j is averaged over time. In contrast to the expression (3.31) of $\hat{\pi}_j$, the probability π_j that the channel is in state j can be expressed from (3.30) as follows:

$$\pi_j=\sum_{i=0}^{K}p_{i,j}=\sum_{i=0}^{K}P_i\hat{\pi}_j(i,0). \tag{3.32}$$

If the loading is very low, meaning $P_0\rightarrow 1$, then it is clear that both expressions of $\hat{\pi}_j$ and π_j approach $\hat{\pi}_j(0,0)$, which is consistent with (3.26) that $\hat{\pi}_j=\pi_j$ when

$\lambda \to 0$. On the other hand, if arrival packets always find a large number of packets waiting in the buffer, then (3.29) and (3.31) reveal that the transmission of a packet is started in state j with a probability close to η_j, which is again consistent with (3.26) that $\hat{\pi}_j = \eta_j$ when $\lambda \to \infty$. Thus, the expression (3.31) of the start-service probability $\hat{\pi}_j$ covers the two extremes originally derived in [32] as special cases.

The expression (3.29) implies that the state transition factor β indicates how fast the channel state probability tends to be stable. This point can be illustrated by the following two extreme cases:

1. $\beta \to 1$, when the state transition rate f_j is very small, corresponding to a system with very slow changes of channel rate. The probability $\hat{\pi}_j(i, m)$ reaches η_j only after a large number m of packets have been served in a busy period.

2. $\beta \to 0$, when the state transition rate f_j is extremely large compared to service rate μ_j. According to (3.29), the probability $\hat{\pi}_j(i, m)$ reaches η_j after a small number m of packets have been served. In this case, the *M/MMSP/1/K* queue behaves as an *M/G/1/K* queue with the same first and second moments of service time, because the dependency among service times becomes insignificant and, therefore, can be ignored. A detailed discussion of this point is provided in subsection D of Section 3.4.

In the next section, we show that the state transition factor β also plays an essential role in the derivation of queue length distribution.

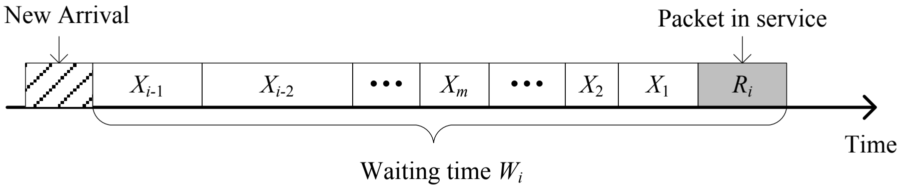

Figure 3.7: Waiting time of an arrival seeing i packets in the system.

3.4 Generalized Pollaczek-Khinchin Formula for the *M/MMSP/1/K* Queue

In this section, we derive the generalized P-K formula for the *M/MMSP/1/K* queue, which shows that the queueing delay is proportional to $\frac{1}{1-\beta}$. It should be noted that the analysis described in this section is independent of the number of states of the Markov channels. The waiting time analysis is based on the residual service time similar to the method described in [5]. As illustrated in Fig. 3.7, we define

- Q_c = The number of packets in the system seen by an arrival ($0 \leq Q_c \leq K$).
- W_i = Waiting time of an arrival seeing i ($0 \leq i \leq K-1$) packets in the system. An arrival that sees K packets in the system is blocked with probability P_K, in which case there is no waiting time.
- R_i = Residual service time of an arrival that sees i ($1 \leq i \leq K-1$) packets in the system. That is, R_i is the remaining time until the completion of current service. If $i = 0$, then the system is empty and there is no residual service time.
- X_m = Service time of the m^{th} ($1 \leq m \leq i-1$) packet in the queue, starting from the first packet behind the head-of-line (HOL) packet, which is in service.

According to the above definitions, the waiting time of an arrival is given by

$$W_i = \begin{cases} R_i + \sum_1^{i-1} X_m & 2 \leq i \leq K-1, \\ R_i & i = 1, \\ 0 & i = 0. \end{cases}$$

Taking expectation, we have

$$W = \frac{1}{1-P_K} \sum_{i=0}^{K-1} P_i W_i = R + \frac{1}{1-P_K} \sum_{i=2}^{K-1} P_i \sum_{m=1}^{i-1} E[X_m | Q_c = i], \tag{3.33}$$

where W is the average waiting time and $R = \frac{1}{1-P_K} \sum_{i=1}^{K-1} P_i R_i$ is the mean residual service time. In the rest of this section, we will derive the generalized P-K formula of *M/MMSP/1/K* queue from the expression (3.33).

3.4.1 Moments of Service time

The derivations of the mean residual service time R and the service time of the m^{th} waiting packet X_m require the first and second moments, $E[T]$ and $E[T^2]$, of the service time of the *M/MMSP/1/K* queue. They can be obtained from the first and second conditional moments $E[T_j]$ and $E[T_j^2]$ $(j = 0, 1)$ defined in [32] as follows:

$$E[T] = \sum_{j=0,1} \hat{\pi}_j E[T_j], \tag{3.34a}$$

$$E[T^2] = \sum_{j=0,1} \hat{\pi}_j E[T_j^2], \tag{3.34b}$$

where $E[T_j]$ and $E[T_j^2]$ are the first and second conditional moments of service time given that the service begins with state j. According to [32], the closed-form expressions of $E[T_j]$ and $E[T_j^2]$, expressed in terms of μ_0, μ_1, f_0 and f_1, are obtained as following:

Consider an arbitrary packet whose service starts in channel state j. Let T_j be the random variable denoting the total service time for this packet. During the service of this packet, the channel state j may change to $\overline{j}$ after time T_f when the serve speed alternates, or even stay at j until the serve is finished after time T_μ. Conditioning on steady state, T_μ and T_f are exponentially distributed with parameters μ_j and f_j respectively. Then, T_j can be calculated as

$$T_j = min(T_f, T_\mu) + \begin{cases} 0 & \text{with probability } \frac{\mu_j}{\mu_j+f_j} \\ T_{\bar{j}} & \text{with probability } \frac{f_j}{\mu_j+f_j} \end{cases}$$

where $j = 0, 1$.

Taking the Laplace Stieltjes transform (LST) on both sides, we get

$$E(e^{-sT_j}) = \frac{\mu_j + f_j}{s + \mu_j + f_j} \left(\frac{\mu_j}{\mu_j + f_j} + \frac{f_j}{\mu_j + f_j} E(e^{-sT_{\bar{j}}}) \right).$$

Arranging terms, we have

$$(s + \mu_j + f_j)E(e^{-sT_j}) = \mu_j + f_j E(e^{-sT_{\bar{j}}}). \tag{3.35}$$

Solving (3.35) at both $j = 0, 1$ states, and substituting $\mu_j + f_j = \mu_j + f_j$, we get

$$E(e^{-sT_j}) = \frac{\mu_j s + \mu_0\mu_1 + \mu_0 f_1 + \mu_1 f_0}{s^2 + (\mu_0 + \mu_1 + f_0 + f_1)s + \mu_0\mu_1 + \mu_0 f_1 + \mu_1 f_0}. \tag{3.36}$$

Taking the first and second derivatives of (3.36) with respect to s, and substituting $s = 0$, we have

$$E[T_j] = \frac{\mu_{\bar{j}} + f_0 + f_1}{\mu_0\mu_1 + \mu_0 f_1 + \mu_1 f_0}, \tag{3.37a}$$

$$E[T_j^2] = 2\frac{\mu_{\bar{j}}^2 + 2\mu_{\bar{j}} f_{\bar{j}} + f_j(\mu_0 + \mu_1) + (f_0 + f_1)^2}{(\mu_0\mu_1 + \mu_0 f_1 + \mu_1 f_0)^2}. \tag{3.37b}$$

When the arrival rate goes to infinity, the mean service time becomes $1/\hat{\mu}$, as shown in the following expression:

$$\lim_{\lambda\to\infty} E[T] = \sum_{j=0}^{1} \lim_{\lambda\to\infty} \hat{\pi}_j E[T_j] = \sum_{j=0}^{1} \eta_j E[T_j] = \tfrac{1}{\hat{\mu}}, \tag{3.38}$$

which can be obtained by combining (3.26), (3.34a) and (3.37a).

The channel capacity $\hat{\mu}$ defined in (3.7) is also the bound of the maximum arrival rate input to a stable *M/MMSP/1* queue. With an infinite input buffer, the server

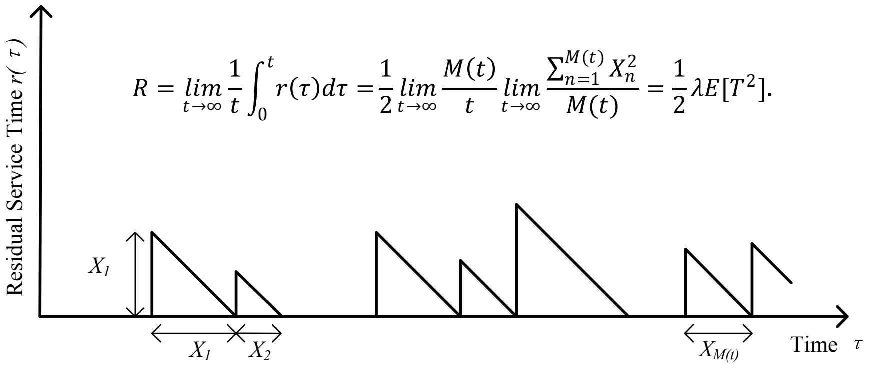
$R = \lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{2}\lim_{t\to\infty}\frac{M(t)}{t}\lim_{t\to\infty}\frac{\sum_{n=1}^{M(t)} X_n^2}{M(t)} = \frac{1}{2}\lambda E[T^2].$
Residual Service Time $r(\tau)$
X_1
X_1
X_2
$X_{M(t)}$
Time τ

Figure 3.8: Derivation of the mean residual service time (page 143-144 in [5]).

utilization can be readily obtained from Little's Law as follows:

$$\rho = \lambda E[T]. \tag{3.39}$$

The server utilization is also the probability that the channel is busy, or $\rho = 1 - P_0$. From (3.17) and (3.10), we obtain the following expression after extensive algebra:

$$\rho = 1 - (\hat{\mu} - \lambda)\tfrac{z_0}{1-z_0}\tfrac{\lambda(1-z_0)+f_0+f_1}{\mu_0\mu_1}. \tag{3.40}$$

Since $z_0 \in (0,1)$, the stability condition ($\rho < 1$) of the *M/MMSP/1* queue implies $\lambda < \hat{\mu}$.

3.4.2 Residual service time

We need the mean residual service time R to complete the derivation of the P-K formula for the *M/MMSP/1/K* queue. There is no general expression of the mean residual service time R for finite buffer size K. In the case of infinite input buffer, the following expression of mean residual service time

$$R = \tfrac{1}{2}\lambda E[T^2] \tag{3.41}$$

of *M/MMSP/1* queue can be derived in the same manner as that of the *M/G/1* queue, which was described in pages 143-144 of [5]. Fig. 3.8 illustrates the relationship between the residual service time and the arrivals. Suppose X_n is the service

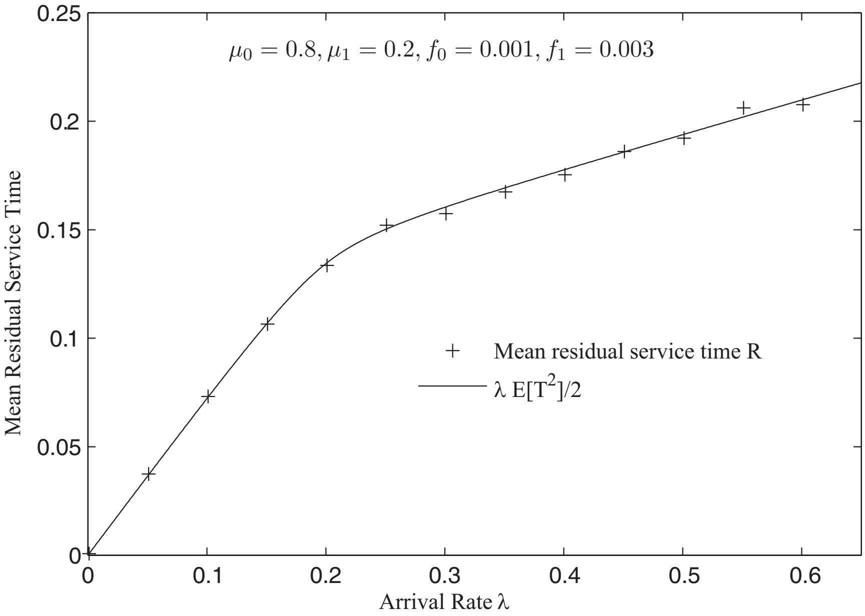

Figure 3.9: Simulations of mean residual service time R and $\frac{1}{2}\lambda E[T^2]$ of the *M/MMSP/1* queue.

time of the n^{th} customer, then its residual service time $r(\tau)$ at time τ must fall linearly (with slope -1) to 0. A simple derivation of (3.41) given in [5] is displayed in Fig. 3.8. Since this fact has nothing to do with the property whether the service time is independently distributed (in the *M/G/1* model) or state dependent (in our *M/MMSP/1* model). Thus, the same proof of (3.41) can also be applied to our model, which is also firmly verified by the simulation result shown in Fig. 3.9. Note that, due to the state dependency of service time, the mean residual service time R of an *M/MMSP/1/K* queue cannot be obtained by simply extending (3.41) to $\frac{1}{2}\lambda(1-P_K)E[T^2]$. The derivation of a general expression of the mean residual service time R of the *M/MMSP/1/K* queue remains open.

3.4.3 Generalized P-K formula of Mean Waiting Time

Suppose that for an arrival seeing i packets in the system with probability P_i, the m^{th} packet in the queue will start its service after the previous m packets (including the one in service) were served, as demonstrated in Fig. 3.7. From Theorem 3.1, we know that the m^{th} packet will start the service in state j with probability $\hat{\pi}_j(i,m)$. Then, its averaged service time X_m under the condition that there are i packets in the system when a new packet arrives is given by

$$E[X_m|Q_c = i] = \sum_{j=0}^{1} E[T_j]\hat{\pi}_j(i,m). \tag{3.42}$$

Theorem 3.2. *The mean waiting time of an M/MMSP/1 queue is given by*

$$W = \frac{\frac{\lambda}{\hat{\mu}}E[T] + \frac{1}{1-\beta}\sum_{j=0}^{1} E[T_j](\pi_j - \hat{\pi}_j)}{1 - \frac{\lambda}{\hat{\mu}}}. \tag{3.43}$$

Proof. Substituting (3.29) into (3.42), the total time for all $i-1$ packets in queue that have been served is given by the unconditional (3.38):

$$\begin{aligned}\sum_{m=1}^{i-1} E[X_m|Q_c = i] =& \sum_{j=0}^{1} E[T_j] \sum_{m=1}^{i-1} \hat{\pi}_j(i,m) \\ =& \tfrac{1}{1-\beta} \sum_{j=0}^{1} E[T_j]\left(\hat{\pi}_j(i,0) - \hat{\pi}_j(i,i)\right) + \tfrac{i-1}{\hat{\mu}} - \sum_{j=0}^{1} E[T_j]\left(\hat{\pi}_j(i,0) - \eta_j\right).\end{aligned} \tag{3.44}$$

Averaging over all possible i, we obtain the following averaged time to serve all packets in queue seeing by an arbitrary arrival:

$$\begin{aligned}&\sum_{i=2}^{K-1} P_i \sum_{m=1}^{i-1} E[X_m|Q_c = i] \\ =& \tfrac{1}{1-\beta} \sum_{i=2}^{K-1} \sum_{j=0}^{1} E[T_j]\left(P_i\hat{\pi}_j(i,0) - P_i\hat{\pi}_j(i,i)\right) + \\ &\sum_{i=2}^{K-1} P_i \tfrac{i-1}{\hat{\mu}} - \sum_{i=2}^{K-1} \sum_{j=0}^{1} E[T_j]\left(P_i\hat{\pi}_j(i,0) - P_i\eta_j\right)\end{aligned}$$

$$= \tfrac{1}{1-\beta} \sum_{j=0}^{1} E[T_j] \left(\pi_j - p_{K,j} - \hat{\pi}_j(1-P_K)\right) + \tfrac{L_q-(K-1)P_K}{\hat{\mu}}$$

$$- \sum_{j=0}^{1} E[T_j] \left(\pi_j - P_{K,j} - \eta_j(1-P_K)\right) +$$

$$\sum_{j=0}^{1} E[T_j](p_{0,j} - P_0\eta_j), \tag{3.45}$$

where the second equation is obtained from (3.30), (3.31) and (3.32) and $L_q = \sum_{i=2}^{K}(i-1)P_i$. From Little's Law, we have

$$L_q = \lambda(1-P_K)W. \tag{3.46}$$

Substituting (3.45) and (3.46) into (3.33), we obtain the following mean waiting time of the *M/MMSP/1/K* queue:

$$W = \frac{1}{1-P_K}\frac{1}{1-\frac{\lambda}{\hat{\mu}}}\Big\{R(1-P_K) + \tfrac{1}{1-\beta}\sum_{j=0}^{1} E[T_j]\left(\pi_j - P_{K,j} - \hat{\pi}_j(1-P_K)\right) - \tfrac{(K-1)P_K}{\hat{\mu}} -$$

$$\sum_{j=0}^{1} E[T_j]\left(\pi_j - p_{K,j} - \eta_j(1-P_K)\right) + \sum_{j=0}^{1} E[T_j]\left(p_{0,j} - P_0\eta_j\right)\Big\}. \tag{3.47}$$

Taking the limit of $K \to \infty$, we obtain the following mean waiting time from (3.47):

$$W = \frac{R + \frac{1}{1-\beta}\sum_{j=0}^{1} E[T_j](\pi_j - \hat{\pi}_j) - \sum_{j=0}^{1} E[T_j](\pi_j - \eta_j) + \sum_{j=0}^{1} E[T_j]\left(p_{0,j} - P_0\eta_j\right)}{1-\frac{\lambda}{\hat{\mu}}}. \tag{3.48}$$

Using the mean residual service time R of *M/MMSP/1* given in (3.41) and after some algebra, the expression (3.48) can be simplified to (3.43). To prove the two expressions (3.43) and (3.48) are equivalent, we need to show that the following identity holds:

$$R - \sum_{j=0}^{1} E[T_j](\pi_j - \eta_j) + \sum_{j=0}^{1} E[T_j]\left(p_{0,j} - P_0\eta_j\right) = \tfrac{\lambda}{\hat{\mu}}E[T]. \tag{3.49}$$

From the closed-form expressions of $\hat{\pi}_j$, $E[T_j]$ and $E[T_j^2]$, previously derived in (3.28) and (3.37) respectively, it is straightforward to obtain each term of (3.49) as follows:

- The residual service time $R = \frac{1}{2}\lambda E[T^2]$ is given by (3.41), where the second moment of service time is obtained from (3.28), (3.34b) and (3.37b) as follws:

$$E[T^2] = 2\tfrac{\rho}{\lambda}\tfrac{\mu_0+\mu_1+f_0+f_1-\lambda/\rho}{\mu_0\mu_1+\mu_0 f_1+\mu_1 f_0}. \tag{3.50}$$

- From (3.1), (3.37a) and (3.38), we have

$$\sum_{j=0}^{1} E[T_j](\pi_j - \eta_j) = \sum_{j=0}^{1} E[T_j]\pi_j - \tfrac{1}{\hat{\mu}} = \tfrac{\mu_0+\mu_1+f_0+f_1-\hat{\mu}}{\mu_0\mu_1+\mu_0 f_1+\mu_1 f_0} - \tfrac{1}{\hat{\mu}}. \tag{3.51}$$

- From (3.30) and (3.38), we have

$$\sum_{j=0}^{1} E[T_j]\left(p_{0,j} - P_0\eta_j\right) = \sum_{j=0}^{1} E[T_j]p_{0,j} - P_0\sum_{j=0}^{1} E[T_j]\eta_j = \sum_{j=0}^{1} E[T_j]p_{0,j} - P_0\tfrac{1}{\hat{\mu}}, \tag{3.52}$$

  where

$$\sum_{j=0}^{1} E[T_j]p_{0,j} = \tfrac{(\mu_0+\mu_1+f_0+f_1)(1-\rho)-(\hat{\mu}-\lambda)}{\mu_0\mu_1+\mu_0 f_1+\mu_1 f_0} \tag{3.53}$$

  can be obtained from (3.6) and (3.37a).

Collectively, we establish the identity (3.49) by substituting (3.41), (3.50), (3.51), (3.52) and (3.53) into the left-hand side of (3.49), and (3.34a) and (3.39) into the right-hand side. □

Alternately, from (3.24) and (3.25), the mean waiting time of the system with infinite buffer capacity can be expressed as follows:

$$W = \frac{1}{\hat{\mu}-\lambda} + \frac{\mu_0+\mu_1-\hat{\mu}-\mu_0\mu_1\rho/\lambda}{(f_0+f_1)(\hat{\mu}-\lambda)} - E[T]. \tag{3.54}$$

From (3.1), (3.34a), (3.37a) and (3.39), it can be readily proven that the mean waiting time (3.43) obtained from the generalized P-K formula is the same as (3.54), the one derived from the generating function.

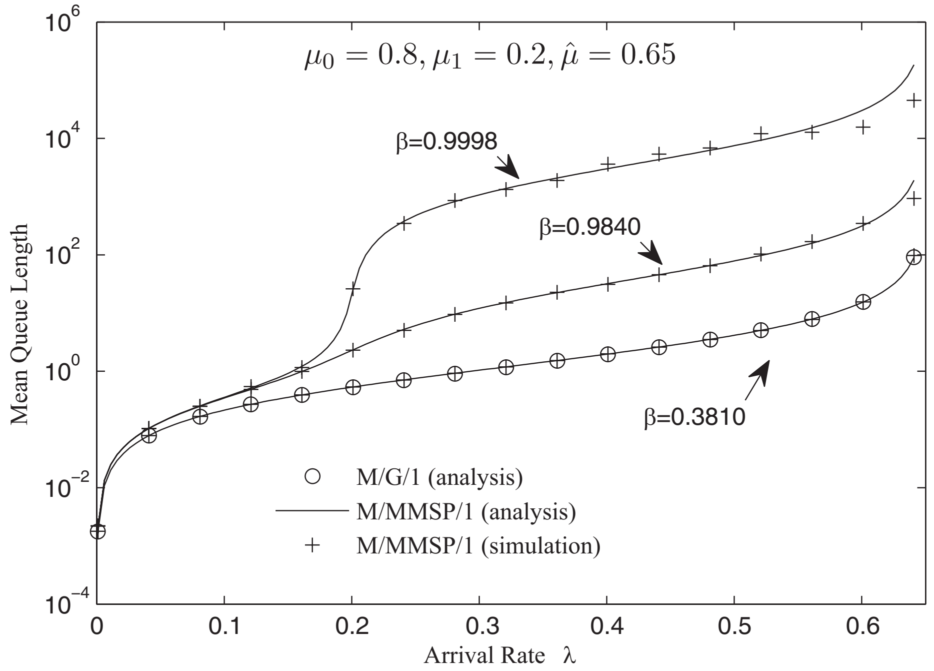

Figure 3.10: Mean queue length of *M/MMSP/1* and *M/G/1* with different state transition factor β.

3.4.4 The impact of state transition factor on queue length

The generalized P-K formula explicitly expresses the impact of the state transition factor on the performance of wireless channels. For a given fixed channel capacity $\hat{\mu}$, the generalized P-K formula (3.43) reveals that the mean waiting time is greatly affected by the state transition factor β. This point is illustrated in Fig. 3.10 by the following mean queue length for different β:

$$L = \lambda(E[T] + W).$$

Note that different values of state transition factor β in Fig. 3.10 correspond to different parameters f shown in Fig. 3.5. Both analytical and simulation results show that this mean queue length of *M/MMSP/1* is very sensitive to the state transition factor β when the arrival rate is larger than the service rate in a *Bad* channel state ($\lambda > \mu_1$). Because some packets arriving in the *Bad* state cannot be transmitted until the channel changes to the *Good* state, resulting in a long queue of backlogged packets. The P-K formula shows that the state transition factor β can be used to characterize the following two extreme cases:

1. $\beta \to 1$, for slow varying channel systems. It can be readily seen from (3.27) that β is close to 1 when the channel state transition rate f_j is much smaller than the service rate μ_j. Since the factor $\frac{1}{1-\beta}$ becomes extremely large when β is close to 1, there is a large mean waiting time of the *M/MMSP/1* queue due to the dominating term $\frac{1}{1-\beta}\sum_{j=0}^{1} E[T_j](\pi_j - \hat{\pi}_j)$ in the nominator of (3.43). In Fig. 3.10, the mean queue length at $\beta = 0.9998$ ($f = 0.0001$) is much longer than the one at $\beta = 0.9840$ ($f = 0.01$).

2. $\beta \to 0$, when the channel state transition rate f_j is extremely large compared to the service rate μ_j. According to (3.27), β is close to 0 and $\frac{1}{1-\beta}$ is close to 1. Furthermore, from (3.29), the conditional start-service probability of m^{th} packet $\hat{\pi}_j(i,m) \approx \eta_j$ for $m \geq 1$, which results in $X_m \approx \frac{1}{\hat{\mu}}$ ($E[T] \approx \frac{1}{\hat{\mu}}$) and $\sum_{i=2}^{\infty} P_i \sum_{m=1}^{i-1} X_m = \frac{L_q}{\hat{\mu}}$. Then the mean waiting time of *M/MMSP/1* queue (3.48) can be approximately given as:

$$W \approx \frac{R}{1-\frac{\lambda}{\hat{\mu}}} \approx \frac{\frac{1}{2}\lambda E[T^2]}{1-\lambda E[T]}. \tag{3.55}$$

   The right-hand side of (3.55) corresponds to the mean waiting time of an *M/G/1* queue, which sometimes is used to approximately model a wireless channel [7, 12]. The expression (3.55) indicates that this approximation is valid only when β is close to 0. As an example, the mean queue length of the *M/G/1* queue plotted in Fig. 3.10 almost coincides with that of *M/MMSP/1* queue when $\beta = 0.3810$ ($f = 1$).

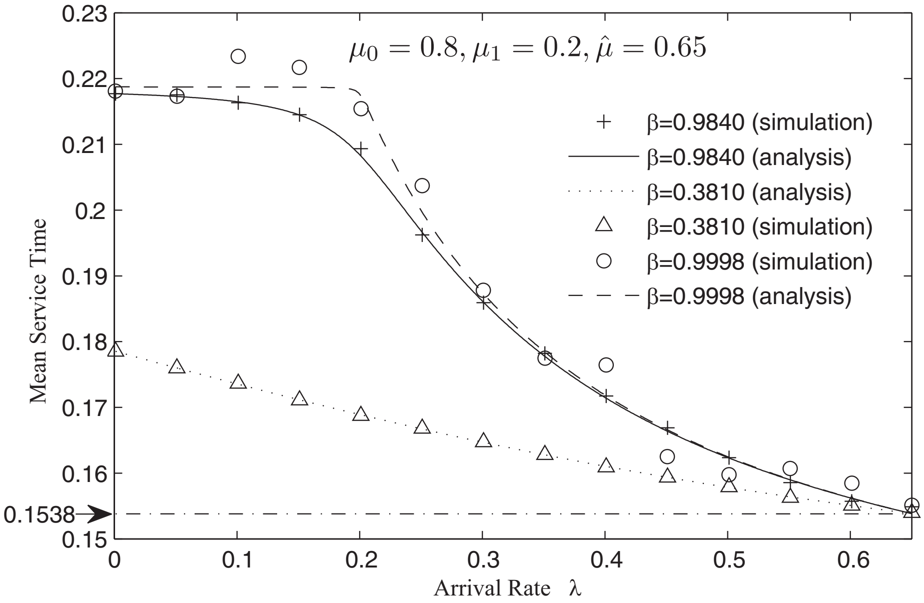

Figure 3.11: Mean service time $E[T]$ of *M/MMSP/1* with different state transition factor β.

Note that a relative smaller state transition factor β corresponds to a smaller $E[T]$ and $\sum_{j=0}^{1} E[T_j](\pi_j - \hat{\pi}_j)$ in the nominator of (3.43), as depicted in Fig. 3.11. We know from (3.26) that $\sum_{j=0}^{1} E[T_j]\pi_j$ is the first moment of service time when λ approaches 0, and the lower bound (0.1538 in Fig. 3.11) is given by $\sum_{j=0}^{1} E[T_j]\eta_j = 1/\hat{\mu}$ when λ approaches ∞. It is obvious that the first moment of service time $E[T]$ is bounded by $\frac{1}{\hat{\mu}} \leq E[T] \leq \sum_{j=0}^{1} E[T_j]\pi_j$.

Our analysis clearly explains the reason that a wireless communication system with a smaller state transition factor β owns a smaller queueing delay. For a given Hybrid ARQ coding scheme, an increase in the packet length leads to a smaller μ_j; while an increase in the Doppler frequency f_D leads to a larger f_j. From the definition of state transition factor β given in (3.27), we know that either of them will generate a smaller β, consequently, a smaller queueing delay. This property also explains the results depicted in Fig. $2 \sim 4$ in [46].

Chapter 4

Generalized Pollaczek-Khinchin Formula for the *M/MMSP/1/K* Queue — Finite-state

The only available method to analyze the average queue length of finite-state Markov channels in the literature is the Matrix Geometric Method [36]. In this chapter, we extend the generalized Pollaczek-Khinchin formula developed in Chapter 3 to Markov channels with finite states.

4.1 Finite-state Markov Model

A N-state Markov chain of a wireless fading channel, Markov modulated service process (MMSP), is shown in Fig. 4.1. The state j ($j = 0, 1, 2, ..., N$) represents that the channel is in state j where the service rate is exponentially distributed with mean μ_j. The state transition rate from state j to $j+1$ is denoted as $f_{j,j+1}$ (> 0). Here the channel states could only change to its adjacent states, because the channel quality of our interest is continuously varying. However, the following derivations

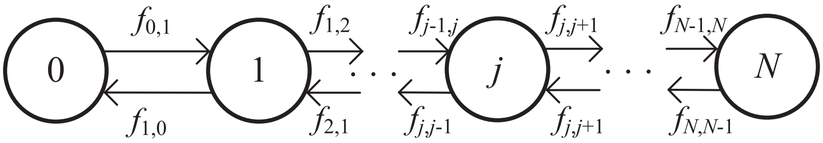

Figure 4.1: Finite-state Markov model.

are still established under the assumption that the channel state could jump to any state j' out of the N states. Then, the infinitesimal generator matrix $\boldsymbol{Q}$ is given by

$$\boldsymbol{Q} = \begin{pmatrix} -f_0 & f_{0,1} & 0 & 0 & 0 & \cdots & 0 \\ f_{1,0} & -f_1 & f_{1,2} & 0 & 0 & \cdots & 0 \\ 0 & f_{2,1} & -f_2 & f_{2,3} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & f_{N-1,N-2} & -f_{N-1} & f_{N-1,N} \\ 0 & 0 & 0 & 0 & \cdots & f_{N,N-1} & -f_N \end{pmatrix}, \tag{4.1}$$

where we denote

$$f_j = \sum_{j' \neq j} f_{j,j'}. \tag{4.2}$$

At time t, the state of the system $X(t)$ is defined as the number of packets in the system, including the one in service, while the channel state is denoted by $V(t)$. The process $\{(X(t), V(t)), t \geq 0\}$ is a Continuous Time Markov Chain (CTMC) with a finite state space $\{(i,j), i = 0, 1, 2, ..., j = 0, 1, ..., N\}$. We denote this queuing model with infinite buffer capacity as $M/MMSP/1$. The steady state probabilities of the $M/MMSP/1$ are defined by:

$$p_{i,j} = \lim_{t \to \infty} Pr\{X(t) = i, V(t) = j\}, \tag{4.3}$$

where $i = 0, 1, 2, ...$ and $j = 0, 1, ..., N$. Thus, the probability that there are i packets in the system in steady state is given by:

$$P_i = \sum_{j=0}^{N} p_{i,j}. \tag{4.4}$$

4.2 Start-service Probability

For a system with finite-state Markov channels, suppose that a new packet arrives into a FIFO buffer and sees i packets in the system. Those packets are sequentially numbered by $m = 0, 1, \cdots, i$, the one in service is numbered 0 and the new arrival is numbered i, as shown in Fig. 3.6. The definitions related to start-service probability $\hat{\pi}_j$ of finite-state Markov channels are independent of the number of channel states and are the same as the ones in two-state Markov channels described in Section 3.3, except that we extend the channel states from 2 to N as following:

- π_j = the steady-state probability that the channel is in state j, for $j = 0, 1, 2, ..., N$. If we denote π_j as a column vector $\boldsymbol{\pi} = [\pi_0, \pi_1, ..., \pi_N]^T$ where the superscript T means the transpose of the matrix, π_j can be obtained by solving

  $$\boldsymbol{Q}^T \boldsymbol{\pi} = \mathbf{0} \tag{4.5}$$

  and

  $$\sum_{j=0}^{N} \pi_j = 1 \tag{4.6}$$

  Then we define the channel capacity $\hat{\mu}$, as

  $$\hat{\mu} = \sum_{j=0}^{N} \pi_j \mu_j. \tag{4.7}$$

  Detailed discussions on the channel capacity can be find in Section 4.3.1.

- $\hat{\pi}_j$ = the probability that an HOL packet's start-service state is in channel state j, for $j = 0, 1, 2, ..., N$. According to the analysis in [32], the following two extreme cases

$$\hat{\pi}_j = \begin{cases} \pi_j & \text{when } \lambda \to 0 \\ \eta_j = \frac{\pi_j \mu_j}{\sum_{j=0}^{N} \pi_j \mu_j} & \text{when } \lambda \to \infty \end{cases} \tag{4.8}$$

are still established for Markov channels with finite states. However, there is no present expression of $\hat{\pi}_j$ in the literature. Here, we express $\hat{\pi}_j$ in forms of π_i and $\hat{\pi}_j(i,m)$ in Theorem 4.1 and from which we define the state transition factor $\beta \in (0,1)$ for finite-state Markov channel. We denote the corresponding column vector as $\hat{\boldsymbol{\pi}} = [\hat{\pi}_0, \hat{\pi}_1, ..., \hat{\pi}_N]^T$ and $\boldsymbol{\eta} = [\eta_0, \eta_1, ..., \eta_N]^T$.

- $\hat{\pi}_j(i,m)$ = the conditional start-service probability of m^{th} packet, which is defined as follows:

$$\hat{\pi}_j(i,m) = P\{\text{start-service state of } m^{th} \text{ packet is } j \\ \mid \text{an arrival sees } i \text{ packets in the system}\}$$

for $i = 0, 1, 2, ...$ and $j = 0, 1, 2, ..., N$. We denote the corresponding column vector as $\hat{\boldsymbol{\pi}}\boldsymbol{(i,m)} = [\hat{\pi}_0(i,m), \hat{\pi}_1(i,m), ..., \hat{\pi}_N(i,m)]^T$.

- $P\{j'|j\}$ = the probability that a packet starts the service in channel state j and finishes the service in channel state j'. We denote the corresponding matrix, called state transition matrix $\hat{\boldsymbol{Q}}$ in this paper, as

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} P\{0|0\} & P\{0|1\} & \cdots & P\{0|N\} \\ P\{1|0\} & P\{1|1\} & \cdots & P\{1|N\} \\ \vdots & \vdots & \ddots & \vdots \\ P\{N|0\} & P\{N|1\} & \cdots & P\{N|N\} \end{pmatrix}. \tag{4.9}$$

4.2.1 State transition matrix

For a packet starts its service in channel state j, the Markov channel may change to its adjacent state before the packet is transmitted or stays in the same channel state till the end of the packet's transmission. For different values of channel state j, the channel state transition during during one packet's service time is given by the following set of equations:

- $j = 0$,

$$\begin{cases} p\{0|0\} = \frac{\mu_0}{f_0+\mu_0} + \frac{f_{0,1}}{f_0+\mu_0} p\{0|1\} \\ p\{j' \neq 0|0\} = \frac{f_{0,1}}{f_0+\mu_0} p\{j'|1\} \end{cases} \tag{4.10a}$$

- $1 \leq j \leq N-1$,

$$\begin{cases} p\{j' = j|j\} = \frac{\mu_j}{f_j+\mu_j} + \frac{f_{j,j-1}}{f_j+\mu_j} p\{j|j-1\} + \frac{f_{j,j+1}}{f_j+\mu_j} p\{j|j+1\} \\ p\{j' \neq j|j\} = \frac{f_{j,j-1}}{f_j+\mu_j} p\{j'|j-1\} + \frac{f_{j,j+1}}{f_j+\mu_j} p\{j'|j+1\} \end{cases} \tag{4.10b}$$

- $j = N$,

$$\begin{cases} p\{j' = N|N\} = \frac{\mu_N}{f_N+\mu_N} + \frac{f_{N,N-1}}{f_N+\mu_N} p\{N|N-1\} \\ p\{j' \neq N|N\} = \frac{f_{N,N-1}}{f_N+\mu_N} p\{j'|N-1\} \end{cases} \tag{4.10c}$$

Multiplying $f_j + \mu_j$ on each side of the equations in (4.10) and expressing (4.10) in forms of matrix, we have

$$\boldsymbol{M} \begin{bmatrix} p\{j'|0\} \\ p\{j'|1\} \\ \vdots \\ p\{j'|N\} \end{bmatrix} = \boldsymbol{e}_{j'}, \tag{4.11}$$

where $\boldsymbol{e}_{j'} = [0, 0, ..., \mu_{j'}, ...]^T$ is a vector with the j'^{th} element be $\mu_{j'}$, and

$$\boldsymbol{M} = \begin{pmatrix} f_0+\mu_0 & -f_{0,1} & 0 & 0 & 0 & \cdots & 0 \\ -f_{1,0} & f_1+\mu_1 & -f_{1,2} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & -f_{N-1,N-2} & f_{N-1}+\mu_{N-1} & -f_{N-1,N} \\ 0 & 0 & 0 & 0 & \cdots & -f_{N,N-1} & f_N+\mu_N \end{pmatrix}. \tag{4.12}$$

Comparing (4.12) with (4.1), we have

$$\boldsymbol{M} = \boldsymbol{D} - \boldsymbol{Q}, \tag{4.13}$$

where we denote

$$\boldsymbol{D} = \mathrm{diag}(\mu_0, \mu_1, ..., \mu_N). \tag{4.14}$$

Then $P\{j'|j\}$ is obtained from Cramer's rule, as

$$P\{j'|j\} = \frac{|\boldsymbol{M_{j',j}}|}{|\boldsymbol{M}|}, \tag{4.15}$$

where $|\boldsymbol{M}|$ is the determinant of the matrix $\boldsymbol{M}$ and $\boldsymbol{M_{j',j}}$ is a matrix obtained from $\boldsymbol{M}$ by replacing the j^{th} column by $\boldsymbol{e}_{j'}$. After some development, $|\boldsymbol{M_{j',j}}|$ is given by

$$|\boldsymbol{M_{j',j}}| = \mu_{j'} C_{j',j}, \tag{4.16}$$

where $C_{j',j}$ is the cofactor of the $(j', j)^{th}$ element of matrix $\boldsymbol{M}$. The corresponding cofactor matrix $\boldsymbol{C} = [C_{j',j}]_{N\times N}$ is key to compute the inverse of $\boldsymbol{M}$, as

$$\boldsymbol{M}^{-1} = \frac{\boldsymbol{C}^T}{|\boldsymbol{M}|}. \tag{4.17}$$

Considering all possible $j', j = 0, 1, 2, ..., N$ of (4.15), we obtain the state transition matrix $\hat{\boldsymbol{Q}}$ by resorting to (4.16) and (4.17)

$$\hat{\boldsymbol{Q}} = \boldsymbol{D}(\boldsymbol{M^T})^{-1}. \tag{4.18}$$

4.2.2 Start service probability

Theorem 4.1. *The probability that the service of an HOL packet of $M/MMSP/1$ starts with channel state j is given by:*

$$\hat{\pi}_j = \sum_i P_i \hat{\pi}_j(i, i), \qquad j = 0, 1, ..., N, \tag{4.19}$$

where

$$\boldsymbol{\hat{\pi}(i, m)} = \boldsymbol{\hat{Q}}^m \boldsymbol{\hat{\pi}(i, 0)} \tag{4.20}$$

and

$$\hat{\pi}_j(i, 0) = p_{i,j}/P_i. \tag{4.21}$$

Proof. From the definition of $\hat{\pi}_i(i, m)$ and $P\{j'|j\}$, we establish the following relationship between $\hat{\pi}_i(i, m)$ and $\hat{\pi}_i(i, m+1)$, as

$$\boldsymbol{\hat{\pi}(i, m+1)} = \boldsymbol{\hat{Q}\hat{\pi}(i, m)} \tag{4.22}$$

From the probability P_i that a newly arrived packet sees i packets in the system and $p_{i,j}$ that a newly arrived packet sees i packets in the system while the current channel is in state j, we obtain the following initial state probability:

$$\hat{\pi}_j(i, 0) = p_{i,j}/P_i. \tag{4.23}$$

Solving recurrence relationship (4.22) for all possible $m \geq 0$ along with the initial condition (4.23), we get

$$\boldsymbol{\hat{\pi}(i, m)} = \boldsymbol{\hat{Q}}^m \boldsymbol{\hat{\pi}(i, 0)} \tag{4.24}$$

Since an arrival that sees i packets in the system starts the service in state j with probability $\hat{\pi}_j(i, i)$, we then obtain the following start-service probability by combining (4.22) and (4.24):

$$\hat{\pi}_j = \sum_i P_i \hat{\pi}_j(i, i) \tag{4.25}$$

□

Theorem 4.2. *The stationary probability vector of the Markov chain characterized by the state transition matrix $\boldsymbol{\hat{Q}}$ is $\boldsymbol{\eta}$. That is*

$$\boldsymbol{\eta} = \boldsymbol{\hat{Q}\eta}. \tag{4.26}$$

Proof. From the definition of inverse matrix, the following equation always hold

$$\boldsymbol{I} = (\boldsymbol{M}^T)^{-1}\boldsymbol{M}^T, \tag{4.27}$$

where $\boldsymbol{I}$ is the identity matrix. Substituting (4.13) into (4.27) and multiplying $\boldsymbol{\pi}$ on both sides of (4.27), we arrive at

$$\begin{aligned}\boldsymbol{\pi} &= (\boldsymbol{M}^T)^{-1}(\boldsymbol{D}-\boldsymbol{Q})^T\boldsymbol{\pi}\\ &= (\boldsymbol{M}^T)^{-1}\boldsymbol{D}\boldsymbol{\pi} - (\boldsymbol{M}^T)^{-1}\boldsymbol{Q}^T\boldsymbol{\pi}\\ &= (\boldsymbol{M}^T)^{-1}\boldsymbol{D}\boldsymbol{\pi} \end{aligned}\tag{4.28}$$

where $\boldsymbol{D}^T = \boldsymbol{D}$ since $\boldsymbol{D}$ is a diagonal matrix and the last equation is due to (4.5). Notice that η_j given in (4.8) could be expressed in forms of π_j, as

$$\boldsymbol{\eta} = \frac{\boldsymbol{D}\boldsymbol{\pi}}{|\boldsymbol{D}\boldsymbol{\pi}|_1}, \tag{4.29}$$

where $|\boldsymbol{D}\boldsymbol{\pi}|_1$ is the 1-norm of the vector $\boldsymbol{D}\boldsymbol{\pi}$ and $|\boldsymbol{D}\boldsymbol{\pi}|_1 = \hat{\mu}$. With (4.29), (4.18) and (4.28), we finally establish the relationship between $\boldsymbol{\eta}$ and $\hat{\boldsymbol{Q}}$ as follows:

$$\begin{aligned}\hat{\boldsymbol{Q}}\boldsymbol{\eta} &= \frac{1}{\hat{\mu}}\boldsymbol{D}(\boldsymbol{M}^{\boldsymbol{T}})^{-1}\boldsymbol{D}\boldsymbol{\pi}\\ &= \frac{1}{\hat{\mu}}\boldsymbol{D}\boldsymbol{\pi}\\ &= \boldsymbol{\eta}. \end{aligned}\tag{4.30}$$

□

If arrival packets always find infinite number of packets waiting in the buffer, then (4.24) and (4.25) reveal that the transmission of a packet is started in state j is approximately given by

$$\lim_{\lambda\to\infty}\hat{\boldsymbol{\pi}} = \lim_{i\to\infty}\hat{\boldsymbol{\pi}}(\boldsymbol{i},\boldsymbol{i}) = \lim_{m\to\infty}\hat{\boldsymbol{\pi}}(\boldsymbol{\infty},\boldsymbol{m}) = \lim_{m\to\infty}\hat{\boldsymbol{Q}}^m\hat{\boldsymbol{\pi}}(\boldsymbol{\infty},\boldsymbol{0}) \tag{4.31}$$

Theorem 4.2 shows that $\boldsymbol{\eta}$ is the stationary probability vector of the Markov chain characterized by the state transition matrix $\hat{\boldsymbol{Q}}$. From the property of Markov chain, we have

$$\lim_{m\to\infty} \hat{\boldsymbol{\pi}}(\boldsymbol{\infty}, \boldsymbol{m}) = \boldsymbol{\eta}. \tag{4.32}$$

Both (4.31) and (4.32) lead to result that $\hat{\pi}_j = \eta_j$ when $\lambda \to \infty$, which is again consistent with (4.8).

On the other hand, in contrast to the expression (4.25) of $\hat{\pi}_j$, the probability π_j that the channel is in state j can be expressed from (4.23) as follows:

$$\pi_j = \sum_i p_{i,j} = \sum_i P_i \hat{\pi}_j(i, 0). \tag{4.33}$$

If the loading is very low, meaning $P_0 \to 1$, then it is clear that both expressions of $\hat{\pi}_j$ and π_j approach $\hat{\pi}_j(0,0)$, which is consistent with (4.8) that $\hat{\pi}_j = \pi_j$ when $\lambda \to 0$. Thus, the expression (4.25) of the start-service probability $\hat{\pi}_j$ covers the two extremes originally derived in [32] as special cases.

4.2.3 State transition factor

The state transition factor β indicates how fast the channel state probability tends to be stable, which is the convergence rate for the Markov chain characterized by the state transition matrix $\hat{\boldsymbol{Q}}$. If $\mu_j > 0$ for all $j = 0, 1, 2, ..., N$, from the definition of $p\{j'|j\}$ we have $p\{j'|j\} \in (0, 1)$. It follows that $\hat{\boldsymbol{Q}}$ is ergodic. Rosenthal [42] shows the following properties related to $\hat{\boldsymbol{Q}}$:

- *Property 1.* The state transition matrix $\hat{\boldsymbol{Q}}$ has one and only one eigenvalue equaling to 1 and the absolute value of every other eigenvalue is less than 1.

  Without loss of generality, we write down the eigenvalues of $\hat{\boldsymbol{Q}}$ as $1 = \xi_0 > |\xi_1| \geq |\xi_2| \geq |\xi_3| \geq \cdots \geq |\xi_N|$. Theorem 4.2 implies that $\boldsymbol{\eta}$ is an eigenvector of $\hat{\boldsymbol{Q}}$ corresponding to the eigenvalue ξ_0.

- *Property 2.* For any initial distribution $\hat{\boldsymbol{\pi}}(\boldsymbol{i}, \boldsymbol{0})$ and $j = 0, 1, 2, ..., N$, there is a constant $C_j > 0$ such that

$$|\hat{\pi}_j(i,m) - \eta_j| \leq C_j m^{J-1} |\xi_1|^{m-J+1}. \tag{4.34}$$

where J is the size of the largest Jordan block of $\hat{\boldsymbol{Q}}$. In particular, if $\hat{\boldsymbol{Q}}$ is diagonalizable, $J = 1$.

The property 2 shows that the converge rate of the Markov chain is highly related to $|\xi_1|$. There is a related work [45] which also studies the importance of the second largest eigenvalue on the convergence rate of Markov chain. Here, we define the state transition factor as

$$\beta = |\xi_1|. \tag{4.35}$$

If one of the service rate in channel state j equals to 0, such as $\mu_0 = 0$, we have $p\{0|j\} = 0$ for all $j = 0, 1, 2, ..., N$ since the packet's service will not end in channel state $j = 0$. Then, the whole first row of $\hat{\boldsymbol{Q}}$ will become all 0s and $\xi_N = 0$. The aforementioned two properties are still established, so is the definition of state transition factor β.

We develop a possible approximation of $\hat{\pi}_j(i,m)$ from the property 2 by combining (4.24), (4.34) and (4.35) as

$$\hat{\pi}_j(i,m) \approx m^{J-1} \beta^{m-J+1} (\hat{\pi}_j(i,0) - \eta_j) + \eta_j. \tag{4.36}$$

If $m >> J$ or $J = 1$, (4.36) could be simplified to

$$\hat{\pi}_j(i,m) \approx \beta^m (\hat{\pi}_j(i,0) - \eta_j) + \eta_j. \tag{4.37}$$

The approximation ignores those components contributed by the other eigenvalues $\xi_j, 2 \leq j \leq N$. If $N = 1$ with no components ignored, then (4.37) exactly holds, which is (3.29) of the two-state Markov channel.

4.3 Generalized Pollaczek-Khinchin Formula for the *M/MMSP/1* Queue

In this section, we derive the generalized P-K formula for the *M/MMSP/1* queue with finite channel states, which shows that the queueing delay is approximately proportional to $\frac{1}{1-\beta}$. It should be noted that the waiting time analysis described in Section 3.4 is independent of the number of states of the Markov channels and we could easily extend the analysis from two states to finite states. Here we reproduce those definitions related to waiting time defined in Section 3.4, as:

- $Q_c =$ The number of packets in the system seen by an arrival ($Q_c \geq 0$).
- $W_i =$ Waiting time of an arrival seeing i ($i \geq 0$) packets in the system.
- $R_i =$ Residual service time of an arrival that sees i ($i \geq 1$) packets in the system. That is, R_i is the remaining time until the completion of current service. If $i = 0$, then the system is empty and there is no residual service time.
- $X_m =$ Service time of the m^{th} ($1 \leq m \leq i-1$) packet in the queue, starting from the first packet behind the head-of-line (HOL) packet, which is in service.

According to the above definitions, the waiting time of an arrival is given by

$$W_i = \begin{cases} R_i + \sum_1^{i-1} X_m & i \geq 2, \\ R_i & i = 1, \\ 0 & i = 0. \end{cases}$$

Taking expectation, we have

$$W = \sum_{i=0}^{\infty} P_i W_i = R + \sum_{i=2}^{\infty} P_i \sum_{m=1}^{i-1} E[X_m | Q_c = i], \tag{4.38}$$

where W is the average waiting time and $R = \sum_{i=1}^{\infty} P_i R_i$ is the mean residual service time. In the rest of this section, we will derive the generalized P-K formula of *M/MMSP/1* queue from the expression (4.38).

4.3.1 Moments of Service time

The derivations of the mean residual service time R and the service time of the m^{th} waiting packet X_m require the first and second moments, $E[T]$ and $E[T^2]$, of the service time of the *M/MMSP/1* queue. They can be obtained from the first and second conditional moments $E[T_j]$ and $E[T_j^2]$ $(j = 0, 1, 2, ..., N)$ defined in [32] as follows:

$$E[T] = \sum_{j=0}^{N} \hat{\pi}_j E[T_j], \tag{4.39a}$$

$$E[T^2] = \sum_{j=0}^{N} \hat{\pi}_j E[T_j^2], \tag{4.39b}$$

where $E[T_j]$ and $E[T_j^2]$ are the first and second conditional moments of service time given that the service begins with state j. If we denote

$$\boldsymbol{E[T]} = \{E[T_0], E[T_2], E[T_2], ..., E[T_N]\}^T$$
$$\boldsymbol{E[T^2]} = \{E[T_0^2], E[T_2^2], E[T_2^2], ..., E[T_N^2]\}^T,$$

then $E[T_j]$ and $E[T_j^2]$ are obtained from

$$\boldsymbol{E[T]} = \boldsymbol{M}^{-1}\mathbf{1}, \tag{4.40a}$$

$$\boldsymbol{E[T^2]} = 2\boldsymbol{M}^{-1}\boldsymbol{E[T]}, \tag{4.40b}$$

where $\boldsymbol{M}$ is defined in (4.12) and $\mathbf{1}$ is an all-ones vector.

According to [32], $E[T_j]$ and $E[T_j^2]$ are obtained as following: Consider an arbitrary packet whose service starts in channel state j. Let T_j be the random variable denoting the total service time for this packet. During the service of this packet, the channel state j may change to $j+1$ (or $j-1$) after time $T_{j,j+1}$ (or $T_{j,j-1}$) when the serve speed alternates, or even stay at j until the serve is finished after time T_μ. Conditioning on steady state, T_μ, $T_{j,j+1}$ and $T_{j,j-1}$ are exponentially distributed with parameters μ_j, $f_{j,j+1}$ and $f_{j,j-1}$ respectively. Then, T_j can be calculated as

$$T_j = min(T_\mu, T_{j,j+1}, T_{j,j-1}) + \begin{cases} 0 & \text{with probability} \quad \frac{\mu_j}{\mu_j+f_j} \\ T_{j,j+1} & \text{with probability} \quad \frac{f_{j,j+1}}{\mu_j+f_j} \\ T_{j,j-1} & \text{with probability} \quad \frac{f_{j,j-1}}{\mu_j+f_j} \end{cases}$$

where $j = 0, 1, 2, ..., N$.

Taking the Laplace Stieltjes transform (LST) on both sides, we get

$$E(e^{-sT_j}) = \frac{\mu_j + f_j}{s + \mu_j + f_j} \left(\frac{\mu_j}{\mu_j + f_j} + \frac{f_{j,j+1}E(e^{-sT_{j+1}}) + f_{j,j-1}E(e^{-sT_{j-1}})}{\mu_j + f_j} \right).$$

Arranging terms, we have

$$(s + \mu_j + f_j)E(e^{-sT_j}) = \mu_j + f_{j,j+1}E(e^{-sT_{j+1}}) + f_{j,j-1}E(e^{-sT_{j-1}}). \tag{4.41}$$

Taking derivative of (4.41) with respect to s and substituting $s = 0$, we have

$$(\mu_j + f_j)E[T_j] - f_{j,j+1}E[T_{j+1}] - f_{j,j-1}E[T_{j-1}] = 1 \tag{4.42}$$

By solving (4.42), we get the conditional first moment of service time $E[T_j]$ for every channel state j. Taking the second derivative of LST in (4.41) with respect to s and substituting $s = 0$ and (4.42), we get

$$(\mu_j + f_j)E[T_j^2] - f_{j,j+1}E[T_{j+1}^2] - f_{j,j-1}E[T_{j-1}^2] = 2E[T_j] \tag{4.43}$$

By solving (4.43), we get $E[T_j^2]$ for every $j = 0, 1, 2, ..., N$. Reorganizing (4.42) and (4.43) in forms of matrix with (4.12), it is straightforward that (4.40) are the solutions of $E[T_j]$ and $E[T_j^2]$.

When the arrival rate goes to infinity, from (4.8) we have

$$\lim_{\lambda\to\infty} E[T] = \sum_{j=0}^{N} \lim_{\lambda\to\infty} \hat{\pi}_j E[T_j] = \sum_{j=0}^{N} \eta_j E[T_j]. \tag{4.44}$$

The following we show that

$$\sum_{j=0}^{N} \eta_j E[T_j] = \tfrac{1}{\hat{\mu}}, \tag{4.45}$$

always holds. Resorting to the matrices,

$$\begin{aligned}
\boldsymbol{\eta}^T \boldsymbol{E}[\boldsymbol{T}] &= \boldsymbol{\eta}^T \boldsymbol{M}^{-1}\mathbf{1} \\
&= \frac{1}{\hat{\mu}}(\boldsymbol{D}\boldsymbol{\pi})^T \boldsymbol{M}^{-1}\mathbf{1} \\
&= \frac{1}{\hat{\mu}}(\boldsymbol{M}^T\boldsymbol{\pi})^T \boldsymbol{M}^{-1}\mathbf{1} \\
&= \frac{1}{\hat{\mu}}\boldsymbol{\pi}^T \boldsymbol{M}\boldsymbol{M}^{-1}\mathbf{1} \\
&= \frac{1}{\hat{\mu}}\boldsymbol{\pi}^T \mathbf{1} \\
&= \frac{1}{\hat{\mu}} \qquad (4.46)
\end{aligned}$$

where (4.40a), (4.29) and (4.28) are used sequentially.

4.3.2 Residual service time

The comparison on residual service time between *M/MMSP/1* and *M/G/1* presented in Section 3.4.2 is independent of the Markov channels states. For *M/MMSP/1* with finite channel states, the mean residual service time is also given by the following expression of

$$R = \tfrac{1}{2}\lambda E[T^2]. \tag{4.47}$$

Except that here the second moment of service time $E[T^2]$ is obtained from (4.39b). Although the equations to compute $E[T^2]$ are all present, the probability P_i that there are i packets in the system is not easy to obtain for $N \geq 2$. One possible way in the literature is using the Matrix Geometric Method, which need large extensive calculation and is not under our consideration.

For the concerned *M/MMSP/1* queueing system, the packets arrive in Poisson process and the service time in each channel state j is exponentially distributed. If a new arriving packet see i ($i > 0$) packets in the system with probability with probability P_i, the residual service time of the packet under service (refer to the 0^{th} packet in Fig. 3.6) will renew. From Theorem 4.1 the 0^{th} packet will restart the service in state j with probability $\hat{\pi}_j(i, 0)$. Then, we get another expression of the mean residual service time

$$R = \sum_{i=1}^{\infty} P_i \sum_{j=0}^{N} E[T_j]\hat{\pi}_j(i, 0) = \sum_{j=0}^{N} E[T_j](\pi_j - p_{0,j}), \tag{4.48}$$

where (4.23) and (4.33) are used.

4.3.3 Generalized P-K formula of Mean Waiting Time

Suppose that for an arrival seeing i packets in the system with probability P_i, the m^{th} packet in the queue will start its service after the previous m packets (including the one in service) were served, as demonstrated in Fig. 3.7. From Theorem 4.1, we know that the m^{th} packet will start the service in state j with probability $\hat{\pi}_j(i, m)$. Then, its averaged service time X_m under the condition that there are i packets in the system when a new packet arrives is given by

$$E[X_m|Q_c = i] = \sum_{j=0}^{N} E[T_j]\hat{\pi}_j(i, m). \tag{4.49}$$

Theorem 4.3. *The mean waiting time of an M/MMSP/1 queue is approximately given by*

$$W \approx \frac{\frac{\lambda}{\hat{\mu}}E[T] + \frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j](\pi_j - \hat{\pi}_j)}{1 - \frac{\lambda}{\hat{\mu}}}. \tag{4.50}$$

Proof. The $\hat{\pi}_j(i, m)$ given by (4.24) need the computation of $\hat{\boldsymbol{Q}}^m$, which provides no physical interpretations of the $M/MMSP/1$. Hence, we adapt its approximation as shown in (4.37). Substituting (4.37) into (4.49), the total time for all $i-1$ packets in queue that have been served is given by the unconditional (4.45):

$$\begin{aligned}
&\sum_{m=1}^{i-1} E[X_m|Q_c=i] = \sum_{j=0}^{N} E[T_j] \sum_{m=1}^{i-1} \hat{\pi}_j(i,m) \\
\approx&\sum_{j=0}^{N} E[T_j] \sum_{m=1}^{i-1} (\beta^m(\hat{\pi}_j(i,0)-\eta_j)+\eta_j) \\
=&\sum_{j=0}^{N} E[T_j] \sum_{m=1}^{i-1} (\beta^m(\hat{\pi}_j(i,0)-\eta_j)) + \sum_{j=0}^{N} E[T_j] \sum_{m=1}^{i-1} \eta_j \\
=&\sum_{j=0}^{N} E[T_j]\left(\hat{\pi}_j(i,0)-\eta_j\right) \sum_{m=1}^{i-1} \beta^m + (i-1)\sum_{j=0}^{N} E[T_j]\eta_j \\
=&\sum_{j=0}^{N} E[T_j]\left(\hat{\pi}_j(i,0)-\eta_j\right) \frac{\beta-\beta^i}{1-\beta} + (i-1)\frac{1}{\hat{\mu}} \\
=&\sum_{j=0}^{N} E[T_j]\left(\hat{\pi}_j(i,0)-\eta_j\right) \frac{1-\beta^i}{1-\beta} - \sum_{j=0}^{N} E[T_j]\left(\hat{\pi}_j(i,0)-\eta_j\right)\frac{1-\beta}{1-\beta} + (i-1)\frac{1}{\hat{\mu}} \\
=&\frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j]\left(\hat{\pi}_j(i,0)-\hat{\pi}_j(i,i)\right) + \frac{i-1}{\hat{\mu}} - \sum_{j=0}^{N} E[T_j]\left(\hat{\pi}_j(i,0)-\eta_j\right), \qquad (4.51)
\end{aligned}$$

Where

$$\begin{aligned}
\left(\hat{\pi}_j(i,0)-\eta_j\right)(1-\beta^i) &= \hat{\pi}_j(i,0)-\eta_j - \left(\hat{\pi}_j(i,0)-\eta_j\right)\beta^i \\
&= \hat{\pi}_j(i,0) - \hat{\pi}_j(i,m).
\end{aligned}$$

Averaging over all possible i, we obtain the following averaged time to serve all packets in queue seeing by an arbitrary arrival:

$$\begin{aligned}
&\sum_{i=2}^{\infty} P_i \sum_{m=1}^{i-1} E[X_m|Q_c=i] \\
\approx&\frac{1}{1-\beta}\sum_{i=2}^{\infty}\sum_{j=0}^{N} E[T_j]\left(P_i\hat{\pi}_j(i,0)-P_i\hat{\pi}_j(i,i)\right) + \sum_{i=2}^{\infty} P_i\frac{i-1}{\hat{\mu}} - \sum_{i=2}^{\infty}\sum_{j=0}^{N} E[T_j]\left(P_i\hat{\pi}_j(i,0)-P_i\eta_j\right) \\
=&\frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j]\left(\pi_j-\hat{\pi}_j\right) + \frac{L_q}{\hat{\mu}} - \sum_{j=0}^{N} E[T_j]\left(\pi_j-\eta_j\right) + \sum_{j=0}^{N} E[T_j](p_{0,j}-P_0\eta_j), \qquad (4.52)
\end{aligned}$$

where the second equation is obtained from (4.23), (4.25) and (4.33) and $L_q = \sum_{i=2}^{\infty}(i-1)P_i$. From Little's Law, we have

$$L_q = \lambda W. \tag{4.53}$$

Substituting (4.48), (4.52) and (4.53) into (4.38), the mean waiting time W given by the following equation

$$\begin{aligned}
W \approx & \sum_{j=0}^{N} E[T_j](\pi_j - p_{0,j}) + \frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j]\left(\pi_j - \hat{\pi}_j\right) + \frac{\lambda W}{\hat{\mu}} \\
& - \sum_{j=0}^{N} E[T_j]\left(\pi_j - \eta_j\right) + \sum_{j=0}^{N} E[T_j](p_{0,j} - P_0\eta_j) \\
= & \frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j]\left(\pi_j - \hat{\pi}_j\right) + \frac{\lambda W}{\hat{\mu}} + \sum_{j=0}^{N} E[T_j]\eta_j - \sum_{j=0}^{N} E[T_j]P_0\eta_j \\
= & \frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j]\left(\pi_j - \hat{\pi}_j\right) + \frac{\lambda W}{\hat{\mu}} + (1-P_0)\sum_{j=0}^{N} E[T_j]\eta_j \\
= & \frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j]\left(\pi_j - \hat{\pi}_j\right) + \frac{\lambda W}{\hat{\mu}} + (1-P_0)\frac{1}{\hat{\mu}} \qquad (4.54)
\end{aligned}$$

Solving (4.54) for W, we obtain the following mean waiting time of the *M/MMSP/1* queue:

$$W \approx \frac{\frac{(1-P_0)}{\hat{\mu}} + \frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j](\pi_j - \hat{\pi}_j)}{1 - \frac{\lambda}{\hat{\mu}}}. \tag{4.55}$$

In fact, the $1 - P_0$ in the nominator of (4.55) is the server utilization $\rho = 1 - P_0$. From Little's law, we have

$$\rho = \lambda E[T]. \tag{4.56}$$

Substituting (4.56) into (4.55), we have (4.50). □

4.4 Three-state MMSP Systems

A three-state Markov modulated service rates model is used to analyze the performance of Hybrid ARQ system [19], where the instantaneous throughput from the source to the destination depends on the current quality of the physical channel. The channel system has three possible states corresponding to every Markov state: a *good* state with high speed data-rate where packets are transmitted successfully on first attempts, a *moderate* state with lower data-rate where packets are transmitted with retransmissions, and a *bad* state with no data-rate where no packets get through. The delay performance is analyzed in [19] by resorting to conditional generating functions methods, similarly to the method we presented in Chapter 3.

The performance of the two-state Gilbert-Elliot model is improved by accounting a diversity of order by deploying two antennas [63]. The signals received at the two antennas fade independently and the channel can be modeled by three states: both channels are *good*, only one of the channels is *good*, and both channels are *bad*.

4.4.1 Model description

The three-state queueing model we analyzed in this paper is proposed in [19]. The system has independent Poisson job and server arrivals (with rates λ_c and λ_s respectively), and independent exponentially-distributed job service time and server life time (with rates μ_c per server and μ_s respectively). The service policy allows only one job in the queue to get served by all servers simultaneously with a First Come First Service (FCFS) manner. The maximum number of servers is 2.

Let $n_c(t)$ and $n_s(t)$ are the number of jobs and servers in the system at time t respectively. Then the process $\{n_c(t), n_s(t)\}$ is a two-dimension birth-death process with a infinite state space $\{(i, j), i = 0, 1, 2, ..., j = 0, 1, 2\}$. At any time t, a job will be served with rate $n_s(t)\mu_c$. The server arrival and depart processes are independent on the number of job process in the system and can be studied by an $M/M/2/2$ process. Fig. 4.2 shows the steady-state rate transition diagram. Let $\{p_{i,j}\}$ be the steady state probabilities, that is

$$p_{i,j} = \lim_{t\to\infty} Pr\{n_c(t) = i, n_s(t) = j\}, \tag{4.57}$$

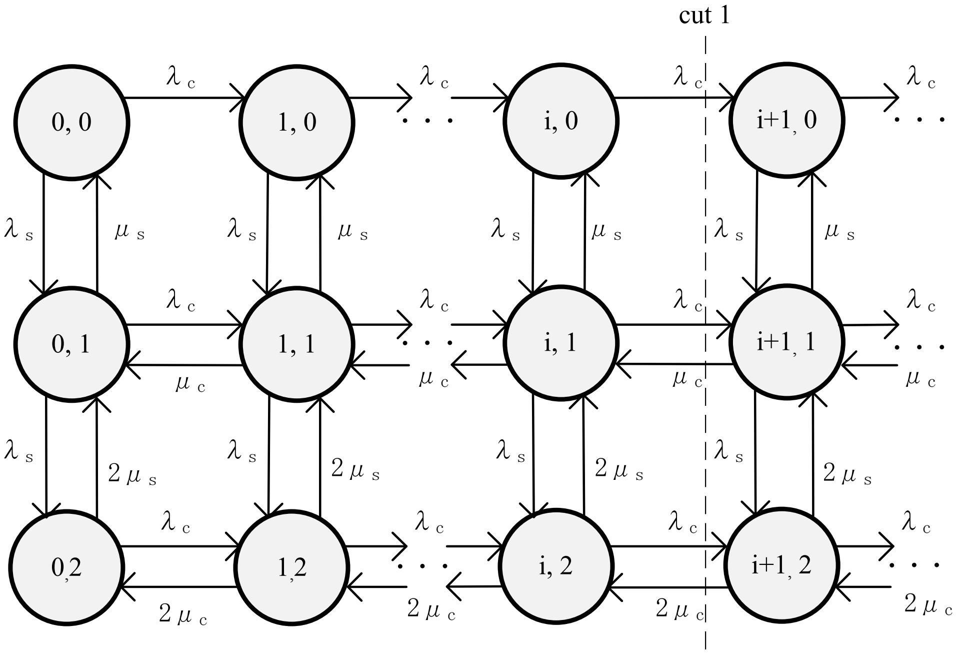

Figure 4.2: Steady-state rate transition diagram for three-state model

where $i = 0, 1, 2, ..., j = 0, 1, 2$. The probability that there are i packets in the system at steady state, $P_i(i = 0, 1, 2, ...)$, is given as

$$P_i = \sum_{j=0}^{2} p_{i,j}. \tag{4.58}$$

We have the set of balance equations from the state transition rate diagram shown in Fig. 4.2, as

$$(\lambda_c + \lambda_s)p_{0,0} = \mu_s p_{0,1} \tag{4.59a}$$

$$(\lambda_c + \lambda_s)p_{i,0} = \lambda_c p_{i-1,0} + \mu_s p_{i,1} \tag{4.59b}$$

$$(\lambda_c + \lambda_s + \mu_s)p_{0,1} = \lambda_s p_{0,0} + 2\mu_s p_{0,2} + \mu_c p_{1,1} \tag{4.59c}$$

$$(\lambda_c + \lambda_s + \mu_s + j\mu_c)p_{i,1} = \lambda_s p_{i,0} + \lambda_c p_{i-1,1} + 2\mu_s p_{i,2} + 1\mu_c p_{i+1,1} \tag{4.59d}$$

$$(\lambda_c + 2\mu_s)p_{0,2} = \lambda_s p_{0,1} + 2\mu_c p_{1,2} \tag{4.59e}$$

$$(\lambda_c + 2\mu_s + 2\mu_c)p_{i,2} = \lambda_s p_{i,1} + \lambda_c p_{i-1,2} 2\mu_c p_{i+1,2} \tag{4.59f}$$

$$i = 1, 2, 3, ...$$

The probability that system is in state j (there are j servers in the system), π_j , is given as

$$\pi_j = \sum_{i=0}^{\infty} p_{i,j}. \tag{4.60}$$

From the state transition diagram in Fig. 4.2, π_j can be solved as

$$\pi_0 = \frac{1}{1+\rho_s+\rho_s^2/2}, \quad \pi_1 = \frac{\rho_s}{1+\rho_s+\rho_s^2/2} \quad \text{and} \quad \pi_2 = \frac{\rho_s^2}{2(1+\rho_s+\rho_s^2/2)} \tag{4.61}$$

where $\rho_s = \lambda_s/\mu_s$. The balance equation with respect to the dashed line cut 2 is

$$\lambda_c P_i = \sum_{j=0}^{2} j\mu_c p_{i+1,j}, \tag{4.62}$$

for all $i = 0, 1, 2, ...$ Sum all these equations together over i, we get

$$\lambda_c = \sum_{i=0}^{\infty}\sum_{j=0}^{2} j\mu_c p_{i+1,j} = \mu_c \sum_{j=0}^{2} j(\pi_j - p_{0,j}). \tag{4.63}$$

The left hand side of (4.63) means the net input mean arrival rate, while the right hand side is the system throughput. Let

$$\hat{\mu} = \sum_{j=0}^{2} \pi_j j \mu_c \tag{4.64}$$

be the channel capacity. Then (4.63) can be rewritten as

$$\hat{\mu} - \lambda_c = \mu_c \pi_1 + 2\mu_c \pi_2. \tag{4.65}$$

For steady state, all the probabilities $p_{i,j}$ should be positive. From (4.65), the necessary condition for system stability is:

$$\lambda_c < \hat{\mu}. \tag{4.66}$$

4.4.2 Start-service probability

With respect to the general $M/MMSP/1$ model we proposed in Section 4.1, for the three-state Markov modulated queueing system we have $\lambda = \lambda_c$, $\mu_j = j\mu_c$,

$f_{j,j+1} = \lambda_s$ and $f_{j,j-1} = j\mu_s$. the infinitesimal generator matrix $\boldsymbol{Q}$ is given by

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda_s & \lambda_s & 0 \\ \mu_s & -\lambda_s - \mu_s & \lambda_s \\ 0 & 2\mu_s & 2\mu_s \end{pmatrix}. \tag{4.67}$$

The probabilities related to start-service probability $\hat{\pi}_j$ of $M/MMSP/1$ are listed as following:

- $\pi_j = \frac{1}{1+\rho_s+\rho_s^2/2}\frac{\rho_s^j}{j!}$. And the channel capacity

$$\hat{\mu} = \sum_{j=0}^{3} \pi_j \mu_j = \frac{\rho_s \mu_c (1+\rho_s)}{1+\rho_s+\rho_s^2/2}. \tag{4.68}$$

- $\hat{\pi}_j$ = the probability that an HOL packet's start-service state is in channel state j, for $j = 0, 1, 2$. And its two extreme cases

$$\hat{\pi}_j = \begin{cases} \pi_j & \text{when } \lambda_c \to 0 \\ \eta_j = \frac{\pi_j \mu_j}{\sum_{j=0}^{3} \pi_j \mu_j} = \frac{j\mu_c \pi_j}{\hat{\mu}} & \text{when } \lambda_c \to \infty \end{cases} \tag{4.69}$$

specifically, the values of η_j are

$$\eta_0 = 0, \quad \eta_1 = \frac{\mu_s}{\lambda_s + \mu_s}, \quad \text{and} \quad \eta_2 = \frac{\lambda_s}{\lambda_s + \mu_s}. \tag{4.70}$$

- $P\{j'|j\}$ = the probability that a packet starts the service in channel state j and finishes the service in channel state j'. We denote the corresponding matrix, called state transition matrix $\hat{\boldsymbol{Q}}$ in this paper, as

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} P\{0|0\} & P\{0|1\} & P\{0|3\} \\ P\{1|0\} & P\{1|1\} & P\{1|3\} \\ P\{3|0\} & P\{3|1\} & P\{3|3\} \end{pmatrix}. \tag{4.71}$$

State transition factor

The diagonal matrix $\boldsymbol{D}$ of the three-state Markov modulated queueing system is

$$\boldsymbol{D} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mu_c & 0 \\ 0 & 0 & 2\mu_c \end{pmatrix} \tag{4.72}$$

From (4.13), we have the matrix $\boldsymbol{M}$, as

$$\boldsymbol{M} = \boldsymbol{D} - \boldsymbol{Q} = \begin{pmatrix} \lambda_s & -\lambda_s & 0 \\ -\mu_s & \lambda_s + \mu_s + \mu_c & -\lambda_s \\ 0 & -2\mu_s & 2(\mu_s + \mu_c) \end{pmatrix}. \tag{4.73}$$

From (4.15), we solve all those values of $P\{j'|j\}$, listed as following:

$$p\{0|0\} = \frac{|\boldsymbol{M}_{0,0}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|} \begin{vmatrix} 0 & -\lambda_s & 0 \\ 0 & \lambda_s + \mu_s + \mu_c & -\lambda_s \\ 0 & -2\mu_s & 2(\mu_s + \mu_c) \end{vmatrix} = 0 \tag{4.74a}$$

$$p\{0|1\} = \frac{|\boldsymbol{M}_{0,1}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|} \begin{vmatrix} \lambda_s & 0 & 0 \\ -\mu_s & 0 & -\lambda_s \\ 0 & 0 & 2(\mu_s + \mu_c) \end{vmatrix} = 0 \tag{4.74b}$$

$$p\{0|2\} = \frac{|\boldsymbol{M}_{0,2}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|} \begin{vmatrix} \lambda_s & -\lambda_s & 0 \\ -\mu_s & \lambda_s + \mu_s + \mu_c & 0 \\ 0 & -2\mu_s & 0 \end{vmatrix} = 0 \tag{4.74c}$$

$$\begin{aligned} p\{1|0\} &= \frac{|\boldsymbol{M}_{1,0}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|} \begin{vmatrix} 0 & -\lambda_s & 0 \\ \mu_c & \lambda_s + \mu_s + \mu_c & -\lambda_s \\ 0 & -2\mu_s & 2(\mu_s + \mu_c) \end{vmatrix} \\ &= -\frac{\mu_c}{|\boldsymbol{M}|} \begin{vmatrix} -\lambda_s & 0 \\ -2\mu_s & 2(\mu_s + \mu_c) \end{vmatrix} = \frac{\mu_c}{|\boldsymbol{M}|} C_{1,0} = \frac{\mu_c + \mu_s}{\lambda_s + \mu_s + \mu_c} \end{aligned} \tag{4.74d}$$

$$p\{1|1\} = \frac{|\boldsymbol{M}_{1,1}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & 0 & 0 \\ -\mu_s & \mu_c & -\lambda_s \\ 0 & 0 & 2(\mu_s+\mu_c) \end{vmatrix} \tag{4.74e}$$

$$= \frac{\mu_c}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & 0 \\ 0 & 2(\mu_s+\mu_c) \end{vmatrix} = \frac{\mu_c}{|\boldsymbol{M}|}C_{1,1} = \frac{\mu_c+\mu_s}{\lambda_s+\mu_s+\mu_c} \tag{4.74f}$$

$$p\{1|2\} = \frac{|\boldsymbol{M}_{1,2}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & -\lambda_s & 0 \\ -\mu_s & \lambda_s+\mu_s+\mu_c & \mu_c \\ 0 & -2\mu_s & 0 \end{vmatrix} \tag{4.74g}$$

$$= -\frac{\mu_c}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & -\lambda_s \\ 0 & -2\mu_s \end{vmatrix} = \frac{\mu_c}{|\boldsymbol{M}|}C_{1,2} = \frac{\mu_s}{\lambda_s+\mu_s+\mu_c} \tag{4.74h}$$

$$p\{2|0\} = \frac{|\boldsymbol{M}_{1,2}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|}\begin{vmatrix} 0 & -\lambda_s & 0 \\ 0 & \lambda_s+\mu_s+\mu_c & -\lambda_s \\ 2\mu_c & -2\mu_s & 2(\mu_s+\mu_c) \end{vmatrix} \tag{4.74i}$$

$$= \frac{2\mu_c}{|\boldsymbol{M}|}\begin{vmatrix} -\lambda_s & 0 \\ \lambda_s+\mu_s+\mu_c & -\lambda_s \end{vmatrix} = \frac{2\mu_c}{|\boldsymbol{M}|}C_{1,2} = \frac{\lambda_s}{\lambda_s+\mu_s+\mu_c} \tag{4.74j}$$

$$p\{2|1\} = \frac{|\boldsymbol{M}_{1,2}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & 0 & 0 \\ -\mu_s & 0 & -\lambda_s \\ 0 & 2\mu_c & 2(\mu_s+\mu_c) \end{vmatrix} \tag{4.74k}$$

$$= -\frac{2\mu_c}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & 0 \\ -\mu_s & -\lambda_s \end{vmatrix} = \frac{2\mu_c}{|\boldsymbol{M}|}C_{2,1} = \frac{\lambda_s}{\lambda_s+\mu_s+\mu_c} \tag{4.74l}$$

$$p\{2|2\} = \frac{|\boldsymbol{M}_{1,2}|}{|\boldsymbol{M}|} = \frac{1}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & -\lambda_s & 0 \\ -\mu_s & \lambda_s+\mu_s+\mu_c & 0 \\ 0 & -2\mu_s & 2\mu_c \end{vmatrix} \tag{4.74m}$$

$$= \frac{2\mu_c}{|\boldsymbol{M}|}\begin{vmatrix} \lambda_s & -\lambda_s \\ -\mu_s & \lambda_s+\mu_s+\mu_c \end{vmatrix} = \frac{2\mu_c}{|\boldsymbol{M}|}C_{2,2} = \frac{\lambda_s+\mu_c}{\lambda_s+\mu_s+\mu_c} \tag{4.74n}$$

(4.74o)

where

$$|\boldsymbol{M}| = \begin{vmatrix} \lambda_s & -\lambda_s & 0 \\ -\mu_s & \lambda_s + \mu_s + \mu_c & -\lambda_s \\ 0 & -2\mu_s & 2(\mu_s + \mu_c) \end{vmatrix} = 2\lambda_s\mu_c(\lambda_s + \mu_s + \mu_c). \tag{4.75}$$

Then we have the state transition matrix

$$\begin{aligned} \hat{\boldsymbol{Q}} &= \frac{1}{|\boldsymbol{M}|} \begin{pmatrix} 0 & 0 & 0 \\ \mu_c C_{1,0} & \mu_c C_{1,1} & \mu_c C_{1,2} \\ 2\mu_c C_{2,0} & 2\mu_c C_{2,1} & 2\mu_c C_{2,2} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ \frac{\mu_c+\mu_s}{\lambda_s+\mu_s+\mu_c} & \frac{\mu_c+\mu_s}{\lambda_s+\mu_s+\mu_c} & \frac{\mu_s}{\lambda_s+\mu_s+\mu_c} \\ \frac{\lambda_s}{\lambda_s+\mu_s+\mu_c} & \frac{\lambda_s}{\lambda_s+\mu_s+\mu_c} & \frac{\lambda_s+\mu_c}{\lambda_s+\mu_s+\mu_c} \end{pmatrix} \end{aligned} \tag{4.76}$$

which is consistent with (4.18). Denote the three eigenvalues of $\hat{\boldsymbol{Q}}$ as $1 = \xi_0 > |\xi_1| \geq |\xi_2|$, which can be obtained from (4.76), as

$$\xi_0 = 1, \qquad \xi_1 = \frac{\mu_c}{\lambda_s + \mu_s + \mu_c}, \qquad \text{and} \qquad \xi_2 = 0.$$

From (4.35), we define the state transition factor

$$\beta = |\xi_1| = \frac{\mu_c}{\lambda_s + \mu_s + \mu_c}. \tag{4.77}$$

Start service probability

The Theorem 4.1 provides expressions to calculate exact value of the start service probability. However, the initial channel state probability $\pi_j^{(0)}$ which is essential to compute the start service probability $\hat{\pi}_j$ is unknown variables. It could only be calculated by resorting the Matrix Geometric Method in the literature. Here we introduce a Liner Approximation method to approximate $\hat{\pi}_j$.

- *Linear Approximation.* [32] shows that simple expressions of $\hat{\pi}_j$ can be obtained under two extreme cases as shown in (4.69): $\hat{\pi}_j = \pi_j$ when $\lambda_c \to 0$ and $\hat{\pi}_j = \eta_j$ when $\lambda_c \to \infty$. We take an linear approximation that if a packet arrives when system is idle (with probability $1 - \rho$), it will start its service in state j with probability π_j; if a packet arrives when system is busy (with probability ρ), it will start its service in state j with probability η_j. That is

$$\hat{\pi}_j \approx (1 - \rho)\pi_j + \rho\eta_j, \quad \text{for } j = 1, 2. \tag{4.78}$$

Because there is no service rate in channel state $j = 0$, a packet will start its service in state $j = 0$ only if it arrives in state 0 while system is idle. That is

$$\hat{\pi}_j = p_{0,0}, \tag{4.79}$$

which could be confirmed from (4.76) since the first element of $\boldsymbol{D}$ is 0. In order to handle the special case when $j = 0$, we have to introduce another approximation

$$p_{0,j} \approx C_j(1 - \rho)\pi_j, \quad \text{for } j = 1, 2. \tag{4.80}$$

where C_j is a normalization constant decided by

$$1 - \rho = \sum_{j=0}^{3} p_{0,j}. \tag{4.81}$$

If the arrival rate $\lambda_c \to 0$, there is no packet in the system. It follows $p_{0,j} \to \pi_j$ and $\rho \to 0$, which is (4.80); If the arrival rate $\lambda_c \to \infty$, both $p_{0,j}$ and $1 - \rho$ tends to be 0, as $p_{0,j} \to 0$ and $\rho \to 1$. The left and right sides of (4.80) are balanced. From (4.94a), (4.81) and (4.80), we have

$$p_{0,0} \approx \frac{\mu_s}{\lambda_c + \lambda_s} C_j(1 - \rho)\pi_1 = \frac{\lambda_s}{\lambda_c + \lambda_s} C_j(1 - \rho)\pi_0. \tag{4.82}$$

For simplicity we assume that all C_j have the same value, which gives the following one possible approximation of $p_{0,j}$, as

$$\begin{cases} p_{0,0} \approx \frac{\frac{\lambda_s}{\lambda_c+\lambda_s}}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_0 \\ p_{0,j} \approx \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_j \quad \text{for } j=1,2. \end{cases} \tag{4.83}$$

By combining (4.78) and (4.83), we finally establish the Linear Approximation of the start service probability

$$\begin{cases} \hat{\pi}_0 = \frac{1-\frac{\lambda_c}{\lambda_c+\lambda_s}}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_0 \\ \hat{\pi}_j = \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_j + \rho\eta_j, \quad \text{for } j=1,2. \end{cases} \tag{4.84}$$

where the server utilization ρ is analyzed during service time analysis in Section 4.4.3.

4.4.3 Delay analysis

The Linear Approximation of the start service probability $\hat{\pi}_j$ requires the server utilization ρ. In the following sections, we first show how to obtain the ρ from conditional moment of service time $E[T_j]$ and analyze the delay on different server variations, which is verified by simulations.

Moments of service time

In Section 4.3.1, we show that the closed form expressions of the first and second conditional moments of service time, $E[T_j]$ and $E[T_j^2]$, are derived from (4.40) as

$$E[T_0] = \frac{1}{\lambda_s} + E[T_1] \tag{4.85a}$$

$$E[T_1] = \frac{1}{\mu_c} + \frac{\frac{\mu_s}{\lambda_s}(\mu_s+\mu_c) - \frac{\lambda_s}{2}}{\mu_c(\lambda_c+\mu_c+\mu_s)} \tag{4.85b}$$

$$E[T_2] = \frac{1}{2(\mu_c+\mu_s)} + \frac{\mu_s}{\mu_c+\mu_s}E[T_1] \tag{4.85c}$$

and

$$E[T_0^2] = 2\frac{1}{\lambda_s}E[T_0] + E[T_1^2] \tag{4.86a}$$

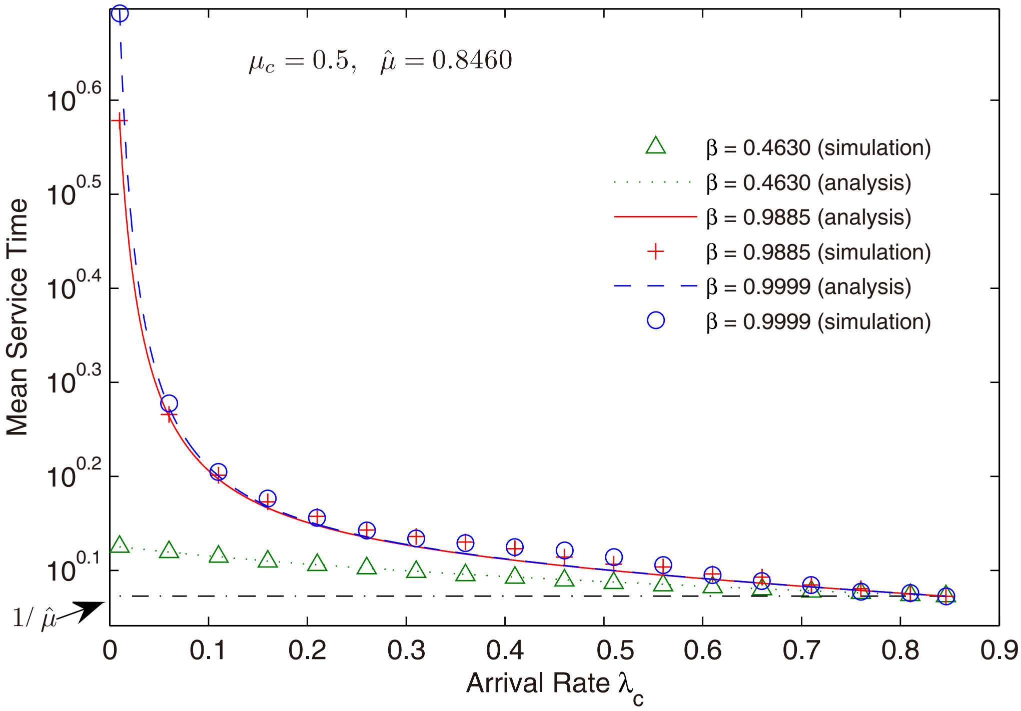

Figure 4.3: Mean service time of three-state queueing model for different arrival rates with different state transition factor β

$$E[T_1^2] = 2\frac{\mu_c + \mu_s}{\mu_c(\lambda_s + \mu_s + \mu_c)}\left(\frac{\mu_s}{\lambda_s}E[T_0] + E[T_1] + \frac{\lambda_s}{2(\mu_s + \mu_c)}E[T_2]\right) \tag{4.86b}$$

$$E[T_2^2] = \frac{1}{\mu_c + \mu_s}E[T_2] + \frac{\mu_s}{\mu_c + \mu_s}E[T_1^2] \tag{4.86c}$$

Then the first and second moments, $E[T]$ and $E[T^2]$, of the service time of the three-state Markov modulated queueing model is obtained from (4.39) as:

$$E[T] = \sum_{j=0}^{2} \hat{\pi}_j E[T_j], \tag{4.87a}$$

$$E[T^2] = \sum_{j=0}^{2} \hat{\pi}_j E[T_j^2], . \tag{4.87b}$$

The first and second moments of service time calculated from our Linear approximately start-service probability fits the simulation results as shown in Fig. 4.3 and Fig. 4.4. The arrival rate λ_c cannot exceed the channel capacity $\hat{\mu}$ for the stability of

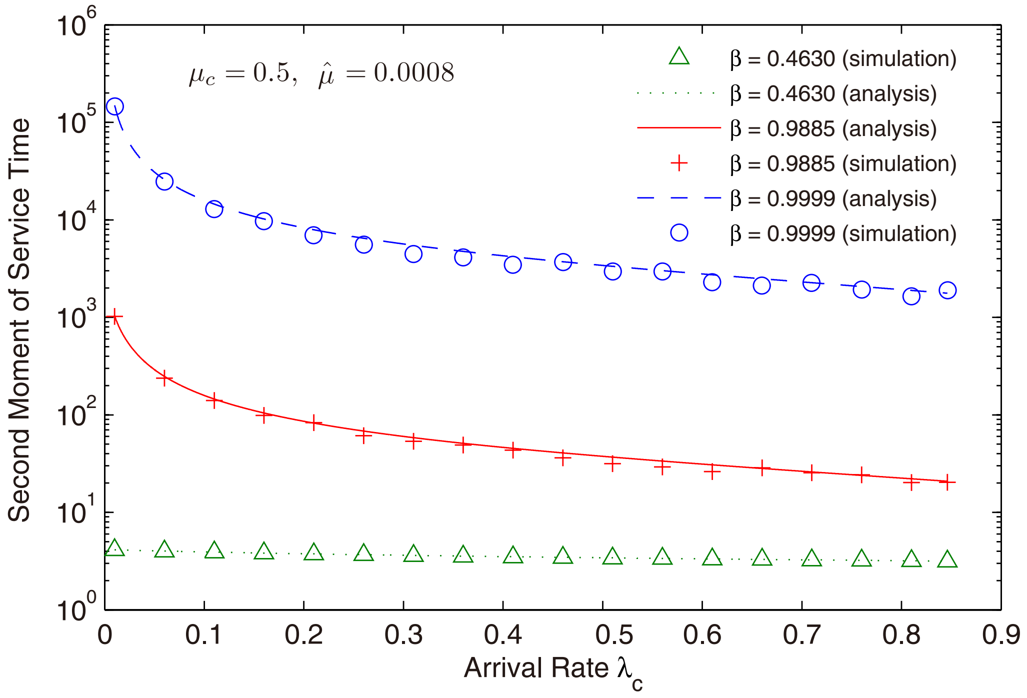

Figure 4.4: The second moment of service time of three-state queueing model for different arrival rates with different state transition factor β

queueing system. As depicted in Fig. 3.11, a relative smaller state transition factor β corresponds to a smaller $E[T]$ and $\sum_{j=0}^{1} E[T_j](\pi_j - \hat{\pi}_j)$ which is essential to the mean waiting time as shown later in (4.91). We know from (4.69) that $\sum_{j=0}^{1} E[T_j]\pi_j$ is the first moment of service time when λ_c approaches 0, and the lower bound (0.1538 in Fig. 4.3) is given by $\sum_{j=0}^{1} E[T_j]\eta_j = 1/\hat{\mu}$ when λ_c approaches ∞. It is obvious that the first moment of service time $E[T]$ is bounded by $\frac{1}{\hat{\mu}} \leq E[T] \leq \sum_{j=0}^{1} E[T_j]\pi_j$. The second moments of service time diverge from each other for different start transition factor β and will not converge to the same lower bound as the arrival rate λ_c increases.

Server utilization ρ

The server utilization ρ is derived from the Little's Law, given as

$$\rho = \lambda_c E[T]. \tag{4.88}$$

With the Linear Approximation $\hat{\pi}_j$ given in (4.84), we have the mean service time from (4.39), as

$$\begin{aligned} E[T] =& \sum_{j=0}^{3} E[T_j]\hat{\pi}_j \\ \approx& \frac{1-\rho}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{3} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right) + \rho\sum_{j=0}^{3} E[T_j]\eta_j \\ =& \frac{1-\rho}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{3} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right) + \rho\frac{1}{\hat{\mu}} \end{aligned} \tag{4.89}$$

where $\sum_{j=0}^{3} E[T_j]\eta_j = 1/\hat{\mu}$.

Solving the joint equations (4.88) and (4.89) for ρ with , we get an approximation of the server utilization

$$\rho \approx \frac{\frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{3} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right)}{\frac{1}{\lambda_c} + \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{3} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right) - \frac{1}{\hat{\mu}}}. \tag{4.90}$$

Mean queue length

From Theorem 4.50, the mean waiting time of the three-state Markov modulated queue model is approximately given by

$$W = \frac{\frac{\lambda_c}{\hat{\mu}}E[T] + \frac{1}{1-\beta}\sum_{j=0}^{2} E[T_j](\pi_j - \hat{\pi}_j)}{1-\frac{\lambda_c}{\hat{\mu}}}. \tag{4.91}$$

The generalized P-K formula explicitly expresses the impact of the state transition factor on the performance of wireless channels. For a given fixed channel capacity $\hat{\mu}$, the generalized P-K formula (4.91) reveals that the mean waiting time is greatly

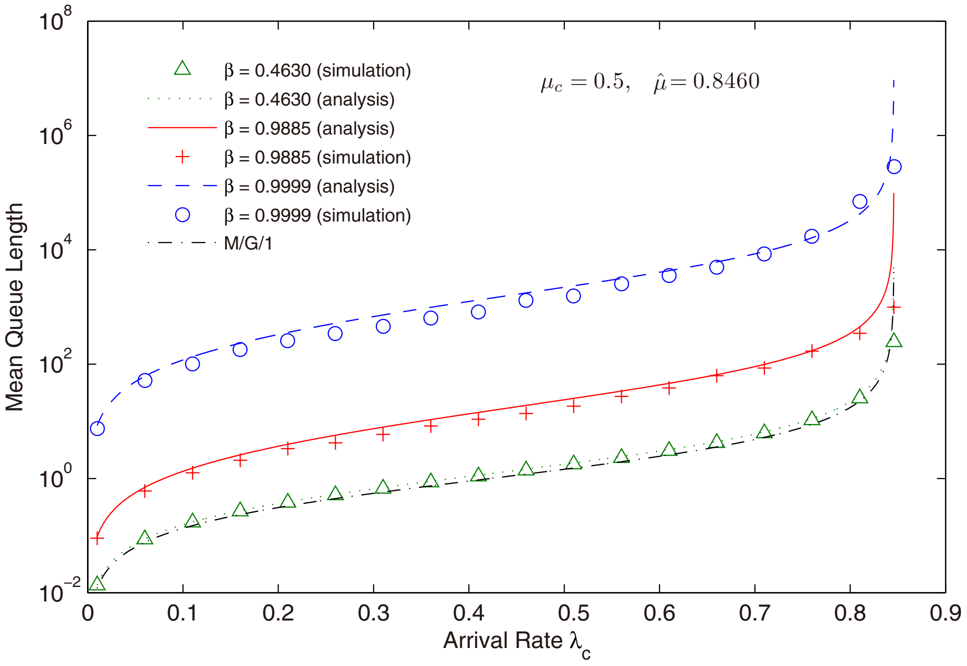

Figure 4.5: Mean queue length of three-state queueing model and $M/G/1$ for different arrival rates with with different state transition factor β

affected by the state transition factor β. This point is illustrated in Fig. 4.5 by the following mean queue length for different β:

$$L = \lambda_c(E[T] + W)$$

Similarly as the results obtained in two-state Markov channel, both analytical and simulation results show that this mean queue length of *M/MMSP/1* with three-state Markov channel is very sensitive to the state transition factor β. When $\beta \to 1$, the mean queue length is dominated by the term $\frac{1}{1-\beta}\sum_{j=0}^{2} E[T_j](\pi_j - \hat{\pi}_j)$. When $\beta \to 0$, the mean queue length is approximately given by

$$W \approx \frac{\frac{\lambda_c}{\hat{\mu}}E[T]}{1 - \frac{\lambda}{\hat{\mu}}},$$

which is close to the mean waiting time of an *M/G/1* queue with mean service time $\hat{\mu}$.

4.5 Peer-to-peer Systems

A general P2P queueing model is proposed by Li et al. [27], where the system has independent Poisson job and server arrivals (with rates λ_c and λ_s respectively), and independent exponentially-distributed job service time and server life time (with rates μ_c per server and μ_s respectively). Perel and Yechiali [37] analyze a similar queueing systems of Markov models with applications in computer networks, where the system is comprised of two connected $M/M/-/-$ type queues with customers of one queue act as servers for the other queue.

In this section, we apply our generalized Pollaczek-Khinchin formula methods to study the Markovian queueing model proposed in [27, 37]. We show that the relationship between server dynamics and waiting time can be characterized by the state transition factor β.

4.5.1 Model description

Here we reproduce the P2P queueing model proposed by [27, 37], as: The system has independent Poisson job and server arrivals (with rates λ_c and λ_s respectively), and independent exponentially-distributed job service time and server life time (with rates μ_c per server and μ_s respectively). The service policy allows only one job in the queue to get served by all servers simultaneously with a First Come First Service (FCFS) manner. Let $n_c(t)$ and $n_s(t)$ are the number of jobs and servers in the system at time t respectively. Then the process $\{n_c(t), n_s(t)\}$ is a two-dimension birth-death process with a infinite state space $\{(i,j), i = 0,1,2,..., j = 0,1,2,..., N\}$. The server number $n_s(t)$ is limited by N, because in a real network system the service rate is constrained by downlink capacity limitation of end users. At any time t, a job will be served with rate $n_s(t)\mu_c$. The server arrival and depart processes are independent on the number of job process in the system and can be studied by an $M/M/\infty$

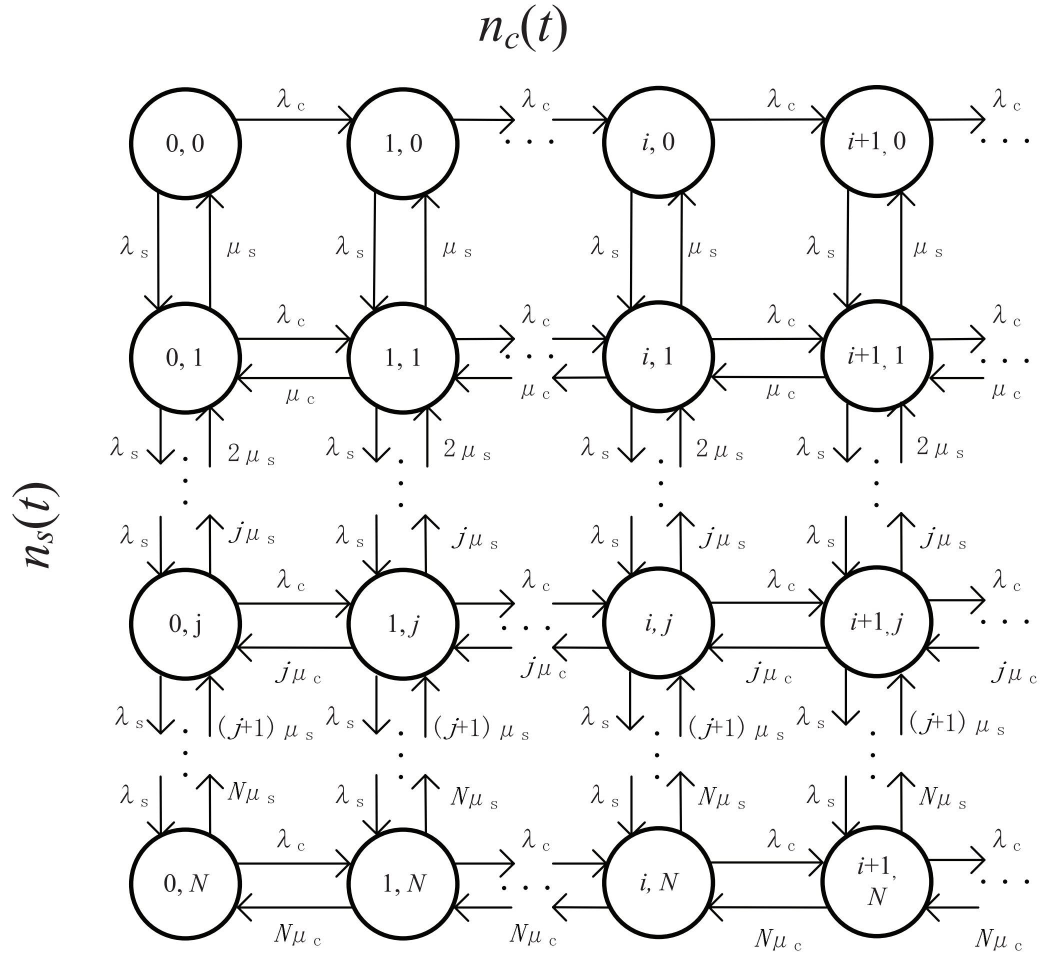

Figure 4.6: Steady-state rate transition diagram for P2P model

process. We model the queueing system of this p2p system as $M/MMSP/1$ and the corresponding steady-state rate transition diagram is shown in Fig. 4.6. The steady state probabilities of the $M/MMSP/1$ are defined by:

$$p_{i,j} = \lim_{t\to\infty} Pr\{n_c(t) = i, n_s(t) = j\}, \tag{4.92}$$

where $i = 0, 1, 2, ...$ and $j = 0, 1, ..., N$. Thus, the probability that there are i packets in the system in steady state is given by:

$$P_i = \sum_{j=0}^{\infty} p_{i,j}. \tag{4.93}$$

The set of balance equations is given by

$$(\lambda_c + \lambda_s)p_{0,0} = \mu_s p_{0,1} \tag{4.94a}$$

$$(\lambda_c + \lambda_s)p_{i,0} = \lambda_c p_{i-1,0} + \mu_s p_{i,1} \tag{4.94b}$$

$$(\lambda_c + \lambda_s + j\mu_s)p_{0,j} = \lambda_s p_{0,j-1} + (j+1)\mu_s p_{0,j+1} + j\mu_c p_{1,j} \tag{4.94c}$$

$$(\lambda_c + \lambda_s + j\mu_s + j\mu_c)p_{i,j} = \lambda_s p_{i,j-1} + \lambda_c p_{i-1,j} + (j+1)\mu_s p_{i,j+1} + j\mu_c p_{i+1,j} \tag{4.94d}$$

$$(\lambda_c + N\mu_s)p_{0,N} = \lambda_s p_{0,N-1} + N\mu_c p_{1,N} \tag{4.94e}$$

$$(\lambda_c + N\mu_s + N\mu_c)p_{i,N} = \lambda_s p_{i,N-1} + \lambda_c p_{i-1,N} + N\mu_c p_{i+1,N} \tag{4.94f}$$

$$i = 1, 2, 3, ..., j = 1, 2, ..., N-1.$$

Let π_j be the probability that system is in state j (there are j servers in the system), we have

$$\pi_j = \sum_{i=0}^{\infty} p_{i,j}. \tag{4.95}$$

From the rate transition diagram shown in Fig. 4.6, the balance equation with respect to the dashed line cut 1 is

$$\lambda_s \pi_j = (j+1)\mu_s \pi_{j+1}, \tag{4.96}$$

for all $j = 0, 1, 2, ..., N$. With $\sum_0^N \pi_j = 1$, π_j can be solved and given as

$$\pi_0 = e^{-\rho_s} \quad \text{and} \quad \pi_j = e^{-\rho_s}\frac{\rho_s^j}{j!}, \tag{4.97}$$

where $\rho_s = \lambda_s/\mu_s$. For convenience, we have assumed a large N such that $\pi_N \to 0$ in the following analysis. If N is a comparative small number and π_N can not be ignored, we only need to add a normalization constant to π_j given in (4.97). (4.97) is the probability distribution for $M/M/\infty$ process. The average number of servers $E[L_s]$ in the P2P system is

$$E[L_s] = \sum_{j=0}^{\infty} j\pi_j = \rho_s. \tag{4.98}$$

The balance equation with respect to the dashed line cut 2 is

$$\lambda_c P_i = \sum_{j=0}^{N} j\mu_c p_{i+1,j}, \tag{4.99}$$

for all $i = 0, 1, 2, \ldots.$ Sum all these equations together over i, we get

$$\lambda_c = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} j\mu_c p_{i+1,j} = \mu_c \sum_{j=0}^{\infty} j(\pi_j - p_{0,j}). \tag{4.100}$$

The left hand side of (4.100) means the net input mean arrival rate, while the right hand side is the system throughput (the mean number of jobs served by this P2P system per unit time). Substituting (4.98) into (4.100), it yields

$$\rho_c = \rho_s - \sum_{j=0}^{\infty} p_{0,j}. \tag{4.101}$$

For steady state, all the probabilities $p_{i,j}$ should be positive. Then, we can deduce from (4.101) that the following relation must hold:

$$\rho_c < \rho_s, \tag{4.102}$$

which is the necessary condition for system stability. [27] shows it is also the sufficient condition.

4.5.2 Start-service probability

With respect to the general $M/MMSP/1$ model we proposed in Section 4.1, for the p2p queueing system we have $\lambda = \lambda_c$, $\mu_j = j\mu_c$, $f_{j,j+1} = \lambda_s$ and $f_{j,j-1} = j\mu_s$. the infinitesimal generator matrix $\boldsymbol{Q}$ is given by

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda_s & \lambda_s & 0 & 0 & 0 & \cdots & 0 \\ \mu_s & -\lambda_s - \mu_s & \lambda_s & 0 & 0 & \cdots & 0 \\ 0 & 2\mu_s & -\lambda_s - 2\mu_s & \lambda_s & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & (N-1)\mu_s & -\lambda_s - (N-1)\mu_s & \lambda_s \\ 0 & 0 & 0 & 0 & \cdots & N\mu_s & -N\mu_s \end{pmatrix}. \tag{4.103}$$

The probabilities related to start-service probability $\hat{\pi}_j$ of $M/MMSP/1$ are listed as following:

- $\pi_j = e^{-\rho_s}\frac{\rho_s^j}{j!}$. And the channel capacity

$$\hat{\mu} = \sum_{j=0}^{N} \pi_j \mu_j = \rho_s \mu_c. \tag{4.104}$$

- $\hat{\pi}_j$ = the probability that an HOL packet's start-service state is in channel state j, for $j = 0, 1, 2, ..., N$. And its two extreme cases

$$\hat{\pi}_j = \begin{cases} \pi_j & \text{when } \lambda_c \to 0 \\ \eta_j = \frac{\pi_j \mu_j}{\sum_{j=0}^{N} \pi_j \mu_j} = \frac{j\pi_j}{\rho_s} & \text{when } \lambda_c \to \infty \end{cases} \tag{4.105}$$

- $P\{j'|j\}$ = the probability that a packet starts the service in channel state j and finishes the service in channel state j'. We denote the corresponding matrix, called state transition matrix $\hat{\boldsymbol{Q}}$ in this paper, as

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} P\{0|0\} & P\{0|1\} & \cdots & P\{0|N\} \\ P\{1|0\} & P\{1|1\} & \cdots & P\{1|N\} \\ \vdots & \vdots & \ddots & \vdots \\ P\{N|0\} & P\{N|1\} & \cdots & P\{N|N\} \end{pmatrix}. \tag{4.106}$$

State transition factor

The diagonal matrix $\boldsymbol{D}$ of the p2p queueing system is

$$\boldsymbol{D} = \text{diag}(0, \mu_c, 2\mu_c, ..., N\mu_c). \tag{4.107}$$

From (4.13), we have the matrix $\boldsymbol{M}$, as

$\boldsymbol{M} = \boldsymbol{D} - \boldsymbol{Q} =$

$$\begin{pmatrix} \lambda_s & -\lambda_s & 0 & 0 & 0 & \cdots & 0 \\ -\mu_s & \lambda_s+\mu_s+\mu_c & \lambda_s & 0 & 0 & \cdots & 0 \\ 0 & -2\mu_s & \lambda_s+2(\mu_s+\mu_c) & -\lambda_s & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & -(N-1)\mu_s & \lambda_s+(N-1)(\mu_s+\mu_c) & -\lambda_s \\ 0 & 0 & 0 & 0 & \cdots & -N\mu_s & N(\mu_s+\mu_c) \end{pmatrix}. \tag{4.108}$$

Then the state transition matrix $\hat{\boldsymbol{Q}}$ is given by (4.18)

$$\hat{\boldsymbol{Q}} = \boldsymbol{D}(\boldsymbol{M}^{\boldsymbol{T}})^{-1}. \tag{4.109}$$

If we write down the eigenvalues of $\hat{\boldsymbol{Q}}$ as $1 = \xi_0 > |\xi_1| \geq |\xi_2| \geq |\xi_3| \geq \cdots \geq |\xi_N|$. From (4.35), we have the state transition factor

$$\beta = |\xi_1| \tag{4.110}$$

and an approximation of $\hat{\pi}_j(i,m)$, as

$$\hat{\pi}_j(i,m) \approx \beta^m(\hat{\pi}_j(i,0) - \eta_j) + \eta_j. \tag{4.111}$$

Start service probability

The Theorem 4.1 provides expressions to calculate exact value of the start service probability. However, the initial channel state probability $\pi_j^{(0)}$ which is essential to compute the start service probability $\hat{\pi}_j$ is unknown variables. It could only be calculated by resorting to the Matrix Geometric Method in the literature. Hence, we seek some simple approximations to compute $\hat{\pi}_j$. The Linear Approximation method proposed in Section 4.4.2 is the simplest among all possible approximations but only suitable for small ρ_s. If ρ tends to be large, e.g. $\rho = 500$, then the mean service time will always be a number close to $1/500\mu_c$ with little variations. That is ρ is approximately a linear function of the arrival rate λ_c. However, the start service probability is not linear with λ_c. Besides the Linear Approximation method, here we introduce the Cumulative Distribution Function (CDF) Approximation, which is applicable to different values of ρ_s, especially when ρ_s is large, such as $\rho_s > 100$. A comparison of the two approximations is provided in Section 4.5.3.

- *Linear Approximation.* [32] shows that simple expressions of $\hat{\pi}_j$ can be obtained under two extreme cases as shown in (4.105): $\hat{\pi}_j = \pi_j$ when $\lambda_c \to 0$ and $\hat{\pi}_j = \eta_j$ when $\lambda_c \to \infty$. We take an linear approximation that if a packet arrives when system is idle (with probability $1 - \rho$), it will start its service in state j with probability π_j; if a packet arrives when system is busy (with probability ρ), it will start its service in state j with probability η_j. That is

$$\hat{\pi}_j \approx (1-\rho)\pi_j + \rho\eta_j, \quad \text{for } j = 1, 2, ..., N \tag{4.112}$$

Because there is no service rate in channel state $j = 0$, a packet will start its service in state $j = 0$ only if it arrives in state 0 while system is idle. That is

$$\hat{\pi}_j = p_{0,0}, \tag{4.113}$$

which could be confirmed from (4.109) since the first element of $\boldsymbol{D}$ is 0. In order to handle the special case when $j = 0$, we have to introduce another approximation

$$p_{0,j} \approx C_j(1-\rho)\pi_j, \quad \text{for } j = 1, 2, ..., N, \tag{4.114}$$

where C_j is a normalization constant decided by

$$1 - \rho = \sum_{j=0}^{N} p_{0,j}. \tag{4.115}$$

If the arrival rate $\lambda_c \to 0$, there is no packet in the system. It follows $p_{0,j} \to \pi_j$ and $\rho \to 0$, which is (4.114); If the arrival rate $\lambda_c \to \infty$, both $p_{0,j}$ and $1-\rho$ tends to be 0, as $p_{0,j} \to 0$ and $\rho \to 1$. The left and right sides of (4.114) are balanced. From (4.94a), (4.115) and (4.114), we have

$$p_{0,0} \approx \frac{\mu_s}{\lambda_c + \lambda_s} C_j (1-\rho)\pi_1 = \frac{\lambda_s}{\lambda_c + \lambda_s} C_j (1-\rho)\pi_0. \tag{4.116}$$

For simplicity we assume that all C_j have the same value, which gives the following one possible approximation of $p_{0,j}$, as

$$\begin{cases} p_{0,0} \approx \frac{\frac{\lambda_s}{\lambda_c+\lambda_s}}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_0 \\ p_{0,j} \approx \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_j \quad \text{for } j = 1, 2, ..., N, \end{cases} \tag{4.117}$$

By combining (4.112) and (4.117), we finally establish the Linear Approximation of the start service probability

$$\begin{cases} \hat{\pi}_0 = \frac{1-\frac{\lambda_c}{\lambda_c+\lambda_s}}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_0 \\ \hat{\pi}_j = \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\rho)\pi_j + \rho\eta_j, \quad \text{for } j = 1, 2, ..., N \end{cases} \tag{4.118}$$

where the server utilization ρ is analyzed during service time analysis in Section 4.5.3.

- *Cumulative Distribution Function (CDF) Approximation.* The steady state probabilities

$$\pi_j = e^{-\rho_s} \frac{\rho_s^j}{j!}$$

are discrete variables. Define a continuous function based on the distribution of π_j, as

$$f(x) = e^{-\rho_s} \frac{\rho_s^x}{x!}. \tag{4.119}$$

It is the probability dense function of the channel state if the channel state changes continuously with x instead of j. Define the cumulative distribution function (CDF)

$$F(x) = \int_0^x f(t)dt,$$

and a variable γ

$$\gamma = \frac{F(\lambda_c)}{F(\rho_s)}. \tag{4.120}$$

All those packets arriving into the system are approximately divided into two groups: $1-\gamma$ of them start their service in channel state j with probability π_j and γ of them start service in channel state j with probability η_j. To derive the approximation of the start service probability $\hat{\pi}$ based on γ is similar to the ones conducted in the Linear Approximation, except that we replace the ρ by γ.That is

$$\hat{\pi}_j \approx (1-\gamma)\pi_j + \gamma\eta_j, \quad \text{for } j = 1, 2, ..., N \tag{4.121}$$

The channel state $j = 0$ in which there is no service rate should also be handled separately. A packet will start its service in state $j = 0$ only if it arrives in state 0 while system is idle.

$$\hat{\pi}_j = p_{0,0}, \tag{4.122}$$

which could be confirmed from (4.109) since the first element of $\boldsymbol{D}$ is 0. In order to handle the special case when $j = 0$, we have to introduce another approximation

$$p_{0,j} \approx C_j(1-\gamma)\pi_j, \quad \text{for } j = 1, 2, ..., N, \tag{4.123}$$

where C_j is a normalization constant decided by

$$1 - \gamma = \sum_{j=0}^{N} p_{0,j}. \tag{4.124}$$

If the arrival rate $\lambda_c \to 0$, there is no packet in the system. It follows $p_{0,j} \to \pi_j$ and $\gamma \to 0$, which is (4.123); If the arrival rate $\lambda_c \to \infty$, both $p_{0,j}$ and $1-\gamma$ tends to be 0, as $p_{0,j} \to 0$ and $\gamma \to 1$. The left and right sides of (4.123) are balanced. From (4.94a), (4.124) and (4.123), we have

$$p_{0,0} \approx \frac{\mu_s}{\lambda_c + \lambda_s} C_j(1-\gamma)\pi_1 = \frac{\lambda_s}{\lambda_c + \lambda_s} C_j(1-\gamma)\pi_0. \tag{4.125}$$

For simplicity we assume that all C_j have the same value, which gives the following one possible approximation of $p_{0,j}$, as

$$\begin{cases} p_{0,0} \approx \frac{\frac{\lambda_s}{\lambda_c+\lambda_s}}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\gamma)\pi_0 \\ p_{0,j} \approx \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\gamma)\pi_j \quad \text{for } j=1,2,...,N, \end{cases} \tag{4.126}$$

By combining (4.121) and (4.126), we establish the CDF approximation of the start service probability

$$\begin{cases} \hat{\pi}_0 = \frac{1-\frac{\lambda_c}{\lambda_c+\lambda_s}}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\gamma)\pi_0 \\ \hat{\pi}_j = \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}(1-\gamma)\pi_j + \gamma\eta_j, \quad \text{for } j=1,2,...,N \end{cases} \tag{4.127}$$

4.5.3 Delay analysis

The Linear Approximation of the start service probability $\hat{\pi}_j$ requires the server utilization ρ. In the following sections, we provide reasonable expressions of $E[T_j]$ and show how to obtain the ρ from conditional moment of service time $E[T_j]$. The delay on different server variations are discussed, which is verified by simulations.

Moments of service time

In Section 4.3.1, we show that the first moments of service time $E[T_j]$ could be obtained from (4.40a), which requires solving the reverse of the matrix $\boldsymbol{M}$. However, we are trying to provide expressions of interested performance metrics with physical interpretations on system parameters throughout our analysis. We state a theorem to obtain the moments of service time in series expansions.

Theorem 4.4. *The first and second moments of service time conditional on the channel state being j at the beginning of service, are given by*

$$\begin{cases} E[T_0] = \frac{1}{\lambda_s} + E[T_1] \\ E[T_j] = \frac{1}{j\mu_c} + \frac{\mu_s}{j^2\mu_c^2} + \frac{1}{j^3}\left(\frac{\mu_s-\lambda_s}{\mu_c^2} + \frac{2\mu_s^2}{\mu_c^3}\right) + \frac{1}{j^4}\left(\frac{\mu_s+\lambda_s}{\mu_c^2} + \frac{6\mu_s^2-5\lambda_s\mu_s}{\mu_c^3} + \frac{6\mu_s^3}{\mu_c^4}\right) + \cdots \end{cases} \tag{4.128a}$$

$$\begin{cases} E[T_0^2] = \frac{2}{\lambda_s}E[T_0] + E[T_1^2] \\ E[T_j^2] = \frac{2}{j^2\mu_c^2} + \frac{6\mu_s}{j^3\mu_c^3} + \frac{1}{j^4}\left(\frac{8\mu_s - 6\lambda_s}{\mu_c^3} + \frac{22\mu_s^2}{\mu_c^4}\right) + \cdots \end{cases} \tag{4.128b}$$

for $j = 1, 2, 3, ..., N$.

The prove of the Theorem 4.4 is provided by Prof. Tony LEE, as detailed in Appendix B. The simplified approximations with only the first-order component are given by:

$$\begin{cases} E[T_0] = \frac{1}{\lambda_s} + E[T_1] \\ E[T_j] = \frac{1}{j\mu_c} \end{cases} \tag{4.129a}$$

$$\begin{cases} E[T_0^2] = \frac{2}{\lambda_s}E[T_0] + E[T_1^2] \\ E[T_j^2] = \frac{2}{j^2\mu_c^2} \end{cases} \tag{4.129b}$$

for $j = 1, 2, 3, ..., N$. Numerical results show that there are little differences between the full series expansion (4.128) and the first-order approximation (4.129). The latter is enough for this p2p model, which is adapted in the following simulations.

Server utilization ρ

The server utilization ρ is derived from the Little's Law, given as

$$\rho = \lambda_c E[T]. \tag{4.130}$$

With the Linear Approximation $\hat{\pi}_j$ given in (4.118), we have the mean service time from (4.39), as

$$\begin{aligned} E[T] =& \sum_{j=0}^{N} E[T_j]\hat{\pi}_j \\ \approx& \frac{1-\rho}{1 - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{N} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right) + \rho\sum_{j=0}^{N} E[T_j]\eta_j \\ =& \frac{1-\rho}{1 - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{N} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right) + \rho\frac{1}{\hat{\mu}} \end{aligned} \tag{4.131}$$

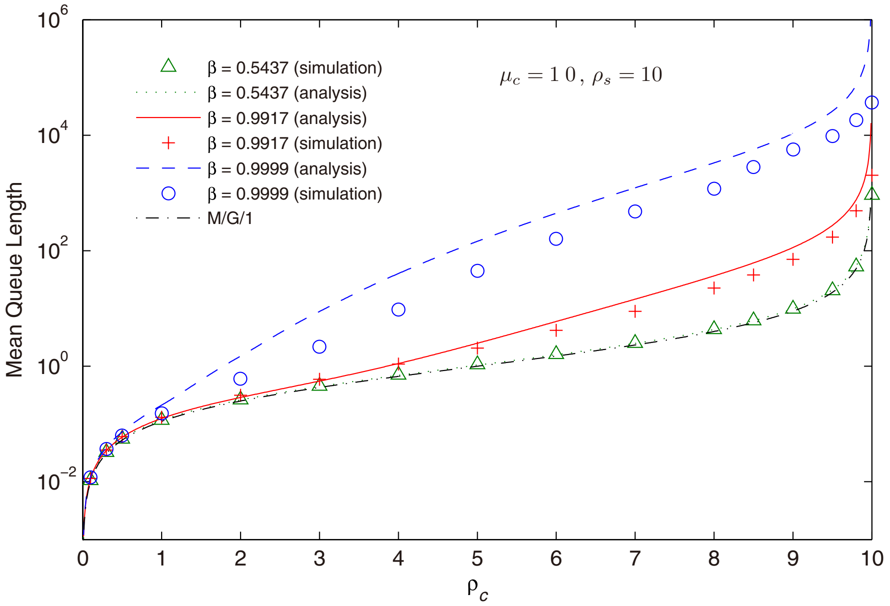

Figure 4.7: Mean queue length of p2p queueing model and $M/G/1$ for different arrival rates with different state transition factor β

where $\sum_{j=0}^{N} E[T_j]\eta_j = 1/\hat{\mu}$.

Solving the joint equations (4.130) and (4.89) for ρ with , we get an approximation of the server utilization

$$\rho \approx \frac{\frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{N} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right)}{\frac{1}{\lambda_c} + \frac{1}{1-\frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0}\left(\sum_{j=0}^{N} E[T_j]\pi_j - \frac{\lambda_c}{\lambda_c+\lambda_s}\pi_0 E[T_0]\right) - \frac{1}{\hat{\mu}}}. \tag{4.132}$$

Mean queue length

From Theorem 4.50, the mean waiting time of the p2p queue model is approximately given by

$$W \approx \frac{\frac{\lambda_c}{\hat{\mu}} E[T] + \frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j](\pi_j - \hat{\pi}_j)}{1 - \frac{\lambda_c}{\hat{\mu}}}. \tag{4.133}$$

The generalized P-K formula explicitly expresses the impact of the state transition factor on the performance of wireless channels. For a given fixed channel capacity

$\hat{\mu}$, the generalized P-K formula (4.133) reveals that the mean waiting time is greatly affected by the state transition factor β. This point is illustrated in Fig. 4.7 by the following mean queue length for different β:

$$L = \lambda_c(E[T + \lambda_c W)$$

With the assumption that the number of channel state $N \to 0$, we fix the $\rho_s = 0$, and adjust the value λ_s and μ_s proportionally to form several systems with different server dynamics [27]. Both analytical and simulation results show that this mean queue length of *M/MMSP/1* is very sensitive to the state transition factor β. When $\beta \to 1$, the mean queue length is dominated by the term $\frac{1}{1-\beta}\sum_{j=0}^{N} E[T_j](\pi_j - \hat{\pi}_j)$. When $\beta \to 0$, the mean queue length is approximately given by

$$W \approx \frac{\frac{\lambda_c}{\hat{\mu}}E[T]}{1 - \frac{\lambda_c}{\hat{\mu}}},$$

which is close to the mean waiting time of an *M/G/1* queue with mean service time $\hat{\mu}$. Our state transition factor and the generalized P-K formula can well explain that observation that a job would spend less time in systems with high server dynamics than in systems with low server dynamics conditional on fixed average service capacity [27].

The mean queue lengths (4.133) with respect to different approximations of the start-service probability $\hat{\pi}_j$ when $\rho_s = 200$ are shown in Fig. 4.8. The Linear Approximation failed to predict the mean queue length while the CDF Approximation fits the simulation results well. Because the Linear Approximation (4.112) is a function of the server utilization ρ. For large ρ_s, the mean service time is almost a constant, as shown in Fig. 4.9. It follows that ρ_s is a linear function of the arrival rate λ_c, which appears as a linear mean service time in Fig. 4.9. However, the mean waiting time is not linearly decreasing with the increasing of λ_c.

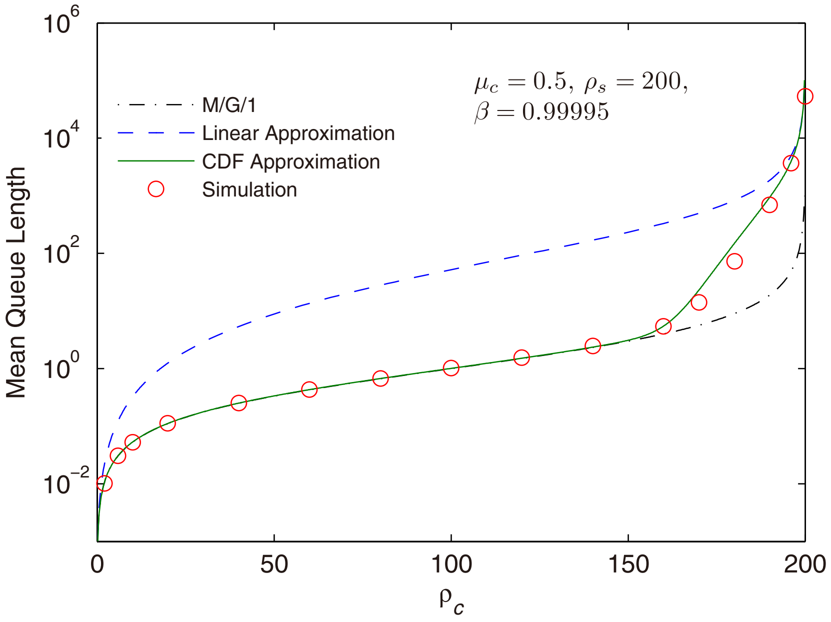

Figure 4.8: Mean queue length with the Linear Approximation and the CDF Approximation for large ρ_s.

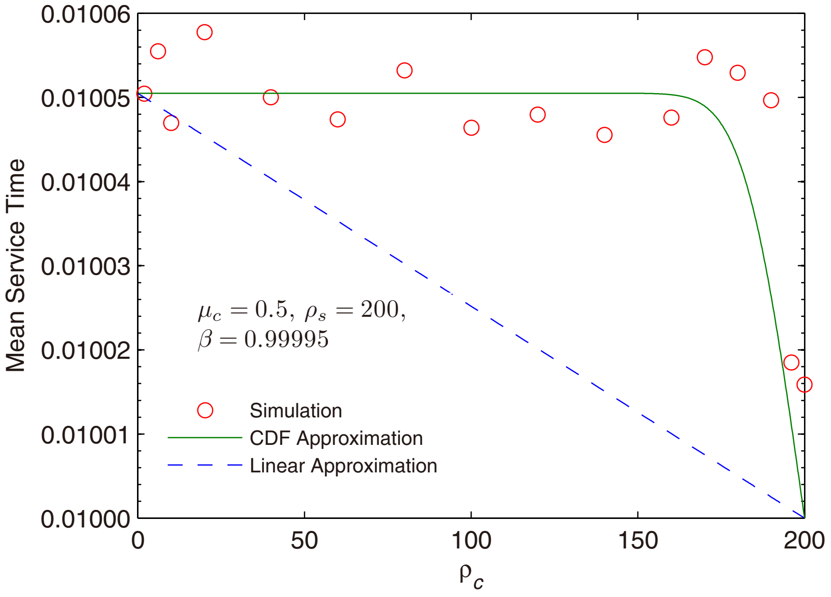

Figure 4.9: Mean service time with the Linear Approximation and the CDF Approximation for large ρ_s.

Chapter 5

Summary and Future works

In this chapter, we will first summarize the main achievements of this dissertation which based on the proposed generalized Pollaczek-Khinchin formula to analyze the queueing systems with Markov modulated service process. We hope this methodology can be extended to more general queueing systems, which can be considered as the future directions of our work.

5.1 Contribution Summary

In this dissertation, we concentrated on the performance analysis of the wireless channel by modeling the channel as a Markov modulated service process (MMSP). The main contribution is to characterize the performance of the *M/MMSP/1* queue by introducing the sate transition factor β, which indicates how fast the channel state changes in comparison with service rates. From the generalized P-K formula for the *M/MMSP/1* queue derived by using the start-service probability, we show that the queueing delay is very sensitive to this state transition factor β. When this factor is close to 1, the queueing delay of the wireless channel becomes extremely large. On the other hand, the performance of the *M/MMSP/1* queueing system can be approximated by an *M/G/1* queue when this factor β is close to 0.

We derived the exact expressions of those performance metrics for Markov channels with two states, by taking an ideal Type I Hybrid ARQ scheme as an example. We illustrate the procedure to obtain the service rate in each channel state and model the queueing system as an *M/MMSP/1/K* queue. We first derive the closed-from expressions of buffer overflow probability and queueing delay from conditional generating functions. A simple expression of state transition factor β is derived and closed-form expressions of start-service probabilities are provided. The closed-form expression of the generalized P-K formula method is derived based on finite buffer capacity.

We extend our generalized P-K formula for Markov channel with two-state to Markov modulated service process (MMSP) with finite-state. The definitions of the start-service probability and the state transition factor are well established, and an approximation of the generalized P-K formula for mean waiting time is derived. For a special three-state Markov channel with no service rate in one of the three states, we show that a simple closed-form expression of the state transition factor is available. We provide a Linear Approximation method to approximately estimate the start service probability. With the approximate start service probability, the generalized P-K formula provides a good approximation to the mean waiting time, as shown by simulations. For Markov channels with large number of states, we illustrate our work through a general model in peer-to-peer systems. Besides the Linear Approximation method for channels with small number of states, we propose a CDF Approximation method to approximately calculate the start service probability for channels with large number of states. Our generalized P-K formula is a good approximation of the mean queue length of the P2P model, as verified by simulations.

5.2 Future Work

In this dissertation, we have established the generalized Pollaczek-Khinchin formula method to analyze the queueing systems with Markov modulated service process. Throughout those listed examples we most concerned, the Markovian service state could only change to its neighbor states. However, our derivations of the state transition factor, start-service probability, conditional moments of service time and the generalized P-K formula are still valid if the channel is allowed to jump to any random state. The closed-from expressions of delay and the state transition factor are obtained under the assumptions of Poisson arrivals and exponentially distributed service time. Hence, extending this approach to general queueing models, such as with renewal arrival process and generally distributed service time, will be our future directions.

We have proposed two methods, the Linear Approximation and the CDF Approximation, to approximate the start-service probability for Markov channels with small number and large number of states separately. A general approximation for start-service probability is expected to fit all Markov channels.

The analysis curves shown in Fig. 4.7 are a little higher than those simulation curves. Because the state transition factor is defined as the second largest eigenvalue of the state transition matrix $\hat{\boldsymbol{Q}}$ and those contributions of other eigenvalues on the generalized P-K formula are ignored. Just like the large-deviation approximation [9,19,32] for buffer overflow probability as we shown in Fig. 3.4 with $\alpha \approx P\{X(\infty) > 0\}$. We expect an adjustment constant on the approximate generalized P-K formula, similar to the asymptotic constant in large-deviation approximation, to provide a more accurate mean waiting time.

We do believe that the generalized Pollaczek-Khinchin formula for Markov channels is a compliment of the finite-state Markov channel model proposed by Wang and Moayeri [50] which will benefit the queueing analysis of system models based on [50]; and the state transition factor is essential to characterize the performance of communication systems over Markov channels which will bring convenience to the design of different communication systems when the delay is considered.

Appendix A

Two-state Markov Channel with No Service Rate in *Bad* State

Here we study the two-state Markov channel with one of its service rate is 0. We derive the interested performance metrics with different methods and show that it is consistat with our general two-state model analyzed in Chapter 3.

A.1 Model Discription

For a fixed service rate (when the channel is in *Good* state), the corresponding service rate distribute exponentially with parameter μ_0. There is no service rate when the channel is in *Bad* state. That is $\mu_1 = 0$. Figure A.1 shows the state rate transition diagram of $M/MMSP/1/K$ with $\mu_1 = 0$, from which we derive the following set of Kolmogorov forward equations

$$(\lambda + f_0)p_{0,0} = f_1 p_{0,1} + \mu_0 p_{1,0} \tag{A.1a}$$

$$(\lambda + f_1)p_{0,1} = f_0 p_{0,0} \tag{A.1b}$$

$$(\lambda + f_0 + \mu_0)p_{i,0} = f_1 p_{i,1} + \lambda p_{i-1,0} + \mu_0 p_{i+1,0} \tag{A.1c}$$

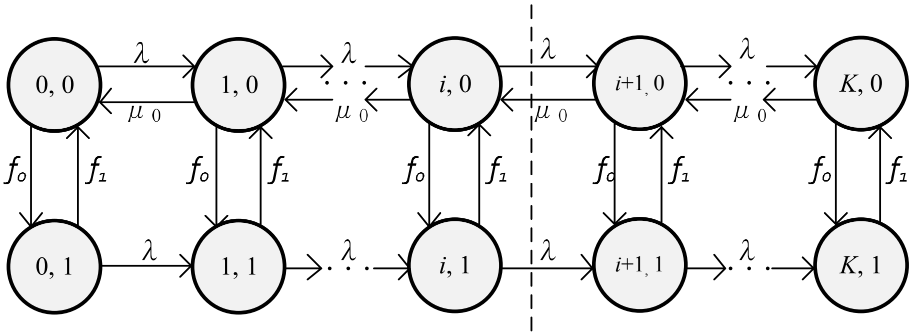

Figure A.1: Rate transition diagram with finite waiting rooms K and $\mu_1 = 0$

$$(\lambda + f_1)p_{i,1} = f_0 p_{i,0} + \lambda p_{i-1,1} \tag{A.1d}$$

$$(f_0 + \mu_0)p_{K,0} = f_1 p_{K,1} + \lambda p_{K-1,0} \tag{A.1e}$$

$$f_1 p_{K,1} = f_0 p_{K,0} + \lambda p_{K-1,1}. \tag{A.1f}$$

for all $i = 1, 2, ..., K - 1$. From the rate transition diagram shown in figure A.1, the balance equation with respect to the dashed line cut is given by

$$\lambda(p_{i,0} + p_{i,1}) = \mu_0 p_{i+1,0} \tag{A.2}$$

for all $i = 0, 1, 2, ..., K - 1$. Summing (A.2) over i, we obtain

$$\lambda(1 - P_K) = \mu_0(\pi_0 - p_{0,0}). \tag{A.3}$$

Corresponding to (3.7), the system capacity when $\mu_1 = 0$ is given by

$$\hat{\mu} = \pi_0 \mu_0, \tag{A.4}$$

which is the average serve rate when the system is always busy. Then (A.3) can be rewritten as $\hat{\mu} - \lambda' = \mu_0 p_{0,0}$.

A.2 Generating Functions

The system generating function

$$G(z) = \sum_{i=0}^{i=K} z^i P_i, \quad |z| \leq 1, j = 0, 1.$$

is obtained by applying $\mu_1 = 0$ to (3.11) as follows

$$\begin{aligned} G(z) = \frac{1}{g(z)} \Big(&(\lambda(1-z) + f_0 + f_1)(\hat{\mu} - \lambda(1 - P_K)) \\ &+ \lambda z^K \left(\mu_0 p_{K,1}(1-z) - (\lambda(1-z) + f_0 + f_1) P_K z\right) \Big), \end{aligned} \tag{A.5}$$

where

$$g(z) = \lambda^2 z^2 - \lambda(\lambda + f_0 + f_1 + \mu_0) z + \mu_0(\lambda + f_1).$$

The function $g(z)$ in the denominator processes two roots z_1 and z_2 ($z_1 < z_2$ by default), as

$$\begin{aligned} z_1 &= \frac{\lambda + f_0 + f_1 + \mu_0 + \sqrt{(\lambda + f_0 + f_1 + \mu_0)^2 - 4\mu_0(\lambda + f_1)}}{2\lambda} \\ z_2 &= \frac{\lambda + f_0 + f_1 + \mu_0 - \sqrt{(\lambda + f_0 + f_1 + \mu_0)^2 - 4\mu_0(\lambda + f_1)}}{2\lambda} \end{aligned}$$

After extensive mathematical development, it turns out that both z_1 and z_2 are larger than 1 under the system stability condition $\lambda < \hat{\mu}$. Otherwise, rearrange the function $g(z)$ as

$$g(z) = \lambda(\lambda z - \mu_0)(z - 1) - (\lambda z - \hat{\mu})(f_0 + f_1). \tag{A.6}$$

It follows that the both of the roots of $g(z)$ locate at the two points of intersections of the following two curves, as plotted in Fig. A.2:

$$y_1(z) = \lambda(\lambda z - \mu_0)(z - 1)$$

and

$$y_2(z) = (\lambda z - \hat{\mu})(f_0 + f_1).$$

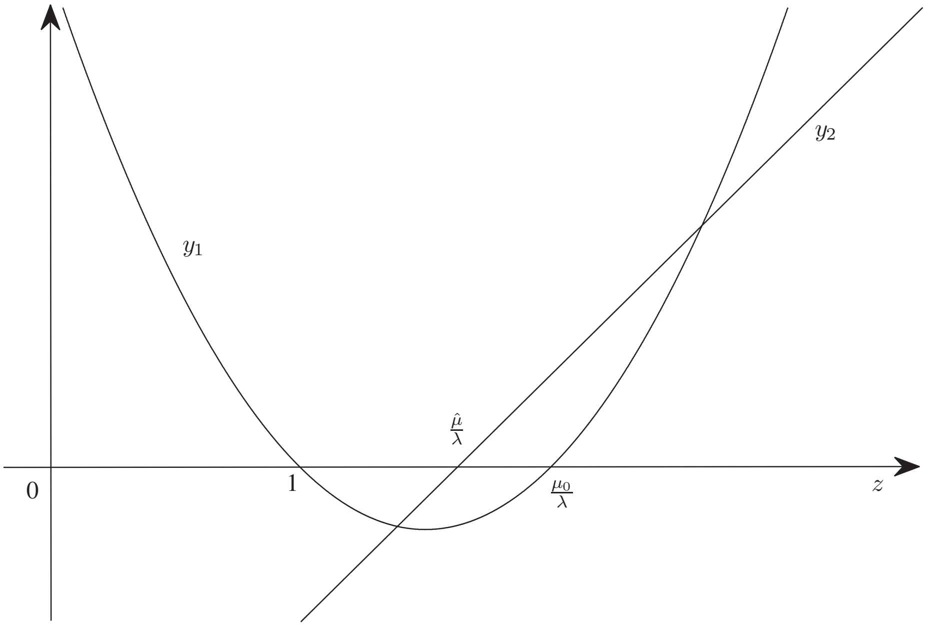

Figure A.2: z_1 and z_2 locate at the three points of intersections of y_1 and y_2.

For positive λ, μ_0, f_0 and f_1, the following relationship always holds under the system stability condition $\lambda < \hat{\mu}$:

$$1 < z_1 < \frac{\hat{\mu}}{\lambda} < \frac{\mu_0}{\lambda} < z_2. \tag{A.7}$$

A.3 Buffer overflow probabilities

In the following we show an alternative method to obtain the blocking probability of the finite queueing systems based on the probabilities of queueing systems with infinite waiting rooms.

A.3.1 Infinite waiting rooms

Taking the limit $K \to \infty$ in (A.5), we obtain the generating function for system with infinite waiting rooms as follows:

$$G(z) = \frac{-\lambda(\hat{\mu}-\lambda)z + (\lambda+f_0+f_1)(\hat{\mu}-\lambda)}{\lambda^2 z^2 - \lambda(\lambda+f_0+f_1+\mu_0)z + \mu_0(\lambda+f_1)}. \tag{A.8}$$

The probability that there are i packets in the system

$$P_i = Az_1^{-i} + Bz_2^{-i}, \quad i = 0, 1, 2, ... \tag{A.9}$$

can be obtained from $G(z)$ expressed as follows:

$$G(z) = \frac{A}{1-z_1^{-1}z} + \frac{B}{1-z_2^{-1}z},$$

where A and B can be easily derived from (A.8), as following

$$\begin{aligned} A &= \frac{\hat{\mu}-\lambda}{\lambda}\frac{z_1^{-1}}{z_1-z_2}\left(\frac{\mu_0}{\lambda}-z_2\right) \\ B &= \frac{\hat{\mu}-\lambda}{\lambda}\frac{z_2^{-1}}{z_2-z_1}\left(\frac{\mu_0}{\lambda}-z_1\right). \end{aligned}$$

More specifically, we could express both parameters A and B in forms of channel model parameters

$$\begin{aligned} A &= \frac{\hat{\mu}-\lambda}{2\mu_0(\lambda+f_1)}\left(\lambda+f_0+f_1+\frac{\mu_0(\lambda+f_1-f_0)-(\lambda+f_0+f_1)^2}{\sqrt{(\lambda+f_0+f_1+\mu_0)^2-4\mu_0(\lambda+f_1)}}\right) \\ B &= \frac{\hat{\mu}-\lambda}{2\mu_0(\lambda+f_1)}\left(\lambda+f_0+f_1+\frac{(\lambda+f_0+f_1)^2-\mu_0(\lambda+f_1-f_0)}{\sqrt{(\lambda+f_0+f_1+\mu_0)^2-4\mu_0(\lambda+f_1)}}\right). \end{aligned}$$

Taking the inverse z-transform of $G(z)$, we get (A.9). It follows that the following probabilities $p_{i,0}$ and $p_{i,1}$ can be calculated from (A.9) along with (3.3) and (A.2):

$$p_{i,0} = \frac{\lambda}{\mu_0}Az_1^{1-i} + \frac{\lambda}{\mu_0}Bz_2^{1-i} \tag{A.10a}$$

$$p_{i,1} = (1-\frac{\lambda}{\mu_0}z_1)Az_1^{-i} + (1-\frac{\lambda}{\mu_0}z_2)Bz_2^{-i} \tag{A.10b}$$

The server utilization is given as follows:

$$\rho = 1 - P_0 = \frac{\lambda}{\mu_0} \frac{\lambda + f_0 + f_1 + \mu_0 - \hat{\mu}}{\lambda + f_1}. \tag{A.11}$$

For system to be stable, the server utilization must satisfy $\rho < 1$. From (A.11), the stable condition implies $\lambda < \hat{\mu}$. It is the system stability condition for the infinite queue and it is why $\hat{\mu}$ is defined as the system capacity.

A.3.2 Finite buffer capacity K

An arriving packet will be blocked if there are K packets in the system. The following theorem gives the blocking probability.

Theorem A.1. *For finite queueing system with K waiting rooms, the blocking probability is given by*

$$P_K = \frac{\hat{\mu} - \lambda}{\hat{\mu}/S - \lambda}, \tag{A.12}$$

where

$$S = \frac{A z_1^{-K}}{1 - z_1^{-1}} + \frac{B z_2^{-K}}{1 - z_2^{-1}}.$$

Proof. Comparing the balance equations of both finite and infinite queueing systems, the first $K - 1$ equations are the same. Hence, we assume the ratios between the first $K - 1$ state probabilities are constant. Then the first $K - 1$ state probabilities can be expressed as

$$p_{i,0} = \frac{1}{S_0} \left(\frac{\lambda}{\mu_0} A z_1^{1-i} + \frac{\lambda}{\mu_0} B z_2^{1-i} \right) \tag{A.13}$$

$$p_{i,1} = \frac{1}{S_0} \left((1 - \frac{\lambda}{\mu_0} z_1) A z_1^{-i} + (1 - \frac{\lambda}{\mu_0} z_2) B z_2^{-i} \right) \tag{A.14}$$

$$i = 0, 1, 2, ..., K - 1,$$

where S_0 is a normalization constant. For the state $(K, 0)$, probabilities should satisfy (A.2) when $i = K - 1$,

$$\lambda(p_{K-1,0} + p_{K-1,1}) = \mu_0 p_{K,0},$$

which yields

$$p_{K,0} = \frac{1}{S_0}\left(\frac{\lambda}{\mu_0}Az_1^{1-K} + \frac{\lambda}{\mu_0}Bz_2^{1-K}\right). \tag{A.15}$$

For the state $(K, 1)$, the balance equation (A.1f) gives

$$\begin{aligned}
p_{K,1} &= \frac{1}{f_1}\left(f_0 p_{K,0} + \lambda p_{K-1,1}\right) \\
&= \frac{1}{S_0 f_1}\left(f_0\left(\frac{\lambda}{\mu_0}Az_1^{1-K} + \frac{\lambda}{\mu_0}Bz_2^{1-K}\right) + \lambda\left((1-\frac{\lambda}{\mu_0}z_1)Az_1^{1-K} + (1-\frac{\lambda}{\mu_0}z_2)Bz_2^{1-K}\right)\right) \\
&= \frac{\lambda}{S_0 f_1 \mu_0}\left((f_0+\mu_0-\lambda z_1)Az_1^{1-K} + (f_0+\mu_0-\lambda z_2)Bz_2^{1-K}\right) \\
&= \frac{1}{S_0}(1+\frac{\lambda}{f_1})\left((1-\frac{\lambda}{\mu_0}z_1)Az_1^{-K} + (1-\frac{\lambda}{\mu_0}z_2)Bz_2^{-K}\right).
\end{aligned} \tag{A.16}$$

The normalization constant S_0 is determined from the property that the summation of all the probabilities goes to unity:

$$\sum_{i=0}^{K}(p_{i,0} + p_{i,1}) = 1. \tag{A.17}$$

We have all the probabilities in state $j = 0$ from (A.13) and (A.15), as

$$\begin{aligned}
\sum_{i=0}^{K} p_{i,0} =& \frac{1}{S_0}\frac{\lambda}{\mu_0}\left(A\frac{z_1\left(1-z_1^{-(K+1)}\right)}{1-z_1^{-1}} + B\frac{z_2\left(1-z_2^{-(1+K)}\right)}{1-z_2^{-1}}\right) \\
=& \frac{1}{S_0}\left(\frac{f_1}{f_0+f_1} - \frac{\lambda}{\mu_0}\left(\frac{Az_1^{-K}}{1-z_1^{-1}} + \frac{Bz_2^{-K}}{1-z_2^{-1}}\right)\right)
\end{aligned} \tag{A.18}$$

and all the probabilities in state $j = 1$ from (A.14) and (A.16)

$$\begin{aligned}
\sum_{i=0}^{K} p_{i,1} =& \sum_{i=0}^{K-1} p_{i,1} + p_{K,1} \\
=& \frac{1}{S_0}\left((1-\frac{\lambda}{\mu_0}z_1)A\frac{1-z_1^{-K}}{1-z_1^{-1}} + (1-\frac{\lambda}{\mu_0}z_2)B\frac{1-z_2^{-K}}{1-z_2^{-1}}\right) \\
&+ \frac{\lambda}{f_1\mu_0}\left((f_0+\mu_0-\lambda z_1)Az_1^{1-K} + (f_0+\mu_0-\lambda z_2)Bz_2^{1-K}\right) \\
=& \frac{1}{S_0}\left(\frac{f_0}{f_0+f_1} - \frac{\lambda f_0}{\mu_0 f_1}\left(\frac{Az_1^{-K}}{1-z_1^{-1}} + \frac{Bz_2^{-K}}{1-z_2^{-1}}\right)\right) \\
=& \frac{f_0}{f_1}\sum_{i=0}^{K} p_{i,0}
\end{aligned} \tag{A.19}$$

Substituting (A.18) and (A.19) into (A.17), we have

$$\begin{aligned}\sum_{i=0}^{K}(p_{i,0}+p_{i,1}) =&\frac{1}{S_0}\left(1-\frac{f_0+f_1}{f_1}\frac{\lambda}{\mu_0}\left(\frac{Az_1^{-K}}{1-z_1^{-1}}+\frac{Bz_2^{-K}}{1-z_2^{-1}}\right)\right)\\ =&1,\end{aligned}$$

which results in

$$\begin{aligned}S_0 =&1-\frac{f_0+f_1}{f_1}\frac{\lambda}{\mu_0}\left(\frac{Az_1^{-K}}{1-z_1^{-1}}+\frac{Bz_2^{-K}}{1-z_2^{-1}}\right)\\ =&1-\frac{\lambda}{\hat{\mu}}\left(\frac{Az_1^{-K}}{1-z_1^{-1}}+\frac{Bz_2^{-K}}{1-z_2^{-1}}\right).\end{aligned}$$

The blocking probability that packets arriving when there are K packets in the system is given by

$$\begin{aligned}P_B =&p_{K,0}+p_{K,1}\\ =&\frac{1}{S_0}\left(\left(1+\frac{\lambda}{f_1}(1-\frac{\lambda}{\mu_0}z_1)\right)Az_1^{-K}+\left(1+\frac{\lambda}{f_1}(1-\frac{\lambda}{\mu_0}z_2)\right)Bz_2^{-K}\right) \qquad \text{(A.20)}\end{aligned}$$

To simplify the expression of P_B, we need the following equations:

$$1+\frac{\lambda}{f_1}(1-\frac{\lambda}{\mu_0}z_1)=\frac{\hat{\mu}-\lambda}{\hat{\mu}}\frac{1}{1-z_1^{-1}}. \qquad \text{(A.21a)}$$

$$1+\frac{\lambda}{f_1}(1-\frac{\lambda}{\mu_0}z_2)=\frac{\hat{\mu}-\lambda}{\hat{\mu}}\frac{1}{1-z_2^{-1}}. \qquad \text{(A.21b)}$$

Substituting (A.4) into (A.21a), we have

$$\frac{1}{\mu_0 f_1}(\mu_0 f_1+\lambda\mu_0-\lambda^2 z_1)=\frac{\mu_0 f_1-\lambda(f_0+f_1)}{\mu_0 f_1}\frac{1}{1-z_1^{-1}}, \qquad \text{(A.22)}$$

Reorganizing (A.22), it gives

$$\lambda^2 z_1-\lambda(\lambda+\mu_0+f_0+f_1)+\mu_0(\lambda+f_1)z_1^{-1}=0. \qquad \text{(A.23)}$$

Please note that (A.23) always holds since z_1 is one of the two roots of $g(z)$. Similarly, (A.21b) is derived since z_2 is the other root of $g(z)$. Substituting both (A.21a) and (A.21b) into (A.20), we have (A.12). □

A.3.3 Large-deviation approximation

The blocking probability P_K is a combination of two exponential series z_1^{-K} and z_2^{-K}. Since $0 < z_2^{-1} < \lambda/\hat{\mu} < z_1^{-1}$ as shown in Fig. A.2, S_0 is dominated by z_1^{-K} for large K(approximately > 10), or specifically, we have $S_0 \approx A\frac{z_1^{-K}}{1-z_1^{-1}}$. It follows that K can be approximately given by

$$K \approx \frac{ln\left(\frac{\lambda}{\hat{\mu}} + \frac{\hat{\mu}-\lambda}{\hat{\mu}}\frac{1}{P_K}\right) + ln\frac{A}{1-z_1^{-1}}}{lnz_1}. \tag{A.24}$$

In practice, the lagging bound can be approximately determined by expression (A.24) for a given blocking probability.

A.4 Delay analysis

Comparing infinite and finite queueing systems, the only difference is the net arrival rate changing from λ to λ'. Hence, the server utilization ρ for finite queueing system can be obtained by replacing λ in (A.11) by λ'. Let $E(T)$ be the mean service time of the system. From Little's Law, $\rho = \lambda' E(T)$, we have

$$E(T) = \frac{\lambda' + f_0 + f_1 + \mu_0 - \hat{\mu}}{\mu_0(\lambda' + f_1)}.$$

The mean number of packets in the system (the mean queue length) $L = \sum_{i=0}^{i=K} iP_i$ can be derived from (A.5) and given as follows:

$$L = \frac{\lambda(1-(K+1)P_K)}{\hat{\mu}-\lambda} + \frac{(1-p_{K,1})\lambda\mu_0 - \hat{\mu}}{(f_0+f_1)(\hat{\mu}-\lambda)},$$

where $p_{K,1}$ is given in (A.16). From Little's Law, we obtain the following expression of mean waiting time:

$$W = \frac{L}{\lambda'} - \frac{\lambda' + f_0 + f_1 + \mu_0 - \hat{\mu}}{\mu_0(\lambda' + f_1)}.$$

The service time is the time that starts from the moment packet is at the head of the queue till the service is finished. For those packets arrived when the server is idle and the channel state is *Bad*, we include the time between the arrival and the channel state changes to *Good* in their service time. However, this interval is considered as waiting time in [59].

A.5 Consistency with $M/MMSP/1/K$

The two-state system with no service rate in *Bad* state is a special case of the $M/MMSP/1/K$ system previously analyzed in 3.2, by replacing $\mu_1 = 0$. Similarly, the steady state probabilities can be obtained by taking the inverse z-transform of the generating function $G(z)$. The function in the denominator of the generating function, $g_(z)$ in (A.6), is consistent with the one in (3.10), if we consider the third root of (3.10) to be zero, as $z_0 = 0$. If the service rate at *bad* state reduces to zero, we will have $z_0 = 0$. Then the probability that there are i packets in the system is given by

$$P_i = \begin{cases} Az_1^{-i} + Bz_2^{-i} & \text{for } i = 0, 1, ..., K-1 \\ Az_1^{-i} + Bz_2^{-i} + C & \text{for } i = K \end{cases} \tag{A.25}$$

The expression of the buffer overflow probability (when $i = K$) is different from other ones (when $0 \leq i \leq K-1$). There exists an additional component C in P_K and this agrees with [20]'s work.

A.6 Numerical and simulation results

For all simulations presented in this paper, packets arrive according to Poisson process with parameter λ and we generate a large enough number of arriving packets to make sure that the system reaches steady state. The wireless channel changes between two states *Good* and *Bad* alternatively with parameters $f_0 = 0.002$ and $f_1 = 0.02$ respectively.

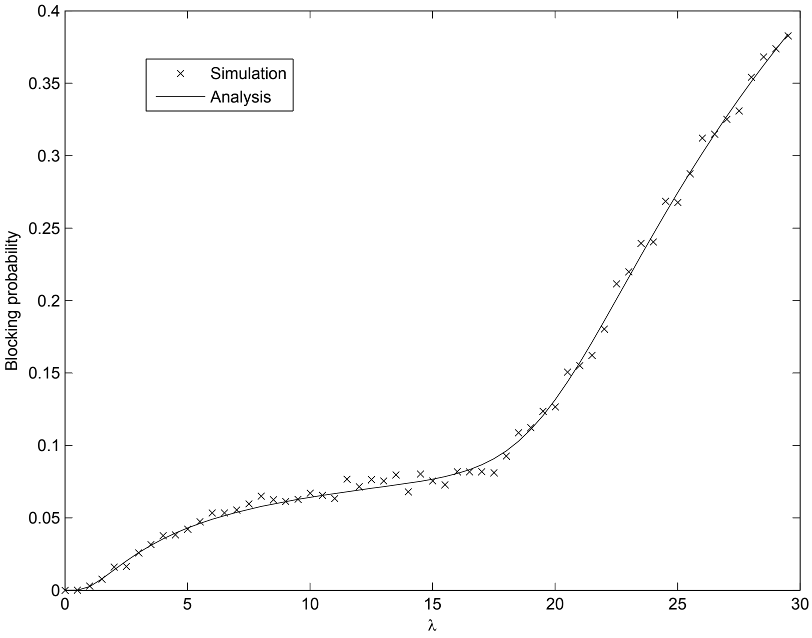

Figure A.3: Blocking probability with lagging bound $K = 20$

We first consider the case of channels with constant lagging bound K. If there are K packets in the system including the one in service upon the arrival of a packet, then the packet will be blocked. The mean system service rate at *Good* state is 2. The total time that the channel is in *Good* state (with probability $\pi_0 = 10/11$) is 10 times longer than that of the *Bad* state (with probability $\pi_1 = 1/11$), from which the system capacity should be $\hat{\mu} = 1.82$.

Figure A.3 shows the corresponding blocking probability curve versus λ. Our analysis is confirmed by the simulation results. In general, the blocking probability increase with respect to λ and in the practical range $0 < \lambda < \hat{\mu}$ of interest, P_K increases slowly.

Figure A.4 shows the blocking probabilities at different lagging bounds and arrive rates. All other parameters remain the same as before. When $\lambda < \hat{\mu}$, the blocking probability will decrease to 0 if given large enough lagging bound K. On the other hand, the blocking probability always exists for any K when $\lambda > \hat{\mu}$. At this time,

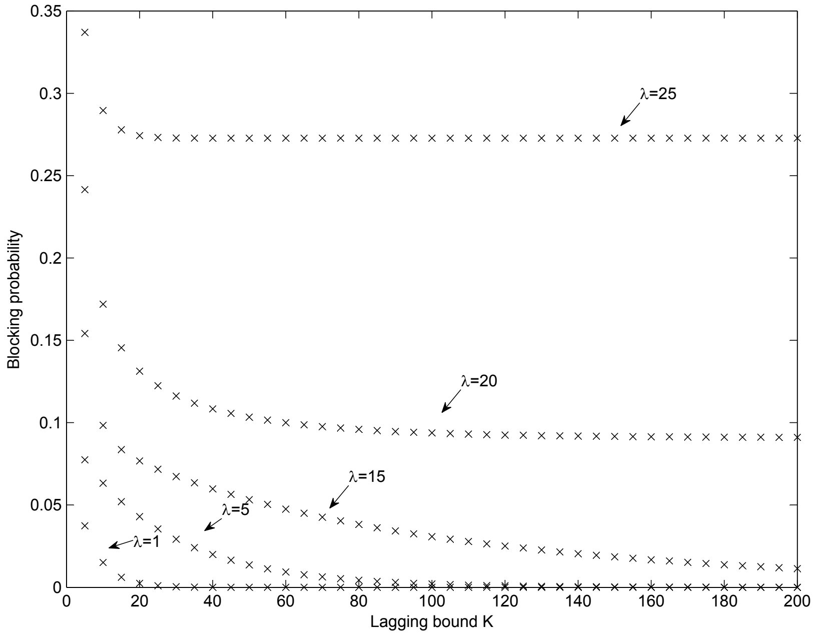

Figure A.4: Blocking probabilities at different K and λ.

the blocking probability will decrease to $1 - \hat{\mu}/\lambda$ when K goes to infinity.

Appendix B

Series expansions of service time Moments in P2P model

We derive the expressions of the conditional first and second moments of service time, $E[T_j]$ and $E[T_j^2]$, given that the number of server at the beginning of the service is j, by serial expansion.

B.1 First moment of service time

We have the conditional first moment of service time $E[T_j]$ from (4.42) as

$$(\lambda_s + j\mu_s + j\mu_c)E[T_j] = 1 + \lambda_s E[T_{j+1}] + j\mu_s E[T_{j-1}]. \tag{B.1}$$

Applying the following Taylor series approximations to (B.1)

$$E[T_{j+1}] \approx E[T_j] + \frac{dE[T_j]}{dj} \quad \text{and} \quad E[T_{j-1}] \approx E[T_j] - \frac{dE[T_j]}{dj}, \tag{B.2}$$

we have

$$j\mu_c E[T_j] \cong 1 + (\lambda_s - j\mu_s)\frac{dE[T_j]}{dj}. \tag{B.3}$$

When $j = 1$, the combination of (B.2 and (B.3) yields:

$$\mu_c E[T_1] \cong 1 + (\lambda_s - \mu_s)(E[T_2] - E[T_1]). \tag{B.4}$$

Also (B.1) becomes

$$(\lambda_s + j\mu_s + j\mu_c)E[T_1] = 1 + \lambda_s E[T_2] + j\mu_s E[T_0]. \tag{B.5}$$

If there is no server available at the beginning of the service, $j = 0$, then the mean service time is given by:

$$E[T_0] = \frac{1}{\lambda_s} + E[T_1]. \tag{B.6}$$

Solving (B.4), (B.5) and (B.6) simultaneously, we have

$$E[T_0] = \frac{\mu_s + \mu_c}{\lambda_s \mu_c} \tag{B.7a}$$

$$E[T_1] = \frac{\mu_s}{\lambda_s \mu_c} \tag{B.7b}$$

$$E[T_2] = \frac{\mu_s - \mu_c}{\lambda_s \mu_s} \tag{B.7c}$$

From (B.1) and (B.7), we can recursively calculate $E[T_j], j = 3, 4, ..., N$. For large j, the conditional mean service time can also be obtained by the series expansion:

$$E[T_j] = \sum_{k=1}^{N} \frac{\tau_k}{j^k}.$$

The coefficients τ_k in the above series are the solution of a set of linear equations. We need the following identities to establish those equations:

$$\begin{aligned} \frac{1}{j+1} &= \frac{1}{j}\left(\frac{1}{1+\frac{1}{j}}\right) = \frac{1}{j}\left(1 - \frac{1}{j} + \frac{1}{j^2} - \frac{1}{j^3} + \cdots\right) \\ &= \frac{1}{j} - \frac{1}{j^2} + \frac{1}{j^3} - \frac{1}{j^4} + \cdots \end{aligned} \tag{B.8a}$$

$$\begin{aligned} \frac{1}{j-1} &= \frac{1}{j}\left(\frac{1}{1-\frac{1}{j}}\right) = \frac{1}{j}\left(1 + \frac{1}{j} + \frac{1}{j^2} + \frac{1}{j^3} + \cdots\right) \\ &= \frac{1}{j} + \frac{1}{j^2} + \frac{1}{j^3} + \frac{1}{j^4} + \cdots \end{aligned} \tag{B.8b}$$

$$\frac{1}{(j+1)^2} = -\frac{d}{dn}\frac{1}{j+1} = \frac{1}{j^2} - \frac{2}{j^3} + \frac{3}{j^4} - \frac{4}{j^5} + \cdots \tag{B.8c}$$

$$\frac{1}{(j-1)^2} = -\frac{d}{dn}\frac{1}{j-1} = \frac{1}{j^2} + \frac{2}{j^3} + \frac{3}{j^4} + \frac{4}{j^5} + \cdots \tag{B.8d}$$

$$\frac{1}{(j+1)^3} = -\frac{1}{2}\frac{d}{dn}\frac{1}{(j+1)^2} = \frac{1}{j^3} - \frac{3}{j^4} + \frac{6}{j^5} - \frac{10}{j^6} + \frac{15}{j^7} - \cdots \tag{B.8e}$$

$$\frac{1}{(j-1)^3} = -\frac{1}{2}\frac{d}{dn}\frac{1}{(j-1)^2} = \frac{1}{j^3} + \frac{3}{j^4} + \frac{6}{j^5} + \frac{10}{j^6} + \frac{15}{j^7} + \cdots \tag{B.8f}$$

$$\frac{1}{(j+1)^4} = -\frac{1}{3}\frac{d}{dn}\frac{1}{(j+1)^3} = \frac{1}{j^4} - \frac{4}{j^5} + \frac{10}{j^6} - \frac{20}{j^7} + \frac{35}{j^8} - \cdots \tag{B.8g}$$

$$\frac{1}{(j-1)^4} = -\frac{1}{3}\frac{d}{dn}\frac{1}{(j-1)^3} = \frac{1}{j^4} + \frac{4}{j^5} + \frac{10}{j^6} + \frac{20}{j^7} + \frac{35}{j^8} + \cdots \tag{B.8h}$$

Substituting (B.8) into (B.1), we have

$$\begin{aligned}
&(\lambda_s + j\mu_s + j\mu_c)\left[\frac{\tau_1}{j} + \frac{\tau_2}{j^2} + \frac{\tau_3}{j^3} + \frac{\tau_4}{j^4} + \cdots\right] \\
&=1 + \lambda_s \left[\begin{array}{c} \tau_1\left(\frac{1}{j} - \frac{1}{j^2} + \frac{1}{j^3} - \frac{1}{j^4} + \frac{1}{j^5} - \cdots\right) \\ +\tau_2\left(\frac{1}{j^2} - \frac{2}{j^3} + \frac{3}{j^4} - \frac{4}{j^5} + \frac{5}{j^6} - \cdots\right) \\ +\tau_3\left(\frac{1}{j^3} - \frac{3}{j^4} + \frac{6}{j^5} - \frac{10}{j^6} + \frac{15}{j^7} - \cdots\right) \\ + \cdots \end{array}\right] \\
&\quad + j\mu_s \left[\begin{array}{c} \tau_1\left(\frac{1}{j} + \frac{1}{j^2} + \frac{1}{j^3} + \frac{1}{j^4} + \frac{1}{j^5} + \cdots\right) \\ +\tau_2\left(\frac{1}{j^2} + \frac{2}{j^3} + \frac{3}{j^4} + \frac{4}{j^5} + \frac{5}{j^6} + \cdots\right) \\ +\tau_3\left(\frac{1}{j^3} + \frac{3}{j^4} + \frac{6}{j^5} + \frac{10}{j^6} + \frac{15}{j^7} + \cdots\right) \\ + \cdots \end{array}\right]
\end{aligned} \tag{B.9}$$

which holds for all $j \geq 3$. Comparing coefficients of (B.9), we obtain the following set of equations:

$$\tau_1(\mu_s + \mu_c) = 1 + \tau_1\mu_s \tag{B.10a}$$

$$\tau_1\lambda_s + \tau_2(\mu_s + \mu_c) = \tau_1\lambda_s + \tau_1\mu_s + \tau_2\mu_s \tag{B.10b}$$

$$\tau_2\lambda_s + \tau_3(\mu_s + \mu_c) = -\tau_1\lambda_s + \tau_2\lambda_s + \tau_1\mu_s + 2\tau_2\mu_s + \tau_3\mu_s \tag{B.10c}$$

$$\tau_3\lambda_s + \tau_4(\mu_s + \mu_c) = \tau_1\lambda_s - 2\tau_2\lambda_s + \tau_3\lambda_s + \tau_1\mu_s + 3\tau_2\mu_s + 3\tau_3\mu_s + \tau_4\mu_s \tag{B.10d}$$

$$\vdots$$

Solving this set of linear equations (B.10), we obtain:

$$\begin{aligned}
\tau_1 =& \frac{1}{\mu_c} \\
\tau_2 =& \frac{\mu_s}{\mu_c^2} \\
\tau_3 =& \frac{\mu_s - \lambda_s}{\mu_c^2} + \frac{2\mu_s^2}{\mu_c^3} \\
\tau_4 =& \frac{\mu_s + \lambda_s}{\mu_c^2} + \frac{6\mu_s^2 - 5\lambda_s\mu_s}{\mu_c^3} + \frac{6\mu_s^3}{\mu_c^4} \\
&\vdots
\end{aligned}$$

It follows that

$$E[T_j] = \frac{1}{j\mu_c} + \frac{\mu_s}{j^2\mu_c^2} + \frac{1}{j^3}\left(\frac{\mu_s - \lambda_s}{\mu_c^2} + \frac{2\mu_s^2}{\mu_c^3}\right) + \frac{1}{j^4}\left(\frac{\mu_s + \lambda_s}{\mu_c^2} + \frac{6\mu_s^2 - 5\lambda_s\mu_s}{\mu_c^3} + \frac{6\mu_s^3}{\mu_c^4}\right) + \cdots$$

for $j = 3, 4, 5, ..., N$

B.2 Second moment of service time

From (4.43), the conditional second moment $E[T_j^2]$ satisfies the following equation:

$$(\lambda_s + j\mu_s + j\mu_c)E[T_j^2] = 2E[T_j] + \lambda_s E[T_{j+1}^2] + j\mu_s E[T_{j-1}^2] \tag{B.12}$$

Applying the following Taylor series approximations to (B.12)

$$E[T_{j+1}^2] \approx E[T_j^2] + \frac{dE[T_j^2]}{dj} \quad \text{and} \quad E[T_{j-1}^2] \approx E[T_j^2] - \frac{dE[T_j^2]}{dj}, \tag{B.13}$$

we have

$$j\mu_c E[T_j^2] \cong 2E[T_j] + (\lambda_s - j\mu_s)\frac{dE[T_j^2]}{dj}. \tag{B.14}$$

When $j = 1$, the combination of (B.13 and (B.14) yields:

$$\mu_c E[T_1^2] \cong 2E[T_1] + (\lambda_s - \mu_s)(E[T_2^2] - E[T_1^2]). \tag{B.15}$$

Also (B.12) becomes

$$(\lambda_s + j\mu_s + j\mu_c)E[T_1^2] = 2E[T_1] + \lambda_s E[T_2^2] + j\mu_s E[T_0^2]. \tag{B.16}$$

If there is no server available at the beginning of the service, $j = 0$, then the mean service time is given by:

$$E[T_0^2] = \frac{2}{\lambda_s^2} + \frac{2E[T_1]}{\lambda_s} + E[T_1^2]. \tag{B.17}$$

Solving (B.15), (B.16) and (B.17) simultaneously, we have

$$E[T_0^2] = \frac{2}{\lambda_s^2} + \frac{2\mu_s}{\lambda_s\mu_c} + \frac{\mu_s(2\mu_s^2 - \lambda_s^2) + 2\mu_c^2(\mu_s - \lambda_s)}{\lambda_s^2\mu_s\mu_c^2} \tag{B.18a}$$

$$E[T_1^2] = \frac{\mu_s(2\mu_s^2 - \lambda_s^2) + 2\mu_c^2(\mu_s - \lambda_s)}{\lambda_s^2\mu_s\mu_c^2} \tag{B.18b}$$

$$E[T_2^2] = \frac{\mu_s(2\mu_s^2 - \lambda_s^2) + 2\mu_c^2(\mu_s - \lambda_s)}{\lambda_s^2\mu_s\mu_c^2} - \frac{2}{\lambda_s^2} - \frac{2\mu_s}{\lambda_s\mu_c} \tag{B.18c}$$

For large j, we can assume

$$E[T_j^2] = \sum_{k=2}^{\infty} \frac{v_k}{j^k}. \tag{B.19}$$

Substituting identities (B.19) into (B.12), we have

$$(\lambda_s + j\mu_s + j\mu_c)\left[\frac{v_2}{j^2} + \frac{v_3}{j^3} + \frac{v_4}{j^4} + \cdots\right] \tag{B.20}$$

$$= 2E[T_j] + \lambda_s \left[\begin{array}{c} v_2\left(\frac{1}{j^2} - \frac{2}{j^3} + \frac{3}{j^4} - \frac{4}{j^5} + \frac{5}{j^6} - \cdots\right) \\ +v_3\left(\frac{1}{j^3} - \frac{3}{j^4} + \frac{6}{j^5} - \frac{10}{j^6} + \frac{15}{j^7} - \cdots\right) \\ +v_4\left(\frac{1}{j^4} - \frac{4}{j^5} + \frac{10}{j^6} - \frac{20}{j^7} + \frac{35}{j^8} - \cdots\right) \\ + \cdots \end{array}\right] \tag{B.21}$$

$$+ j\mu_s \left[\begin{array}{c} v_2\left(\frac{1}{j^2} + \frac{2}{j^3} + \frac{3}{j^4} + \frac{4}{j^5} + \frac{5}{j^6} + \cdots\right) \\ +v_3\left(\frac{1}{j^3} + \frac{3}{j^4} + \frac{6}{j^5} + \frac{10}{j^6} + \frac{15}{j^7} + \cdots\right) \\ +v_4\left(\frac{1}{j^4} + \frac{4}{j^5} + \frac{10}{j^6} + \frac{20}{j^7} + \frac{35}{j^8} + \cdots\right) \\ + \cdots \end{array}\right]. \tag{B.22}$$

Comparing coefficients of (B.20), we obtain the following set of equations:

$$v_2(\mu_s + \mu_c) = 2\tau_1 + v_2\mu_s \tag{B.23a}$$

$$v_2\lambda_s + v_3(\mu_s + \mu_c) = 2\tau_2 + v_2\lambda_s + 2v_2\mu_s + v_3\mu_s \tag{B.23b}$$

$$v_2\lambda_s + v_3\left(\mu_s + \mu_c\right) = 2\tau_2 + v_2\lambda_s + 2v_2\mu_s + v_3\mu_s \tag{B.23c}$$

$$v_3\lambda_s + v_4\left(\mu_s + \mu_c\right) = 2\tau_3 - 2v_2\lambda_s + v_3\lambda_s + 3v_2\mu_s + 3v_3\mu_s + v_4\mu_s \tag{B.23d}$$

$$\vdots$$

Solving this set of linear equations (B.23), we obtain:

$$\begin{aligned} v_2 =& \frac{2}{\mu_c^2} \\ v_3 =& \frac{6\mu_s}{\mu_c^3} \\ v_4 =& \frac{(8\mu_s - 6\lambda_s)}{\mu_c^3} + \frac{22\mu_s^2}{\mu_c^4} \\ \vdots& \end{aligned}$$

It follows that

$$E[T_j^2]\frac{2}{j^2\mu_c^2} + \frac{6\mu_s}{j^3\mu_c^3} + \frac{1}{j^4}\left(\frac{8\mu_s - 6\lambda_s}{\mu_c^3} + \frac{22\mu_s^2}{\mu_c^4}\right) + \cdots \tag{B.25}$$

for $j = 3, 4, 5, ..., N$

Bibliography

[1] N. Akar, N. Oguz, and K. Sohraby. Matrix-geometric solutions of M/G/1-type Markov chains: a unifying generalized state-space approach. *Selected Areas in Communications, IEEE Journal on*, 16(5):626 –639, Jun. 1998.

[2] L. Badia, M. Rossi, and M. Zorzi. SR ARQ packet delay statistics on Markov channels in the presence of variable arrival rate. *Wireless Communications, IEEE Transactions on*, 5(7):1639 –1644, Jul. 2006.

[3] M. Baykal-Gursoy and W. Xiao. Stochastic decomposition in $M/M/\infty$ queues with Markov modulated service rates. *Queueing Syst. Theory Appl.*, 48(1-2):75–88, Sept. 2004.

[4] R. Berry and E. Yeh. Cross-layer wireless resource allocation. *Signal Processing Magazine, IEEE*, 21(5):59 – 68, Sept. 2004.

[5] D. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall, second edition, 1992.

[6] O. J. Boxma and I. A. Kurkova. The M/G/1 queue with two service speeds. *Advances in Applied Probability*, 33(2):pp. 520–540, 2001.

[7] I. Cerutti, A. Fumagalli, and P. Gupta. Delay models of single-source single-relay cooperative ARQ protocols in slotted radio networks with poisson frame arrivals. *IEEE/ACM Trans. Netw.*, 16(2):371–382, Mar. 2008.

[8] J. Chamberland, H. Pfister, and S. Shakkottai. First-passage time analysis for digital communication over erasure channels with delay-sensitive traffic. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 399 –405, Oct. 2010.

[9] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *Communications, IEEE Transactions on*, 44(2):203 –217, Feb. 1996.

[10] M. Eisen and M. Tainiter. Stochastic variations in queuing processes. *Operations Research*, 11(6):922–927, 1963.

[11] E. O. Elliott. Estimates of error rates for codes on bursty-noise channels. *Bell System Tech. J.*, 42(9):1977 –97, Sept. 1963.

[12] N. Ewald and A. Kemp. Modelling the performance of a cross-layer TCP NewReno-HARQ system. *Int'l J. of Communications, Network and System Sciences*, 3(1):19–31, Jan. 2010.

[13] B. Fan, D.-M. Chiu, and J. Lui. Stochastic differential equation approach to model bittorrent-like p2p systems. In *Communications, 2006. ICC '06. IEEE International Conference on*, volume 2, pages 915 –920, Jun. 2006.

[14] B. Fritchman. A binary channel characterization using partitioned Markov chains. *Information Theory, IEEE Transactions on*, 13(2):221 –227, Mar. 1967.

[15] L. G. and R. V. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* Society for Industrial and Applied Mathematics, 1999.

[16] R. G. Gallager. *Information Theory and Reliable Communication.* John Wiley & Sons, Inc., New York, NY, USA, 1968.

[17] E. N. Gilbert. Capacity of a bursty-noise channel. *Bell System Tech. J.*, 39(9):1253 –65, Sept. 1960.

[18] A. Goldsmith and P. Varaiya. Capacity of fading channels with channel side information. *Information Theory, IEEE Transactions on*, 43(6):1986 –1992, Nov. 1997.

[19] N. Gunaseelan, L. Liu, J.-F. Chamberland, and G. Huff. Performance analysis of wireless Hybrid-ARQ systems with delay-sensitive traffic. *Communications, IEEE Transactions on*, 58(4):1262 –1272, Mar. 2010.

[20] L. Huang and T. Lee. Performance analysis of two-state wireless channel with lagging bound. In *Communications and Mobile Computing (CMC), 2011 Third International Conference on*, pages 246 –249, Mar. 2011.

[21] L. Huang and T. Lee. Queueing behavior of Hybrid ARQ wireless system with finite buffer capacity. In *Wireless and Optical Communications Conference (WOCC), 2012 21st Annual*, pages 32 –36, Mar. 2012.

[22] C.-D. Iskander and P. Takis Mathiopoulos. Analytical level crossing rates and average fade durations for diversity techniques in Nakagami fading channels. *Communications, IEEE Transactions on*, 50(8):1301 – 1309, Aug. 2002.

[23] W. C. Jakes and D. C. Cox, editors. *Microwave Mobile Communications*. Wiley-IEEE Press, New York, 1994.

[24] J. G. Kim and M. M. Krunz. Bandwidth allocation in wireless networks with guaranteed packet-loss performance. *IEEE/ACM Trans. Netw.*, 8:337–349, Jun. 2000.

[25] M. Krunz and J. G. Kim. Fluid analysis of delay and packet discard performance for QoS support in wireless networks. *Selected Areas in Communications, IEEE Journal on*, 19(2):384 –395, Feb. 2001.

[26] K. Lee and S. Chanson. Analysis of a delay-constrained Hybrid ARQ wireless system. *Communications, IEEE Transactions on*, 54(8):1514, Aug. 2006.

[27] T. Li, M. Chen, D.-M. Chiu, and M. Chen. Queuing Models for Peer-to-peer Systems. In *8th International Workshop on Peer-to-Peer Systems (IPTPS '09)*, Mar. 2009.

[28] S. Lin, D. Costello, and M. Miller. Automatic-repeat-request error-control schemes. *Communications Magazine, IEEE*, 22(12):5 –17, Dec. 1984.

[29] Q. Liu, S. Zhou, and G. Giannakis. Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design. *Wireless Communications, IEEE Transactions on*, 4(3):1142 – 1153, May 2005.

[30] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with applications to call admission. *J. ACM*, 44(3):366–394, May 1997.

[31] C. Lott, O. Milenkovic, and E. Soljanin. Hybrid ARQ: Theory, state of the art and future directions. In *Information Theory for Wireless Networks, 2007 IEEE Information Theory Workshop on*, pages 1 –5, Jul. 2007.

[32] S. Mahabhashyam and N. Gautam. On queues with Markov modulated service rates. *Queueing Systems*, 51(1-1):89–113, Oct. 2005.

[33] J. McDougall and S. Miller. Sensitivity of wireless network simulations to a two-state Markov model channel approximation. In *Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE*, volume 2, pages 697 – 701 Vol.2, 1-5 2003.

[34] I. L. Mitrany and B. Avi-Itzhak. A many-server queue with service interruptions. *Operations Research*, 16(3):628–638, 1968.

[35] M. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Johns Hopkins University Press, Baltimore, 1981.

[36] M. F. Neuts. A queue subject to extraneous phase changes. *Advances in Applied Probability*, 3(1):pp. 78–119, 1971.

[37] E. Perel and U. Yechiali. Queues where customers of one queue act as servers of the other queue. *Queueing Syst. Theory Appl.*, 60:271–288, Dec. 2008.

[38] C. Pimentel, T. Falk, and L. Lisboa. Finite-state Markov modeling of correlated Rician-fading channels. *Vehicular Technology, IEEE Transactions on*, 53(5):1491 – 1501, Sept. 2004.

[39] P. Purdue. The m/m/1 queue in a Markovian environment. *Operations Research*, 22(3):pp. 562–569, 1974.

[40] D. Qiu and R. Srikant. Modeling and performance analysis of bittorrent-like peer-to-peer networks. *SIGCOMM Comput. Commun. Rev.*, 34:367–378, Aug. 2004.

[41] G. J. K. Regterschot and J. H. A. d. Smit. The queue M/G/1 with markov modulated arrivals and services. *Mathematics of Operations Research*, 11(3):pp. 465–483, 1986.

[42] J. S. Rosenthal. Convergence rates for markov chains. *SIAM Review*, 37(3):pp. 387–405, 1995.

[43] M. Rossi, L. Badia, and M. Zorzi. On the delay statistics of SR ARQ over Markov channels with finite round-trip delay. *Wireless Communications, IEEE Transactions on*, 4(4):1858 – 1868, Jul. 2005.

[44] P. Sadeghi, R. Kennedy, P. Rapajic, and R. Shams. Finite-state Markov modeling of fading channels - a survey of principles and applications. *Signal Processing Magazine, IEEE*, 25(5):57–80, 2008.

[45] F. Schmitt, F. Schmitt, F. Rothlauf, and F. Rothlauf. On the importance of the second largest eigenvalue on the convergence rate of genetic algorithms. Technical report, Proceedings of the 14th Symposium on Reliable Distributed Systems, 2001.

[46] J.-B. Seo, S.-Q. Lee, N.-H. Park, H.-W. Lee, and C.-H. Cho. Queueing behavior of a Type-II Hybrid-ARQ in a TDMA system over a Markovian channel. In *Wireless Networks, Communications and Mobile Computing, 2005 International Conference on*, volume 1, pages 362 – 367 vol.1, Jun. 2005.

[47] C. Shannon. The mathematical theory of communications. *Bell System Technical Journal*, 27:379–623, 1948.

[48] R. Susitaival, S. Aalto, and J. Virtamo. Analyzing the dynamics and resource usage of p2p file sharing by a spatio-temporal model. In V. Alexandrov, G. van Albada, P. Sloot, and J. Dongarra, editors, *Computational Science ICCS 2006*, volume 3994 of *Lecture Notes in Computer Science*, pages 420–427. Springer Berlin / Heidelberg, 2006.

[49] T. Takine. Single-server queues with Markov-modulated arrivals and service speed. *Queueing Systems*, 49(1):7–22, 2005.

[50] H. S. Wang and N. Moayeri. Finite-state Markov channel-a useful model for radio communication channels. *Vehicular Technology, IEEE Transactions on*, 44(1):163 –171, Feb. 1995.

[51] H. White and L. S. Christie. Queuing with preemptive priorities or with breakdown. *Operations Research*, 6(1):79–95, 1958.

[52] R. W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[53] D. Wu, C. Liang, Y. Liu, and K. W. Ross. View-upload decoupling: A redesign of multi-channel p2p video systems. In *IEEE INFOCOM*, pages 2726–2730, 2009.

[54] D. Wu and R. Negi. Effective capacity: a wireless link model for support of quality of service. *Wireless Communications, IEEE Transactions on*, 2(4):630 – 643, Jul. 2003.

[55] X. Yang and G. de Veciana. Service capacity of peer to peer networks. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 4, pages 2242 – 2252 vol.4, Mar. 2004.

[56] U. Yechiali. A queuing-type birth-and-death process defined on a continuous-time Markov chain. *Operations Research*, 21(2):pp. 604–609, 1973.

[57] U. Yechiali and P. Naor. *Queueing Problems with Heterogeneous Arrivals and Service*. Operations research, statistics and economics mimeograph series. Defense Technical Information Center, 1969.

[58] H. Zhang, J. Wang, M. Chen, and K. Ramchandran. Scaling peer-to-peer video-on-demand systems using helpers. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3053 –3056, Nov. 2009.

[59] L. Zhang and T. Lee. Performance analysis of wireless fair queuing algorithms with compensation mechanism. In *Communications, 2004 IEEE International Conference on*, volume 7, pages 4202 – 4206 Vol.7, 20-24 2004.

[60] Q. Zhang and S. Kassam. Finite-state Markov model for rayleigh fading channels. *Communications, IEEE Transactions on*, 47(11):1688 –1692, Nov. 1999.

[61] Y.-P. Zhou and N. Gans. A single-server queue with Markov modulated service times. Center for financial institutions working papers, Wharton School Center for Financial Institutions, University of Pennsylvania, 1999.

[62] M. Zorzi and R. Rao. Arq error control for delay-constrained communications on short-range burst-error channels. In *Vehicular Technology Conference, 1997, IEEE 47th*, volume 3, pages 1528 –1532 vol.3, may 1997.

[63] M. Zorzi and R. Rao. Lateness probability of a retransmission scheme for error control on a two-state Markov channel. *Communications, IEEE Transactions on*, 47(10):1537 –1548, Oct. 1999.

[64] M. Zorzi, R. Rao, and L. Milstein. Error statistics in data transmission over fading channels. *Communications, IEEE Transactions on*, 46(11):1468 –1477, Nov. 1998.