https://dx.doi.org/10.1287/opre.2017.1707

Management of a Shared Spectrum Network in Wireless Communications

Shining Wu*, Jiheng Zhang†, Rachel Q. Zhang†
*Department of Logistics and Maritime Studies
The Hong Kong Polytechnic University, Hong Kong
†Department of Industrial Engineering and Decision Analytics
The Hong Kong University of Science and Technology, Hong Kong
sn.wu@polyu.edu.hk, jiheng@ust.hk, rzhang@ust.hk

We consider a band of the electromagnetic spectrum with a finite number of identical channels shared by both licensed and unlicensed users. Such a network differs from most many-server, two-class queues in service systems, including call centers, because of the restrictions imposed on the unlicensed users in order to limit interference to the licensed users. We first approximate the key performance indicators, namely the throughput rate of the system and the delay probability of the licensed users under the asymptotic regime, which requires the analysis of both scaled and unscaled processes simultaneously using the averaging principle. Our analysis reveals a number of distinctive properties of the system. For example, sharing does not affect the level of service provided to the licensed users in an asymptotic sense even when the system is critically loaded. We then study the optimal sharing decisions of the system to maximize the system throughput rate while maintaining the delay probability of the licensed users below a certain level when the system is overloaded. Finally, we extend our study to systems with time-varying arrival rates and propose a diffusion approximation to complement our fluid one.

Key words: spectrum management; many-server queues; fluid approximation; averaging principle
History: Manuscript https://doi.org/10.1287/opre.2017.1707

1. Introduction

The radio spectrum refers to the range of frequencies suitable for wireless communications in television and radio broadcasting, aviation, public safety, cell phones, and so on. Until recently, spectrum regulatory bodies including the Federal Communications Commission (FCC) in the US and the European Telecommunications Standards Institute have always allocated spectrum bands exclusively to certain service providers whose users are referred to as primary or licensed users, often based on the radio technologies available at the time of allocation. Such static spectrum allocation mitigates interference to essential services, yet it creates underutilization of the allocated spectrum, which can be below 20% even during high demand periods in certain geographic areas. For instance, during the high demand period of a political convention held in New York City in 2004, only about 13% of the allocated spectrum was utilized (Prasad et al. 2010). Studies conducted by the FCC, universities, and industry also revealed that a major part of the spectrum is not fully

utilized most of the time. On the other hand, over the past decades, the convergence of voice and data in wireless communications triggered by the convergence of wireless and Internet technologies has led to an explosion in the number of bits transmitted over the air (Biglieri et al. 2013). Since it is usually difficult to open up higher frequency bands for mobile applications as transmission becomes less reliable in those bands, the existing radio spectrum for data transmission is reaching its capacity.

A natural approach to alleviate the artificial scarcity of spectrum due to static allocation is to allow opportunistic use of temporarily idle channels by unlicensed or secondary users to increase the throughput of already allocated spectrum. This is referred to as opportunistic spectrum access (Hossain et al. 2009). However, allowing unlicensed users access may cause interference to existing licensed users. Thus, such paradigm of operation requires (1) the knowledge of the state of frequency bands (e.g., channel availability, queues, etc.) in real time and (2) an effective control mechanism to govern spectrum usage by unlicensed users, which led to the development of the concept of cognitive radio, first introduced by Mitola and Maguire (1999). Using advanced radio and signal processing technology, cognitive radio is a software-defined radio device that can intelligently sense and explore the spectrum environment, track changes, communicate information among different transceivers and react according to a control mechanism (Hossain et al. 2009). It is widely regarded as one of the most promising technologies for future wireless communications and may potentially mitigate, through dynamic spectrum access, the problem of radio spectrum scarcity.

It is obvious that implementation of a cognitive radio network involves both technological and operational issues, yet much of the research is focused on the former (see Section 2.1 for some relevant literature). In this paper, we focus on the operational issues by considering a band of spectrum with multiple identical channels shared by both licensed and unlicensed users. Since the spectrum has already been allocated to the licensed users and it is usually difficult to set aside a subset of channels for either groups in reality for technical reasons, we assume all the channels are accessible by both licensed and unlicensed users as in most existing literature in electrical engineering. Furthermore, although concurrent transmission is allowed in some networks under which the main concern is technological (e.g., the power level at which an unlicensed user is allowed to transmit), we focus on systems where each channel serves only one user at a time, referred to as the interweave paradigm (Biglieri et al. 2013). Thus, the network considered is a two-class queue served by a single pool of homogeneous servers as in applications in service systems such as call centers and healthcare but with some distinctive features due to the restrictions imposed on the unlicensed users (Hossain et al. 2009). (1) When all the channels are occupied upon arrival, a licensed user will join a queue along with other waiting licensed users who will be served first-in-first-out (FIFO) as soon as a channel becomes available, while an unlicensed user will join a queue

along with other waiting unlicensed users and will only be allowed to sense channel availability *periodically*. An unlicensed user can only occupy a channel when an available channel is detected and no licensed users are waiting, and may also abandon the system every time he senses but finds no available channel. Such a queue where users wait for retrial is referred to as an orbit queue in the queueing literature and is common in computer and communications networks (Artalejo and Gómez-Corral 2008). (2) When in transmission, a licensed user can transmit until his service requirement is fulfilled, while an unlicensed user is only allocated a fixed amount of time, referred to as *a service session*, approaching the end of which he has to stop transmission to sense the environment as sensing cannot occur simultaneously with data transmission. The unlicensed user will be allowed to continue for another service session only if he senses no waiting licensed users. Otherwise, he has to release the channel and join the orbit queue along with other unlicensed users or abandon the system if he needs more time. Note that data transmission can be interrupted and resumed, hence more complicated control policies than those in call centers are allowed, which leads to new managerial insights.

Assuming that perfect sensing can be achieved in a fixed amount of time and both licensed and unlicensed users arrive according to Poisson processes, we first perform in-depth analysis on the key performance indicators in the management of shared spectrum networks, namely the delay probability of the licensed users and the system throughput rate. We then focus on the restrictions that need to be imposed on the unlicensed users when in service and waiting, i.e., the length of a service session and the sensing frequency while waiting. Intuitively, the longer a service session is, the less sensing an unlicensed user needs to perform and hence a higher system throughput rate. Yet, longer service sessions can cause more interference to the licensed users. Likewise, the more frequently an unlicensed user senses channel availability while waiting, the sooner he is able to find an available channel but the more interference he causes to the licensed users. Thus, there is a tradeoff between the throughput rate and the level of interference to the licensed users when deciding on the length of a service session and the sensing frequency. The goals of this research are to answer the following questions: (1) Should a given band of spectrum be shared with unlicensed users? (2) When sharing is permitted, how long should unlicensed users be allowed to transmit each time they occupy a channel and how frequently should they be allowed to sense channel availability while waiting? (3) Under what conditions is sharing more beneficial? (4) How will the decision change with uncertain arrival rates or time-varying arrivals?

Since the band of spectrum considered usually consists of hundreds or thousands of channels, we can treat the system as a large network, and approximate the performance under the asymptotic regime as in Gupta and Kumar (2000) and El Gamal et al. (2006). Due to the restrictions imposed on the unlicensed users when in service and waiting, we need to analyze both scaled and unscaled

processes *simultaneously* using the averaging principle, i.e., approximating the unscaled process by its long-run average. We then formulate the problem as finding the optimal restrictions on the unlicensed users to maximize the throughput rate while maintaining the delay probability of licensed users below a certain level. Our main findings are as follows:

1. **Sensing frequency of the unlicensed users while waiting:** Surprisingly, sensing frequency does not affect the system performance asymptotically as long as the unlicensed users are required to sense channel availability, which takes time and prevents them from occupying idle channels instantaneously. Thus, there is no need to impose any restriction on the sensing frequency from the operational perspective. The decision thus should primarily be based on technological concerns, for instance, power consumption associated with each sensing activity.
2. **The length of a service session:** Intuitively, shorter service sessions should cause less interference to and hence lower the delay probability of the licensed users. However, with shorter service sessions, the unlicensed users need more service sessions to finish their service and hence need to perform more sensing activities while occupying a channel. Thus, shorter service sessions do not always improve the delay probability.
3. **Optimal sharing decisions:** When the system is under or critically loaded, the interference of the unlicensed users to the licensed users is negligible and there is no need to impose a restriction on the service process of the unlicensed users either. That is, allowing the unlicensed users to complete their transmissions without restriction will not cause any interference to licensed users asymptotically as the delay probability is 0. This result is very different from that of most non-preemptive queueing systems under which the delay probability is strictly between 0 and 1 when the system is critically loaded.

   When the system is overloaded, the delay probability of the licensed users is quasi-convex in the length of the service sessions of the unlicensed users, strictly between 0 and 1 and increasing in the load. Thus, a restriction on the service process of the unlicensed users should be imposed only when the load is above a threshold. Furthermore, a shorter service session should be allocated as the load increases until spectrum sharing is no longer feasible.

   The insight that it is possible to improve spectrum utilization while guaranteeing a very high service level, expected by licensed users in practice, is very encouraging news. Thus, spectrum sharing can potentially be a socially optimal solution to alleviating spectrum scarcity.
4. For a given system load, a shorter service session should be allocated to the unlicensed users (1) as the proportion of the licensed users increases, (2) if there are fewer licensed users with longer service times, or (3) if there are more unlicensed users with shorter service times. As the service session shortens, more unlicensed users will abandon the system, which lowers the throughput rate under scenarios (1) and (2). Therefore, spectrum sharing is beneficial to

systems with a smaller proportion of licensed users or a large number of licensed users with shorter service times.

5. When the arrival rates are time varying, a shorter service session should be allocated to the unlicensed users during busy periods. Although optimal control requires continuous adjustment in real time, near-optimal control can be accomplished with occasional adjustments.

To the best of our knowledge, this is the first comprehensive study of a shared network in wireless communications. Although there have been some attempts by researchers in electrical engineering using relatively simple queueing models, our model captures many more of the features of such a system. We are able to uncover complicated system dynamics and obtain managerial insights different from those drawn from the many well-studied service systems. Our work not only opens the door for new applications of existing queueing theory in wireless communications, but may also stimulate the development of new methodologies.

The remainder of this paper is organized as follows. We review the relevant literature in both electrical engineering and queueing theory in the next section and describe the problem of dynamic spectrum sharing in detail in Section 3. In Section 4, we provide a fluid approximation and study the optimal sharing decisions of the system. In Section 5, we offer the intuition behind the construction of the fluid model and give justifications for the fluid approximation. We extend our analysis to systems with time-varying arrival rates and discuss a diffusion scaled approximation in Section 6. We conclude our paper and provide some future research directions in Section 7. The proofs can all be found in the Appendix.

2. Literature Review

In this section, we will first provide some background on the research on opportunistic spectrum access, mostly in electrical engineering. Since we will model a shared network as a multi-class, many-server queue where the unlicensed users join an orbit queue and analyze it using the averaging principle, we will review the relevant literature in queueing theory and its applications.

2.1. On Opportunistic Spectrum Access

Most of the work on opportunistic spectrum access focuses on the technological issues such as the sensing technology to detect idle channels (Mishra et al. 2006), signal encoding (Devroye et al. 2006) and the control of the transmit power to limit interference (Bansal et al. 2008). For research on various technological issues associated with cognitive radio, readers may refer to Akyildiz et al. (2006) and Goldsmith et al. (2009).

Research on the operational issues under simplified settings, however, remains scant. Huang et al. (2008) perform an analytical study on a single-channel system with one licensed and one

unlicensed user, as well as numerical studies on a multi-channel system. They also consider the decisions on the sensing frequency of unlicensed users and how long unlicensed users should be allowed to transmit in their numerical study. Zhao et al. (2008) study the optimal access strategy of an unlicensed user based on the sensing outcome given that each channel has already been assigned to a specific licensed user, while Capar et al. (2002) compare the system performance in terms of bandwidth utilization and blocking probability when a licensed user can be assigned to any channel randomly or in a controlled way.

For a more comprehensive picture of the various issues in dynamic spectrum management and cognitive radio networks, readers may refer to Hossain et al. (2009) and Biglieri et al. (2013).

2.2. On Queueing Theory and Applications

Multi-Class, Many-Server Queues Since a band of spectrum consists of hundreds or thousands of channels and there are both licensed and unlicensed users, the literature of multi-class, many-server queues is relevant. The study of many-server queues was substantiated by the seminal work of Halfin and Whitt (1981), who derive the steady-state distribution of the diffusion limits and establish the square root law describing the relationship between the system load and delay probability. The mathematical insights of the square root law have since been extended and widely adopted in the daily management of call centers around the world. Later, Puhalskii and Reiman (2000) extend the study to multi-class models.

There is a large body of work on multi-class, many-server systems due to their applications in call centers, manufacturing and computer-communication systems with a focus on asymptotic optimal control of the underlying systems. For example, Atar et al. (2004) study asymptotic optimal schedule policies, Gurvich and Whitt (2009) propose a family of queue-and-idleness-ratio rules for routing and scheduling, and Maglaras and Zeevi (2004, 2005) examine the pricing, capacity sizing and admission control decisions in a differentiated service system with guaranteed (high priority) and best-effort (low priority) users. Our model differs from the existing work in that the service (i.e., data transmission) of the unlicensed users may be fulfilled after multiple interruptions, which is not the case in most other applications.

Since the service of unlicensed users may be interrupted by waiting licensed users, the literature on queues with service interruption caused by preemptive priority, which dates back to White and Christie (1958) in single server settings, is also relevant. For a review on some of the early work, we refer the reader to Jaiswal (1968). Among the existing work, most focuses on characterizing the steady state distributions of the queue length, the sojourn time and so on for a given priority discipline. For example, Brosh (1969) derives the expressions for the expected time from arrival to inception of service and provides bounds for the expected sojourn time for each class when

all classes have the same service rates. Buzen and Bondi (1983) obtain the exact expressions for the mean sojourn times when all classes have the same service rates and provide approximations when different classes have different service rates. Recently, Wang et al. (2015) conduct the exact analysis of the steady state of a preemptive $M/M/c$ queue when different classes have different service rates. In our paper, we focus on the control of the service process of the unlicensed users, i.e., how their service processes should be interrupted.

Orbit Queues Since the unlicensed users join an orbit queue in our setting, the literature along this line is also relevant. Yang and Templeton (1987) and Falin and Templeton (1997) offer a survey and a comprehensive summary of the earlier papers, respectively. Later, Mandelbaum et al. (2002) provide an analytical approximation to the key performance of a many-server queueing system with abandonment and retrials under an asymptotic regime. In all these papers, even though customers may join an orbit queue for retrial if they cannot be served immediately upon arrival, their service cannot be interrupted once started.

Recently, a number of studies consider systems where customers may require repeat service due to unresolved or new issues. For instance, de Véricourt and Zhou (2005) and Zhan and Ward (2014) study a customer-routing problem in call centers with callbacks, while de Véricourt and Jennings (2008) and Yom-Tov and Mandelbaum (2014) examine a staffing problem for membership services and healthcare systems where customers may require multiple rounds of service. These systems differ from ours in that customers will wait in a FIFO queue for retrial if the systems are busy upon arrival, although they will first join an orbit queue after they have had a round of service. Allowing the unlicensed users to retry and join an orbit queue as in our setting significantly complicates the analysis since there may be a large number of customers switching frequently between being in service and being in the orbit queue.

The Averaging Principle Only a few studies in the queueing literature have required the use of the averaging principle. Building on a fundamental theory of the averaging principle by Kurtz (1992), Hunt and Kurtz (1994) study martingales and related random measures of large loss networks. Whitt (2002) summarizes the early studies on scheduling multi-class queues using the averaging principle. Recently, a series of studies by Perry and Whitt (2011a,b, 2013) has applied the averaging principle to obtain both the fluid and diffusion limits for an overloaded X model of many-server queues, and to derive insights about the asymptotic optimal control of the system. Pang and Perry (2015) apply the averaging principle to obtain a logarithmic safety staffing rule for call centers with call blending. We adopt some of the methodologies developed by Hunt and Kurtz (1994) and Perry and Whitt (2011a).

3. Problem Description and Assumptions

3.1. The Sharing Network and Performance Measures

We consider a band of spectrum consisting of n identical channels shared by both licensed and unlicensed users, denoted as user types 1 and 2 respectively, and each channel can only be occupied by one user at a time. That is, concurrent transmission is not allowed. Furthermore, we assume that perfect sensing can be achieved in a fixed amount of time $\frac{1}{\mu_s}$ by an unlicensed user. Type i users arrive according to a Poisson process with the rate λ_i^n and require an exponential amount of service time with the rate μ_i, $i = 1, 2$. If there is an available channel, an arriving licensed user will occupy it immediately until his service requirement is fulfilled. Otherwise, he will join a queue along with other waiting licensed users who will be served FIFO as the channels become available.

Next, we describe the service and waiting processes of the unlicensed users in the shared network illustrated in Figure 1 where $I^n(t)$ is the number of idle channels and $Q_i^n(t)$ is the queue length of type i users at time t. According to the policy,

$$Q_1^n(t)I^n(t) = 0. \tag{1}$$

Upon arrival, an unlicensed user will occupy a channel if there is one available. Otherwise, he will join an orbit queue along with other waiting unlicensed users with probability $1-\phi$ or abandon the system.

- The service process: Once he occupies a channel, an unlicensed user is allocated a fixed amount of uninterrupted time, referred to as a *service session* (Liu and Wang 2010), regardless of his service requirement. If he needs more time and finds no licensed user waiting at the end of a session through sensing, he is allowed to continue for another service session. Since sensing cannot occur simultaneously with data transmission and must be interweaved, he needs to devote the last $\frac{1}{\mu_s}$ amount of time in each service session to sense the environment if he needs more time. Hence, we denote the length of a service session by $\frac{1}{\mu_t} + \frac{1}{\mu_s}$ where $\frac{1}{\mu_t}$ is the amount of time allowed for transmission in a service session. If the unlicensed user completes his transmission within $\frac{1}{\mu_t}$ amount of time in a session, he will release the channel without sensing and leave the system. Otherwise, he will have to sense the environment and his service will be interrupted if he finds a waiting licensed user, in which case he will join the orbit queue with probability $1-\phi$ or abandon the system.
- The waiting process: While waiting in the orbit queue, an unlicensed user will only be allowed to sense channel availability periodically. Let $\frac{1}{\theta}$ denote the time between sensing activities, which includes the time needed for sensing channel availability. After each sensing activity, he will occupy a channel if he finds an idle one. Otherwise, he will abandon the system with probability ϕ, or stay in the orbit queue for another sensing activity with probability $1-\phi$.

As one can see, the network has its distinctive characteristics which are not present in most existing multi-class, many-server queueing systems due to the restrictions on the service and waiting processes of the unlicensed users, i.e., the transmission time $\frac{1}{\mu_t}$ in a service session and the sensing frequency θ. The data transmission of the unlicensed users can be interrupted and resumed for any number of times and sensing for channel availability by the unlicensed users in the queue is only allowed periodically. As a result, an unlicensed user may abandon the system upon arrival, after spending some time in the queue without receiving any service, or after receiving partial service. Furthermore, each unlicensed user in the orbit queue needs to sense channel availability independently, which guarantees certain idleness in the system even when there are waiting unlicensed users. These features are new in the queueing literature and interesting, yet significantly complicate the analysis.

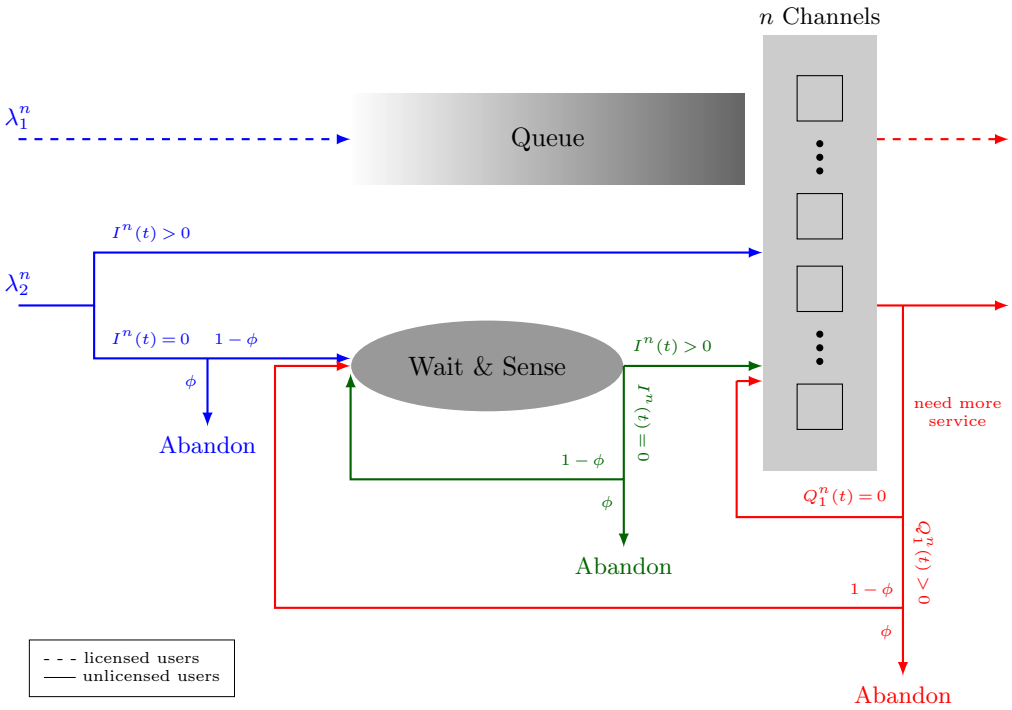


Figure 1 The spectrum sharing network

The performance measures we are concerned with are the throughput rate of the system and the probability that all the channels are occupied upon the arrival of a licensed user, referred to as the delay probability. The goal is to find the transmission time $\frac{1}{\mu_t}$ in a service session and the sensing frequency θ of the unlicensed users that maximize the throughput of the unlicensed users while guaranteeing the delay probability of the licensed users below a certain level.

3.2. Modeling Assumptions

Since the problem is analytically intractable, we will first approximate the deterministic transmission time, sensing time, and the time between consecutive sensing activities in the orbit queue by the exponential distributions with the same means. Table 1 presents a simulation study of the delay probability of the licensed users and the throughput rate of the unlicensed users with deterministic times and exponential times when $\lambda_1 = 0.2$, $\lambda_2 = 0.9$, $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 1$, $\frac{1}{\mu_s} = 0.001$, $\theta = 0.4$, and $\phi = 0.5$. For $\frac{1}{\mu_t} \in \{\infty, 0.6, 0.2\}$, we set $n = 100, 500, 1000, 2000, 4000$ and let $\lambda_i^n = n\lambda_i$. We report the means and 0.95 confidence intervals of the delay probabilities and throughput rates. As one can see, approximating the deterministic times by the exponential times does not reduce the accuracy very much, especially when n is large as in our application where n is in the hundreds or thousands.

$\frac{1}{\mu_t}$	n	delay probability		throughput rate	
		Deterministic	Exponential	Deterministic	Exponential
∞	100	$0.2528 \pm .0036$	$0.2531 \pm .0044$	$0.7682 \pm .0014$	$0.7678 \pm .0024$
	500	$0.2120 \pm .0014$	$0.2142 \pm .0019$	$0.7928 \pm .0007$	$0.7915 \pm .0012$
	1000	$0.2075 \pm .0017$	$0.2075 \pm .0018$	$0.7957 \pm .0008$	$0.7957 \pm .0009$
	2000	$0.2045 \pm .0009$	$0.2044 \pm .0008$	$0.7975 \pm .0005$	$0.7974 \pm .0005$
	4000	$0.2014 \pm .0010$	$0.2021 \pm .0008$	$0.7992 \pm .0004$	$0.7988 \pm .0004$
	Fluid	0.1995		0.8000	
0.6	100	$0.2360 \pm .0040$	$0.2314 \pm .0036$	$0.7656 \pm .0021$	$0.7662 \pm .0022$
	500	$0.1979 \pm .0009$	$0.1953 \pm .0011$	$0.7901 \pm .0007$	$0.7899 \pm .0005$
	1000	$0.1906 \pm .0013$	$0.1892 \pm .0007$	$0.7949 \pm .0006$	$0.7939 \pm .0005$
	2000	$0.1872 \pm .0006$	$0.1854 \pm .0010$	$0.7968 \pm .0004$	$0.7964 \pm .0005$
	4000	$0.1854 \pm .0003$	$0.1838 \pm .0006$	$0.7979 \pm .0002$	$0.7973 \pm .0003$
	Fluid	0.1813		0.7987	
0.2	100	$0.2262 \pm .0031$	$0.2259 \pm .0033$	$0.7643 \pm .0023$	$0.7641 \pm .0019$
	500	$0.1916 \pm .0024$	$0.1918 \pm .0020$	$0.7878 \pm .0014$	$0.7871 \pm .0011$
	1000	$0.1860 \pm .0010$	$0.1855 \pm .0010$	$0.7914 \pm .0005$	$0.7914 \pm .0007$
	2000	$0.1820 \pm .0011$	$0.1820 \pm .0010$	$0.7940 \pm .0007$	$0.7938 \pm .0006$
	4000	$0.1802 \pm .0005$	$0.1804 \pm .0007$	$0.7954 \pm .0003$	$0.7949 \pm .0004$
	Fluid	0.1784		0.7960	

Table 1 Comparison of performance measures with deterministic vs. exponential times.

With the exponential times mentioned above, the probability that an unlicensed user will complete his transmission in a service session is given by $p = \frac{\mu_2}{\mu_2 + \mu_t}$. Furthermore, the actual amount of time an unlicensed user will occupy a channel in each service session follows a phase-type distribution with mean

$$\frac{1}{\mu} = \frac{1}{\mu_2 + \mu_t} + (1 - p) \cdot \frac{1}{\mu_s} = \frac{\mu_t + \mu_s}{(\mu_2 + \mu_t)\mu_s}, \tag{2}$$

which is less than the allocated session time $\frac{1}{\mu_t} + \frac{1}{\mu_s}$. If we let $Z_i^n(t)$ denote the number of channels occupied by type i users at time t, the instantaneous throughput rate at time t is given by $p\mu Z_2^n(t)$.

With hundreds or thousands of channels in a band of spectrum, performing an analytical study of the shared network under a large system scaling to be defined below in Definition 1 is not only for technical tractability, but also appropriate.

Definition 1 (Asymptotic Regime). There exist positive real numbers λ_i, $i=1,2$, such that

$$\lim_{n\to\infty}\frac{\lambda_i^n}{n}=\lambda_i, \text{ and } \frac{\lambda_1}{\mu_1}<1.$$

Here λ_i represents the size of type i users. Different cognitive radio networks have different proportions of licensed and unlicensed users. In IEEE 802.22 wireless regional area networks, unlicensed users outnumber licensed users (Zhang et al. 2009, Jia et al. 2008), i.e., $\lambda_2>\lambda_1$, while in TV white space networks, licensed users are the majority (van de Beek et al. 2012), i.e., $\lambda_1>\lambda_2$. In Gong et al. (2015), the licensed users (from a down-link cellular system) and the unlicensed users (from an *ad hoc* network) have comparable numbers, i.e., $\lambda_1\approx\lambda_2$.

Under the asymptotic regime, we will add a bar to the existing notation to represent the scaled processes in our model, e.g., $\bar{Q}_i^n(t)=\frac{Q_i^n(t)}{n}$, and use the lower case, e.g., $q_i(t)$, to represent the corresponding fluid model, which will be proven to be the fluid limit of the scaled processes.

4. Main Results and Insights

Under the asymptotic regime, the processes involved are scaled and then approximated by tractable ones that preserve the relevant information about the system performance. As in most multi-class queueing systems, the queue length of the licensed users, who have a higher priority, will vanish asymptotically. This is not a problem if the queue length of the licensed users does not affect the users in service in an asymptotic sense, which is the case in most applications, and one can still obtain the managerial insights by analyzing the limit of scaled processes alone. However, whether the number of waiting license users is asymptotically small or exactly zero is important in our setting as it determines whether an unlicensed user should vacate a channel but the scaled processes fail to preserve such important information. Thus, the analysis requires information from both scaled and unscaled processes, involves tracking the two processes simultaneously, and needs to use the averaging principle. These requirements are rare in the literature with only a few exceptions such as Perry and Whitt (2011a), Luo and Zhang (2013), Pang and Perry (2015).

In this section, we first introduce our fluid model $x(t)=(z_1(t),q_1(t),z_2(t),q_2(t))$ which will be used to approximate the stochastic process $X^n(t)=(Z_1^n(t),Q_1^n(t),Z_2^n(t),Q_2^n(t))$ in our system with the justifications to be provided in Section 5. We then derive the steady-state performance and study the optimal sharing decisions of the system in the steady state using the fluid approximations.

4.1. The Fluid Model

DEFINITION 2 (FLUID MODEL). The process $x(t) = (z_1(t), q_1(t), z_2(t), q_2(t))$ evolves according to the constraint

$$0 = [1 - z_1(t) - z_2(t)]q_1(t), \tag{3}$$

and the following differential equations

$$z_1'(t) = [1 - \beta(t)]\lambda_1 + \alpha(t)[\mu_1 z_1(t) + \mu z_2(t)] - \mu_1 z_1(t), \tag{4}$$

$$q_1'(t) = \beta(t)\lambda_1 - \alpha(t)[\mu_1 z_1(t) + \mu z_2(t)], \tag{5}$$

$$z_2'(t) = [1 - \beta(t)][\lambda_2 + \theta q_2(t)] - [p + \alpha(t)(1-p)]\mu z_2(t), \tag{6}$$

$$q_2'(t) = (1-\phi)\beta(t)[\lambda_2 + \theta q_2(t)] + (1-\phi)\alpha(t)(1-p)\mu z_2(t) - \theta q_2(t), \tag{7}$$

where $\beta(t)$ and $\alpha(t)$ depend on how constraint (3) is met. If $q_1(t) > 0$, then $\beta(t) = \alpha(t) = 1$; if $z_1(t) + z_2(t) < 1$, then $\beta(t) = \alpha(t) = 0$; otherwise,

$$\beta(t) = \min\left\{\left(\frac{[\lambda_1 + \lambda_2 + \theta q_2(t) - \mu_1 z_1(t) - p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t)]}{[\lambda_1 + \lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + \mu z_2(t)] - \lambda_1[\mu_1 z_1(t) + p\mu z_2(t)]}\right)^+, 1\right\}, \tag{8}$$

$$\alpha(t) = \min\left\{\frac{\lambda_1 \beta(t)}{\mu_1 z_1(t) + \mu z_2(t)}, 1\right\}. \tag{9}$$

The fluid model defined above is built on the evolution of the system described in Section 3. As we will explain in Section 5 and define formally in Appendix B, $\beta(t)$ is the instantaneous delay probability of the licensed users and $\alpha(t)$ is the instantaneous probability that an unlicensed user has to release the channel after a service session (i.e., there are waiting licensed users in the system), referred to as the interruption probability, under the fluid model. Thus, the differential equations (4)–(7) are quite intuitive. Take equation (4) for an example. The rate of increase in $z_1(t)$ consists of two parts: (1) When the licensed users arrive (at the rate λ_1), there is an available channel (with probability $1 - \beta(t)$); (2) When the licensed users finish service (at the rate $\mu_1 z_1(t)$) or the unlicensed users finish a service session (at the rate $\mu z_2(t)$); there are waiting licensed users (with probability $\alpha(t)$). The rate of decrease in $z_1(t)$ is $\mu_1 z_1(t)$, which is the rate the licensed users occupying the channels finish service. For equation (7), the rate of increase in $q_2(t)$ consists of two parts: (1) When the unlicensed users arrive or those in the orbit queue perform sensing (at the rate $\lambda_2 + \theta q_2(t)$), they find all channels occupied (with probability $\beta(t)$) but decide not to abandon the system (with probability $1 - \phi$); (2) When the unlicensed users finish a service session (at the rate $\mu z_2(t)$), they need another one (with probability $1 - p$) and find licensed users waiting (with probability $\alpha(t)$) but do not abandon the system (with probability $1 - \phi$). The rate of decrease in $q_2(t)$ is $\theta q_2(t)$, which is the rate the unlicensed users in queue sense for available channels.

When $z_1(t)+z_2(t)<1$ or $q_1(t)>0$, the system dynamics is quite simple and resembles that of the many-server queues in call center applications. For example, when $z_1(t)+z_2(t)<1$, the differential equations (4)–(7) reduce to $q_1(t)\equiv 0$ and

$$\begin{aligned} z_1'(t) &= \lambda_1 - \mu_1 z_1(t), \\ z_2'(t) &= \lambda_2 + \theta q_2(t) - p\mu z_2(t), \\ q_2'(t) &= -\theta q_2(t). \end{aligned}$$

Otherwise, the system dynamics is more complicated. Moreover, the process $x(\cdot)$ can move back and forth among different cases, which makes the analysis even more challenging as shown in Appendix A.

Despite the complexity, the fluid model can be solved numerically. Furthermore, we can obtain the steady state of the fluid model in Theorem 1 to approximate the steady state of the original system. For example, $\beta := \lim_{t\to\infty} \beta(t)$ and $TH_2 := \lim_{t\to\infty} p\mu z_2(t)$ can be used to accurately approximate the steady-state delay probability of the licensed users and the throughput rate of the unlicensed users, respectively. Note that fluid models fail to yield probabilistic performance measures in most applications. Similar to Gurvich and Perry (2012), our fluid model actually provides accurate approximations for them.

4.2. The Steady State of the Fluid Model

While the *offered load* of such a system is $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}$, the *effective load* is endogenous as the average time for which an unlicensed user occupies a channel $\frac{1}{\mu}$ defined in (2) depends on the decision $\frac{1}{\mu_t}$. Since $\frac{1}{p}$ is the average number of service sessions needed to fulfill the service requirement of an unlicensed user, the effective service time of an unlicensed user is $\frac{1}{p\mu}$. Thus, the effective load of the system is

$$\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{p\mu},$$

where $p\mu = \frac{\mu_2\mu_s}{\mu_t+\mu_s}$. Note that the effective load is always no less than the offered load and equals the offered load if and only if there is no restriction on the service process of the unlicensed users, i.e., $\frac{1}{\mu_t}=\infty$. The shorter the transmission time in a service session, the more service sessions (and hence sensing) are needed for the unlicensed users to complete their transmissions and the more congested the system is. Depending on the effective load of the system, the steady states of the fluid limits are given in the next theorem whose proof can be found in Appendix A.

THEOREM 1. *There exists a unique solution*[1] *to the fluid model. Moreover, the limiting behavior of the fluid model as* $t\to\infty$ *can be characterized as follows.*

[1] A vector-valued function $x(t)$ is called a solution of the fluid model if it is absolutely continuous on every closed time interval and satisfies equations (4)–(7) almost everywhere.

1. *If* $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{p\mu}>1$, *then* $\lim_{t\to\infty} x(t)=\left(\frac{\lambda_1}{\mu_1},0,1-\frac{\lambda_1}{\mu_1},\frac{1-\phi}{\theta\phi}\left[\lambda_2-p\mu\big(1-\frac{\lambda_1}{\mu_1}\big)\right]\right)$, $TH_2=p\mu\big(1-\frac{\lambda_1}{\mu_1}\big)$, *and* (β,α) *is the unique solution to*

$$\alpha=\frac{\lambda_1}{\lambda_1+\mu\big(1-\frac{\lambda_1}{\mu_1}\big)}\beta, \tag{10}$$

$$\gamma=\beta+(1-\beta)\frac{(1-p)\alpha}{p+(1-p)\alpha}, \tag{11}$$

$$\lambda_2\mathbb{E}[\gamma^K]=\lambda_2-p\mu\big(1-\frac{\lambda_1}{\mu_1}\big), \tag{12}$$

where $K\geq 1$ *follows a geometric distribution with parameter* ϕ.

2. *If* $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{p\mu}\leq 1$, *then* $\lim_{t\to\infty} x(t)=\left(\frac{\lambda_1}{\mu_1},0,\frac{\lambda_2}{p\mu},0\right)$, $TH_2=\lambda_2$, *and* $\alpha=\beta=0$.

We first describe the intuition behind the delay probability β in Equations (10)–(12) before discussing the steady-state behavior in more detail in the next section. Equation (10) is obtained by plugging $\lim_{t\to\infty} x(t)$ into (9). Note that $1-\beta$ is also the probability that an unlicensed user will be served upon arrival or after each sensing activity while waiting in the orbit queue, and $\frac{(1-p)\alpha}{p+(1-p)\alpha}$ is the probability that an unlicensed user in service will be interrupted. Thus, γ in (11) is the probability that an unlicensed user will experience blockage or interruption and hence needs to decide whether or not to abandon the system at least once. Since $K\geq 1$ represents the number of times an unlicensed user needs to decide whether to abandon the system, $\mathbb{E}[\gamma^K]$ is the probability that an unlicensed user will abandon the system. So the left-hand side of (12) can be understood as the abandonment rate of the unlicensed users, while the right-hand side is also the abandonment rate but calculated by subtracting the rate $p\mu\left(1-\frac{\lambda_1}{\mu_1}\right)$ at which unlicensed users complete their service from the total arrival rate λ_2. Given that $\mathbb{E}[\gamma^K]=\frac{\gamma\phi}{1-\gamma(1-\phi)}$, we actually have a closed-form expression (see (29) in Appendix A) for the delay probability β from solving (10)–(12).

Table 1 also presents a comparison between the simulated delay probability and throughput rate and the approximation based on the fluid model. As one can see, the fluid approximation works well, especially when n is large, which is the case in our application. Furthermore, our simulation also reveals that the average queue length of the licensed users is indeed quite short (vanishes asymptotically). For instance, the 0.95 confidence interval of the queue length of the licensed users is $0.0172\pm.0001$ with deterministic times and $0.0211\pm.0001$ with exponential times when $n=4000$ and $\frac{1}{\mu_t}=0.6$.

From Theorem 1, we can see that the system performance is insensitive to θ, the frequency at which the unlicensed users sense for an available channel while waiting in the orbit queue. This is because, although θ affects the transient of the differential equations (4)–(7), it influences the steady state through the total sensing speed θq_2 (i.e., when the derivatives of the left hand side equal 0). As θ increases, the unlicensed users are allowed to sense channel availability more

frequently and hence abandon the system sooner, which lowers the number of waiting unlicensed users q_2. It turns out that, under such a mechanism, the total sensing speed remains constant as θ varies. The insensitivity of θ on the system performance is further confirmed by a simulation study in Appendix C. Thus, the decision on the sensing frequency should be based on technological (e.g., power consumption as sensing consumes power) rather than operational concerns.

When the system is effectively under or critically loaded, in which case the offered load is $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq 1$, all users will be served without delay in the steady state and no restriction needs to be imposed on the unlicensed users. When the system is effectively overloaded, in which case the offered load may or may not be above 1, only $p\mu\left(1 - \frac{\lambda_1}{\mu_1}\right)$ of the unlicensed users will finish service per unit time and the unlicensed users will experience interference with a positive probability.

Theorem 1 also reveals some interesting steady-state behavior that differs from that of the fluid models in most applications such as call centers.

1. It is well understood in the queueing literature that, if a system is critically loaded, there is a positive probability that delay will occur, even with an extra capacity of $O(\sqrt{\lambda^n})$ in most non-preemptive models in applications such as call centers (see Halfin and Whitt 1981). In our application, the requirement for unlicensed users to sense channel availability while waiting in the orbit queue guarantees the availability of idle channels for all licensed users upon arrival even when the system is critically loaded, leading to a zero delay probability for licensed users asymptotically. We note a similar result in Pang and Perry (2015) that, by controlling when outbound calls can be made, reserving a logarithmic order number of servers in a call center can achieve a zero delay probability for inbound calls asymptotically when the system is critically loaded.
2. It is also well understood that, when a system is overloaded, customers will experience delay almost surely in most call center applications because all servers are busy all the time (see Whitt 2006). In our application, however, an arriving licensed user still has a chance to enter service upon arrival even when there is a large number of unlicensed users in the orbit queue as it takes time for them to sense channel availability. Hence, the delay probability of the licensed users, which is endogenously determined by the load through (10)–(12), is strictly less than 1. Even if a licensed user is delayed upon arrival, his waiting time is in the order of $O\left(\frac{1}{\lambda_1^n}\right)$, which is relatively short but may still be significant in data transmission.

In essence, the restriction that the unlicensed users are not allowed to sense channel availability constantly makes the system operate more like a preemptive one for the licensed users than a non-preemptive one.

4.3. Sensitivity of the System Performance

By Theorem 1, sensing frequency does not affect the system performance. Thus, we will focus on the impact of the length of transmission time $\frac{1}{\mu_t}$ (or equivalently the length of the service session) on the throughput of the unlicensed users TH_2 and the delay probability of the licensed users β.

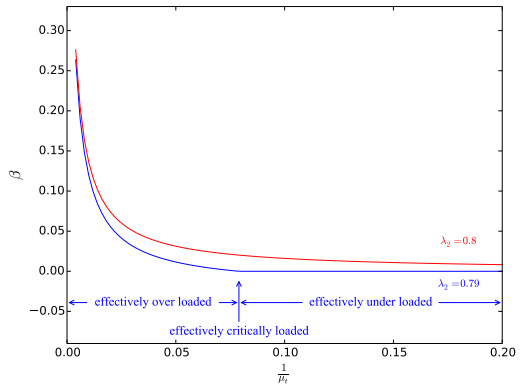
COROLLARY 1. *Throughput TH_2 is always increasing in the transmission time $\frac{1}{\mu_t}$, i.e., allowing the unlicensed users longer service sessions will increase the system throughput rate. β is quasi-convex in $\frac{1}{\mu_t}$. More specifically,*

- *If $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\leq 1$ or $\frac{1}{\mu_s}\geq \frac{1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)}{1+\frac{\mu_2}{\lambda_1}\left(1-\frac{\lambda_1}{\mu_1}\right)}\frac{1}{\mu_2}$, then β decreases in $\frac{1}{\mu_t}$ (see Figure 2(a)–(b)).*
- *Otherwise, there exists a threshold $\frac{1}{\hat{\mu}_t}<\infty$ such that β decreases in $\frac{1}{\mu_t}$ when $\frac{1}{\mu_t}\leq\frac{1}{\hat{\mu}_t}$ and increases in $\frac{1}{\mu_t}$ when $\frac{1}{\mu_t}>\frac{1}{\hat{\mu}_t}$ (see Figure 2(c)–(d)).*

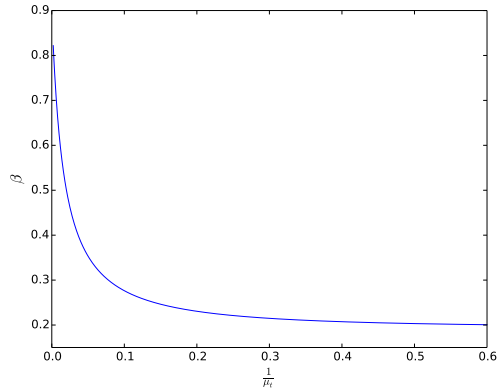
Figure 2 illustrates the delay probability as a function of the transmission time for various λ_1, λ_2 and $\frac{1}{\mu_s}$ when $\frac{1}{\mu_1}=\frac{1}{\mu_2}=1$, $\theta=0.4$ and $\phi=0.5$. Note that the purpose of restricting the amount of time the unlicensed users can occupy a channel is to limit the interference of the unlicensed users to the service of the licensed users. Thus, intuitively, shorter service sessions should always lead to a lower delay probability. The corollary reveals that this is true only if $\frac{1}{\hat{\mu}_t}=0$ which happens when the workload from both types of users are high enough and the sensing time is moderate (Figure 2(d)). When the system is overloaded and sensing is not too time consuming, imposing too short service sessions will only increase the effective load and hence the delay probability, while imposing relatively longer service sessions will increase the delay probability as expected (Figure 2(c)). When the system is under or critically loaded, shorter service sessions will either have no impact on the delay probability or turn the system into an effectively overloaded one, increasing the delay probability (Figure 2(a)). When the system is overloaded and sensing takes a long time, it only makes sense to allow an unlicensed user to transmit for a significant amount of time in order to lower the delay probability (Figure 2(b)).

4.4. Optimal Sharing Decisions in the Steady State

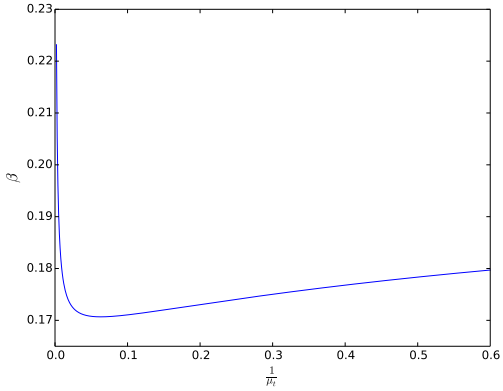
In this section, we investigate whether a given band of spectrum should be shared with unlicensed users and the transmission time $\frac{1}{\mu_t}$ that maximizes the throughput rate of the unlicensed users while keeping the delay probability of the licensed users below a certain level, η. Note that θ does not affect the system performance by Theorem 1 and the transmission time is the only decision. Furthermore, maximizing the throughput rate of the unlicensed users is equivalent to maximizing the throughput rate of the system since the throughput rate of the licensed users is a constant.



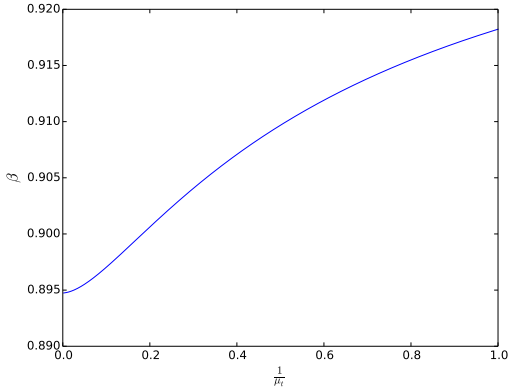
(a) Under ($\lambda_1 = 0.2$, $\lambda_2 = 0.79$) and critically ($\lambda_1 = 0.2$, $\lambda_2 = 0.8$) loaded, $\frac{1}{\mu_s} = 0.001$



(b) Overloaded ($\lambda_1 = 0.2$, $\lambda_2 = 0.9$) and $\frac{1}{\mu_s} = 0.01 > 0.022$



(c) Overloaded ($\lambda_1 = 0.2$, $\lambda_2 = 0.9$) and $\frac{1}{\mu_s} = 0.0001 < 0.022$



(d) Overloaded ($\lambda_1 = 0.8$, $\lambda_2 = 1.7$) and $\frac{1}{\mu_s} = 0.25 \in [0.23, 0.4]$

Figure 2 The delay probability as a function of the transmission time

4.4.1. Whether and How to Share When the system is under or critically loaded, the system may also be effectively overloaded if one allocates shorter service sessions to the unlicensed users. However, by Theorem 1, even if the unlicensed users are allowed to transmit for as long as they need, the delay probability converges to zero and all users are able to complete their transmission without delay as $n \to \infty$. Thus, we do not need to restrict the service session of the unlicensed users when n is large enough.

When the system is overloaded, it is also effectively overloaded regardless of the length of the allocated service session. By Theorem 1, $TH_2 = p\mu\left(1 - \frac{\lambda_1}{\mu_1}\right)$ and β is the solution to (12). Thus, the optimization problem can be written as

$$\max_{\mu_t \geq 0} \quad p\mu \tag{13}$$

$$\begin{aligned} \text{s.t.} \quad & \beta \le \eta, \\ & \mu = \frac{(\mu_2+\mu_t)\mu_s}{\mu_t+\mu_s}, \\ & p = \frac{\mu_2}{\mu_2+\mu_t}. \end{aligned}$$

Since the objective function is increasing in $\frac{1}{\mu_t}$, the optimization problem reduces to one of finding the largest $\frac{1}{\mu_t}$ that satisfies the delay constraint. When η is so small that the feasible region is empty, no unlicensed users should be allowed in the system. Once η is large enough to make the feasible region non-empty, unlicensed users will be allowed in the system. As η increases, the optimal transmission time $\frac{1}{\mu_t^*}$ increases. The optimal transmission time $\frac{1}{\mu_t^*} = \infty$–that is, the unlicensed users are allowed to complete their transmission once they start occupying a channel–if η is larger than the point such that the feasible region becomes unbounded.

Figure 3 demonstrates the optimal spectrum sharing decision as a function of η and λ_2 when $\lambda_1 = 0.2$, $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 1$, $\frac{1}{\mu_s} = 0.001$, $\theta = 0.4$ and $\phi = 0.5$. The upper curve specifies the arrival rate above which the unlicensed users should not be allowed to share the spectrum and the lower one is the threshold below which there is no need to restrict the service session of the unlicensed users, i.e., $\frac{1}{\mu_t^*} = \infty$.

Since our analysis only holds in an asymptotic sense (as the number of channels n becomes large), there is still a non-negligible delay probability when the system is under or critically loaded and n is not sufficiently large. For the same example in Figure 3 with $\frac{\lambda_1^n}{n} = 0.2$, Figure 4 demonstrates the optimal sharing decisions, obtained through simulation, as a function of η and $\frac{\lambda_2^n}{n}$ for $n = 100, 200, 500, 1000$, in which case the system is under or critically loaded when $\frac{\lambda_2^n}{n} \le 0.8$. As one can see, the structure of the optimal sharing decision remains the same, and as n increases, sharing is more likely to occur and the unlicensed users should be allowed longer service sessions.

4.4.2. Sensitivity of the Optimal Decision The optimal decision $\frac{1}{\mu_t^*}$ and the throughput rate of the unlicensed users TH_2^* depend on the system parameters in the following way.

PROPOSITION 1. *The optimal $\frac{1}{\mu_t^*}$ decreases, i.e., the unlicensed users are allowed a shorter transmission time, as*

(1) λ_1 increases while keeping $\lambda_1 + \lambda_2$ constant when $\mu_1 = \mu_2$;

(2) λ_1 and μ_1 decrease while keeping $\frac{\lambda_1}{\mu_1}$ constant; and

(3) λ_2 and μ_2 increase while keeping $\frac{\lambda_2}{\mu_2}$ constant.

Furthermore, the optimal throughput TH_2^* will decrease under (1) and (2).

Note that under all the scenarios, the total offered load $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$ is kept constant. Proposition 1 states that shorter service sessions should be allocated to the unlicensed users (1) as the proportion

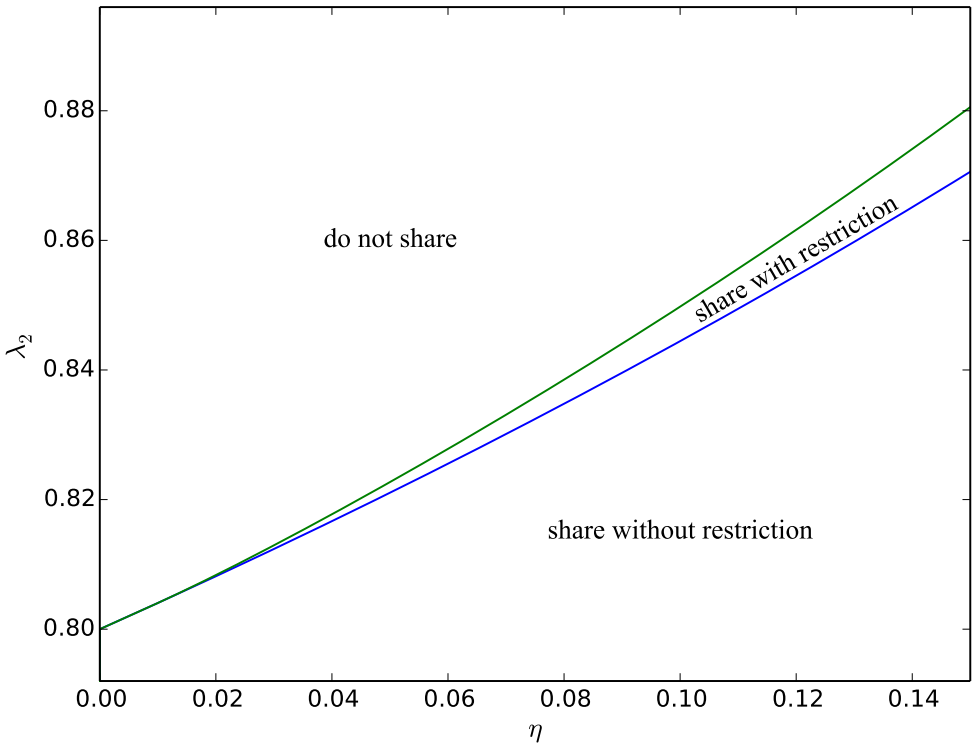
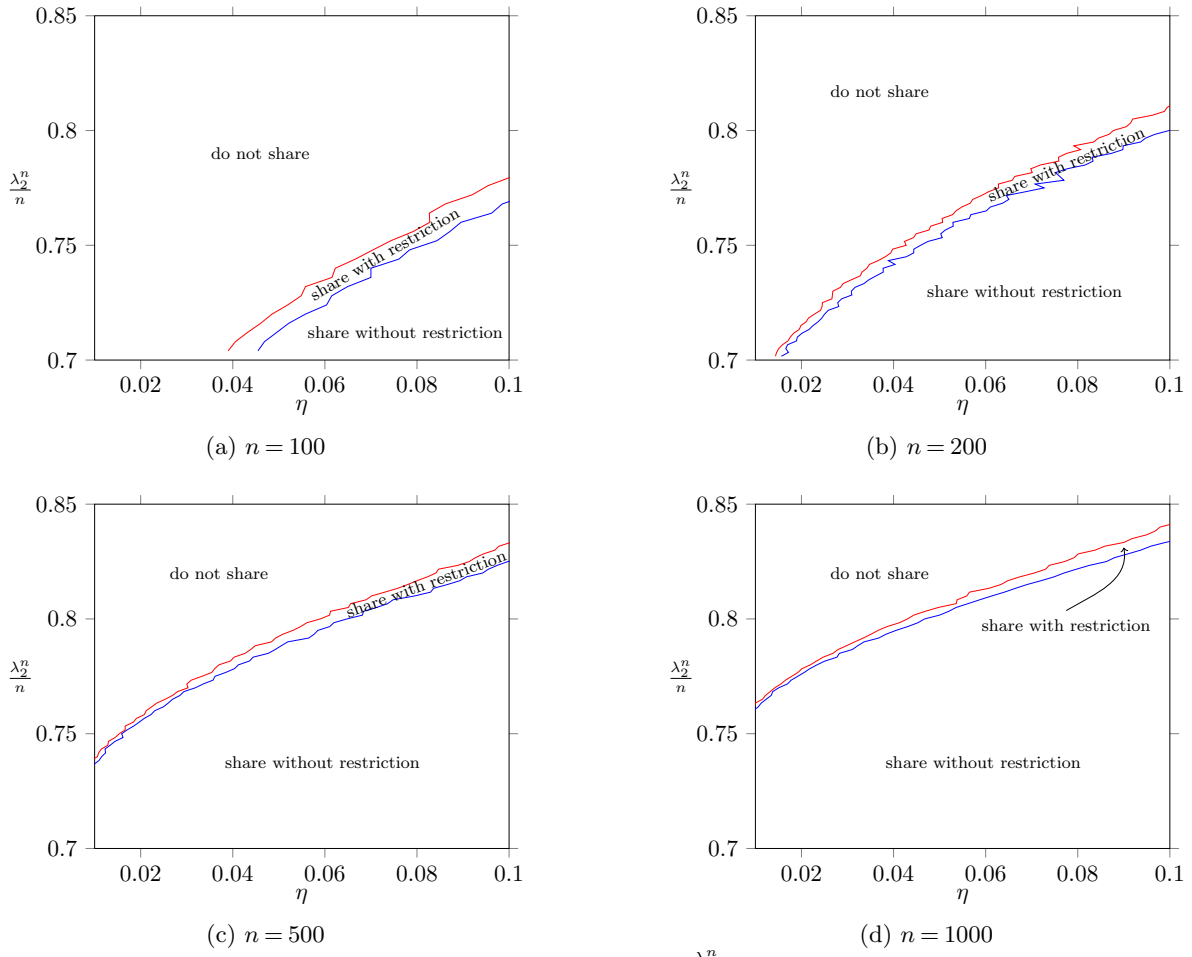


Figure 3 The optimal sharing decision as a function of η and λ_2 for an overloaded system

of licensed users increases when all users have identical service requirements, (2) if there are fewer licensed users but with longer service times; and (3) if there are more unlicensed users but with shorter service times. While (1) and (3) are more intuitive, (2) holds because the delay probability only depends on both $\frac{\lambda_1}{\mu_1}$ and λ_1. A delay incident of a licensed user is counted as one regardless of his service requirement. With fewer licensed users, each delay contributes more to the delay probability and it is easy to show that shorter service sessions should be imposed on the unlicensed users.

As a result, the optimal throughput rate $p\mu^* \left(1 - \frac{\lambda_1}{\mu_1}\right) = \frac{\mu_2 \mu_s}{\mu_t^* + \mu_s} \left(1 - \frac{\lambda_1}{\mu_1}\right)$ decreases under scenarios (1) and (2) as expected. These suggest that spectrum sharing is beneficial to systems with a smaller proportion of licensed users or a large number of licensed users with shorter service times. Under scenario (3), although shorter service sessions have a negative impact on the throughput rate due to more sensing activities, the increase in the number of unlicensed users with shorter service times has a positive impact. Thus, the impact on throughput rate is not monotone.



(a) $n = 100$ (b) $n = 200$

(c) $n = 500$ (d) $n = 1000$

Figure 4 The optimal sharing decisions as a function of η and $\frac{\lambda_2^n}{n}$

5. Justifications for the Fluid Approximation

In this section, we demonstrate in Theorem 2 that the scaled process $\bar{X}^n(t)$ converges to the fluid model $x(t)$ in Section 4. Since the proof of the theorem is quite involved, we describe the main ideas of the proof through the construction of the fluid model, especially the instantaneous delay probability of the licensed users $\beta(t)$ and the instantaneous interruption probability of the unlicensed users $\alpha(t)$. The complete proof can be found in Appendix B. For any $T > 0$, let $\mathcal{D}([0,T],\mathbb{R}^4)$ be the space of all right-continuous $\mathbb{R}^4$ valued functions on $[0,T]$ with left limits, endowed with the Skorohod J_1 topology. Let "$\Rightarrow$" denote convergence in distribution for random objects in $\mathbb{R}^4$ equipped with Euclidian topology or $\mathcal{D}([0,T],\mathbb{R}^4)$ with Skorohod J_1 topology.

THEOREM 2 **(Fluid Approximation)**. *Under the asymptotic regime, if* $\bar{X}^n(0) \Rightarrow x(0)$ *as* $n \to \infty$, *then* $\bar{X}^n(t) \Rightarrow x(t)$ *in* $\mathcal{D}([0,T],\mathbb{R}^4)$, *where* $x(t)$ *is the fluid model specified in Definition 2.*

Need for both scaled and unscaled processes If we let $\Lambda_i^n(t)$ denote the Poisson process with the rate λ_i^n, then $\Lambda_1^n(t+\delta) - \Lambda_1^n(t)$ is the total number of licensed users arriving in a small

interval $[t, t+\delta]$, among which $\int_t^{t+\delta} \mathbf{1}_{\{I^n(s-)=0\}} d\Lambda_1^n(s)$ will find no idle channels upon arrival and have to wait. Thus, the delay probability of the licensed users during this small time interval is $\mathbb{E}\left[\frac{\int_t^{t+\delta} \mathbf{1}_{\{I^n(s-)=0\}} d\Lambda_1^n(s)}{\Lambda_1^n(t+\delta)-\Lambda_1^n(t)}\right]$. That is, the delay probability depends on the information about the unscaled process $I^n(t) \geq 0$ as it determines whether an unlicensed user in service should vacate a channel at the end of a service session. Likewise, we need to keep track of the unscaled process of the queue length of the licensed users $Q_1^n(t)$ and obtain the probability of an unlicensed user in service being interrupted in $[t, t+\delta]$. However, $I^n(t) \geq 0$ vanishes in the asymptotic regime along with the process $Q_1^n(t) \geq 0$ as in most systems with multiple classes and we need to keep track of both the scaled and unscaled processes in order to obtain the system dynamics and asymptotic system performance.

The system dynamics using the averaging principle To obtain the system dynamics, we need to apply the averaging principle by first expressing the probabilities in $[t, t+\delta]$ as a time average using PASTA (Poisson arrivals see time average). For instance, the fraction of time for which there is no idle channel in the system is

$$\frac{1}{\delta}\int_t^{t+\delta} \mathbf{1}_{\{I^n(s-)=0\}} ds = \frac{1}{n\delta}\int_t^{t+n\delta} \mathbf{1}_{\{I^n(t+\frac{s-}{n})=0\}} ds. \tag{14}$$

Let

$$m^n(t) = Q_1^n(t) - I^n(t). \tag{15}$$

We study the system dynamics for the unscaled process $m^n\left(t+\frac{s}{n}\right)$ for $0 \leq s \leq n\delta$. Note that the process $m^n\left(t+\frac{\cdot}{n}\right)$ oscillates around zero in the order of 1. When $m^n\left(t+\frac{s}{n}\right) < 0$ (there are idle channels and no licensed users waiting by (1)), the process increases by 1 when there is a new arrival at the rate $\bar{\lambda}_1^n + \bar{\lambda}_2^n$ or one of the unlicensed users in the orbit queue enters service after sensing the system at the rate $\theta \bar{Q}_2^n(t+\frac{s}{n})$. The process decreases by 1 when a user (licensed or unlicensed) completes service at the rate $\mu_1 \bar{Z}_1^n(t+\frac{s}{n}) + p\mu \bar{Z}_2^n(t+\frac{s}{n})$. When $m^n\left(t+\frac{s}{n}\right) > 0$ (there are licensed users waiting and no idle channels), the process increases by 1 at the rate $\bar{\lambda}_1^n$ and decreases at the rate $\mu_1 \bar{Z}_1^n(t+\frac{s}{n}) + \mu \bar{Z}_2^n(t+\frac{s}{n})$. We refer readers to Appendix B.1 for the detailed system dynamics. It is the long-run average behavior of $m^n\left(t+\frac{\cdot}{n}\right)$ that plays the key role in determining the fraction in (14) when n becomes large in the asymptotic regime.

Explanation for $\boldsymbol{\beta(t)}$ and $\boldsymbol{\alpha(t)}$ As one can see, the process $m^n\left(t+\frac{\cdot}{n}\right)$ is not a Markov process since its evolution depends on a higher dimension process than itself. However, if we approximate the above mentioned rates by their fluid counterparts, i.e., $\bar{Z}_i^n(t+\frac{s}{n})$ by $z_i(t)$, $\bar{Q}_2^n(t+\frac{s}{n})$ by $q_2(t)$, and $\bar{\lambda}_i^n$ by λ_i, we have a Markov process as in Figure 5 whose steady-state distribution π_t can be easily obtained. We use $\pi_t(j)$ to approximate the asymptotic proportion of time for which there are j licensed users in the queue when $j > 0$ and there are $-j$ idle channels when $j < 0$. The delay

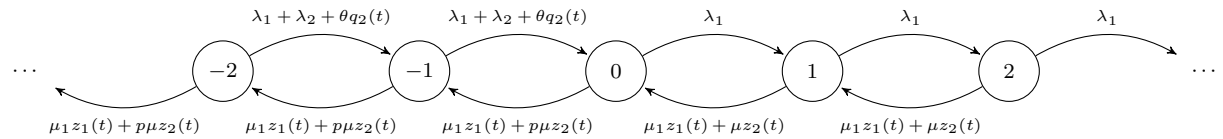


Figure 5 The asymptotic transition diagram of $m^n\left(t+\frac{\cdot}{n}\right)$.

probability of the licensed users in (14) and the interruption probability of the unlicensed users in the asymptotic regime can be approximated by $\sum_{j=0}^{\infty}\pi_t(j):=\beta(t)$ and $\sum_{j=1}^{\infty}\pi_t(j):=\alpha(t)$, respectively.

6. Extensions

In this section, we extend the problem to systems with time-varying arrival rates and propose a diffusion approximation that can lead to better performance in some cases.

6.1. With Time-Varying Arrival Rates

When the arrival rates vary over time, the optimal decision on the transmission time needs to be adjusted dynamically. Suppose that adjustment of the transmission time can be done instantaneously and the initial state $x(0)$ is given. We can extend the fluid model in Definition 2 to allow time-varying arrivals by adding an argument t to λ_i, μ_t, p, and μ to denote their instantaneous values. Following similar arguments in Appendices A and B, we can show that the stochastic processes with time-varying arrival rates converge to the extended fluid model and there exists a unique solution to time-varying differential equations of the fluid model as long as $\lambda_i(t)$'s are bounded and locally Lipschitz continuous. In this case, the instantaneous throughput rate is $p(t)\mu(t)z_2(t)$, the instantaneous delay probability $\beta(t)$ is given by (8) and the optimization problem over a period of time T can then be written as

$$\begin{aligned}\max_{\mu_t(\cdot)}\ & \int_0^T p(t)\mu(t)z_2(t)dt\\ \text{s.t.}\ & \beta(t)\le\eta,\\ & \mu(t)=\frac{[\mu_2+\mu_t(t)]\mu_s}{\mu_t(t)+\mu_s},\\ & p(t)=\frac{\mu_2}{\mu_2+\mu_t(t)}.\end{aligned}$$

Although such a continuous-time dynamic programming problem can be solved numerically using policy iteration, the resulting policy is hard to implement in practice. Thus, we ask whether a periodically adjusted policy will work well. As an example, suppose that the arrival rates change over time as in Figure 6, $\frac{1}{\mu_1}=\frac{1}{\mu_2}=1$ minutes, $\frac{1}{\mu_s}=0.02$ minute, $\theta=0.4$, $\phi=0.5$ and $\eta=0.2$. Figure 6 plots the transmission times adjusted on an hourly basis and the performance over a 10-hour period. As one can see, our heuristic policy performs very well. According to our numerical

experiments, the throughput rate under the hourly adjusted policy is consistently within 0.2% of the optimal throughput rate.

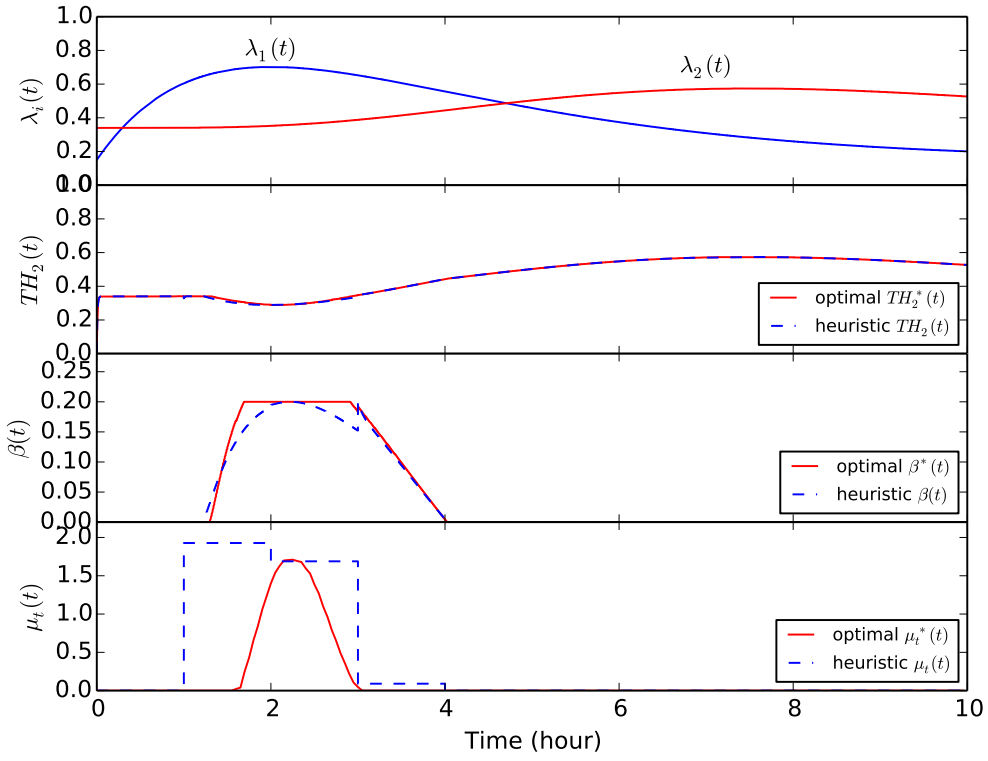


Figure 6 With time-varying arrival rates and periodic adjustments

6.2. A Diffusion Scaling

Although our fluid scaling results in good approximations, it leads to a zero delay probability when the system is under and critically loaded, which is not accurate when n is small. Thus, we ask whether a diffusion scaling may work better for under and critically loaded systems.

Consider the diffusion scaling where the licensed users grow in the order of $O(\sqrt{n})$ and the unlicensed in $O(n)$, i.e.,

$$\begin{aligned}\lambda_1^n &= \tilde{\lambda}_1\sqrt{n},\\ \lambda_2^n &= p\mu n + \tilde{\lambda}_2\sqrt{n},\end{aligned}$$

and $\tilde{Z}_1^n(t) = \frac{Z_1^n(t)}{\sqrt{n}}$, $\tilde{Z}_2^n(t) = \frac{Z_2^n(t)-n}{\sqrt{n}}$ and $\tilde{Q}_i^n(t) = \frac{Q_i^n(t)}{\sqrt{n}}$ are the corresponding diffusion scaled processes. Such a scaling explicitly assumes that there are far more unlicensed users than licensed

users. It can be shown that the diffusion scaled processes converge and there exist coefficients C_1, C_2, C_q, C_β and C_α such that

$$\begin{aligned}
&\mathbb{E}[Z_1^n(\infty)] = C_1\sqrt{n} + o(\sqrt{n}), && \mathbb{E}[Z_2^n(\infty)] = n + C_2\sqrt{n} + o(\sqrt{n}),\\
&\mathbb{E}[Q_1^n(\infty)] = o(1), && \mathbb{E}[Q_2^n(\infty)] = C_q\sqrt{n} + o(\sqrt{n}),\\
&\alpha^n = \frac{C_\alpha}{\sqrt{n}} + o\Big(\frac{1}{\sqrt{n}}\Big), && \beta^n = \frac{C_\beta}{\sqrt{n}} + o\Big(\frac{1}{\sqrt{n}}\Big).
\end{aligned} \tag{16}$$

6.2.1. Estimation of the coefficients First, it is easy to see that $C_1 = \frac{\tilde{\lambda}_1}{\mu_1}$ since the licensed users do not abandon. Since the unlicensed users may abandon, the system is stable in the long run and hence the balance equations are given by letting (4)–(7) to equal to zero and replacing $(\lambda_i, z_i, q_i, \beta, \alpha)$ by $\big(\lambda_i^n, \mathbb{E}[Z_i^n(\infty)], \mathbb{E}[Q_i^n(\infty)], \beta^n, \alpha^n\big)$. Solving the balance equations, we are able to obtain

$$C_\alpha = 0,$$

$$C_\beta = \frac{\theta C_q}{(1-\phi)p\mu}, \tag{17}$$

$$\theta C_q = (1-\phi)\left[\tilde{\lambda}_2 - p\mu C_2 + \theta C_q\right]. \tag{18}$$

It remains to estimate C_2 and C_q. If we are able to derive a closed-form steady state distribution of the limit of the diffusion scaled process, we can obtain the value of these coefficients. Although the four-dimensional diffusion scaled process, $(\tilde{Z}_1^n, \tilde{Q}_1^n, \tilde{Z}_2^n, \tilde{Q}_2^n)$, can be reduced to a three-dimensional process as $\tilde{Q}_1^n$ converges to 0, it has some complicated reflection behavior on the boundary when all channels are busy, i.e., $\tilde{Z}_1^n(t) + \tilde{Z}_2^n(t) = 0$. In general, it is challenging to derive the steady state distribution of a multidimensional diffusion process and closed-form expressions of the coefficients are almost impossible. Thus, we propose a heuristic method to derive closed-form approximations for C_2 and C_q and hence C_β.

We pretend that the licensed users occupy $\frac{\tilde{\lambda}_1}{\mu_1}\sqrt{n}$ channels exclusively and the waiting unlicensed users form a steady source of arrival with the rate $\theta C_q\sqrt{n}$. The unlicensed users are served by the remaining $n - \frac{\tilde{\lambda}_1}{\mu_1}\sqrt{n}$ channels and forms an Erlang-B queue with the arrival rate $\lambda_2^n + \theta C_q\sqrt{n}$ and service rate $p\mu$. In such a network, $\lim_{n\to\infty} \tilde{Z}_2^n(t) = \tilde{z}_2(t)$ is a reflected Brownian motion with an infinitesimal mean $-p\mu\left(\tilde{z}_2 - \frac{\tilde{\lambda}_2 + \theta C_q}{p\mu}\right)$ and infinitesimal variance $2p\mu$. Therefore, $\lim_{t\to\infty}\lim_{n\to\infty} \tilde{Z}_2^n(t)$ follows a truncated normal distribution with mean $\frac{\tilde{\lambda}_2 + \theta C_q}{p\mu}$ and variance 1 on $\left(-\infty, -\frac{\tilde{\lambda}_1}{\mu_1}\right)$ and hence

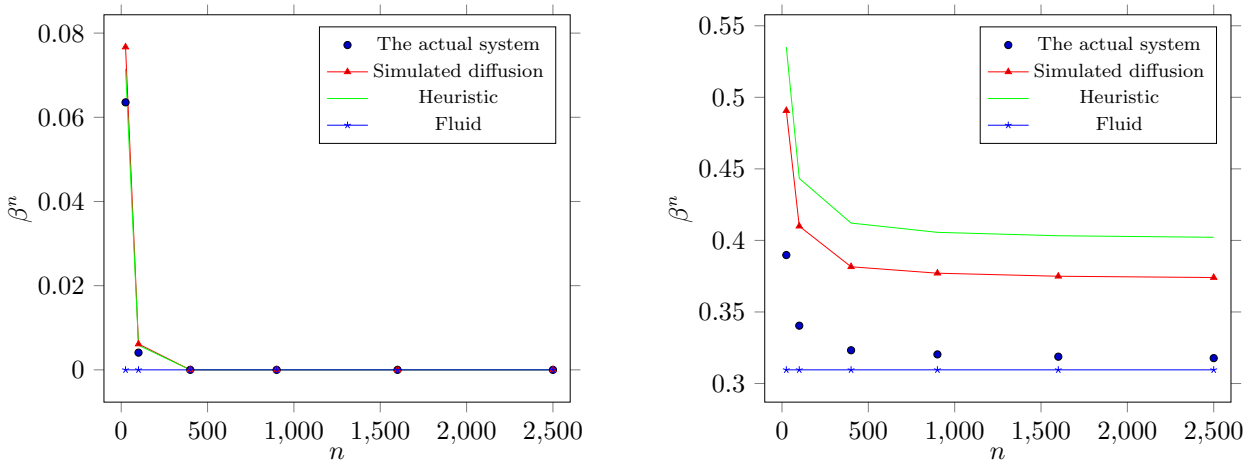
$$C_2 = \mathbb{E}\left[\lim_{t\to\infty}\lim_{n\to\infty} \tilde{Z}_2^n(t)\right] = \frac{\tilde{\lambda}_2 + \theta C_q}{p\mu} - \frac{\Phi'\left(-\frac{\tilde{\lambda}_1}{\mu_1} - \frac{\tilde{\lambda}_2 + \theta C_q}{p\mu}\right)}{\Phi\left(-\frac{\tilde{\lambda}_1}{\mu_1} - \frac{\tilde{\lambda}_2 + \theta C_q}{p\mu}\right)}, \tag{19}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. By (17)–(19), we can obtain C_2, C_q, and

$$C_\beta = \frac{\Phi'\left(-\frac{\tilde{\lambda}_1}{\mu_1} - \frac{\tilde{\lambda}_2}{p\mu} - (1-\phi)C_\beta\right)}{\Phi\left(-\frac{\tilde{\lambda}_1}{\mu_1} - \frac{\tilde{\lambda}_2}{p\mu} - (1-\phi)C_\beta\right)}. \tag{20}$$

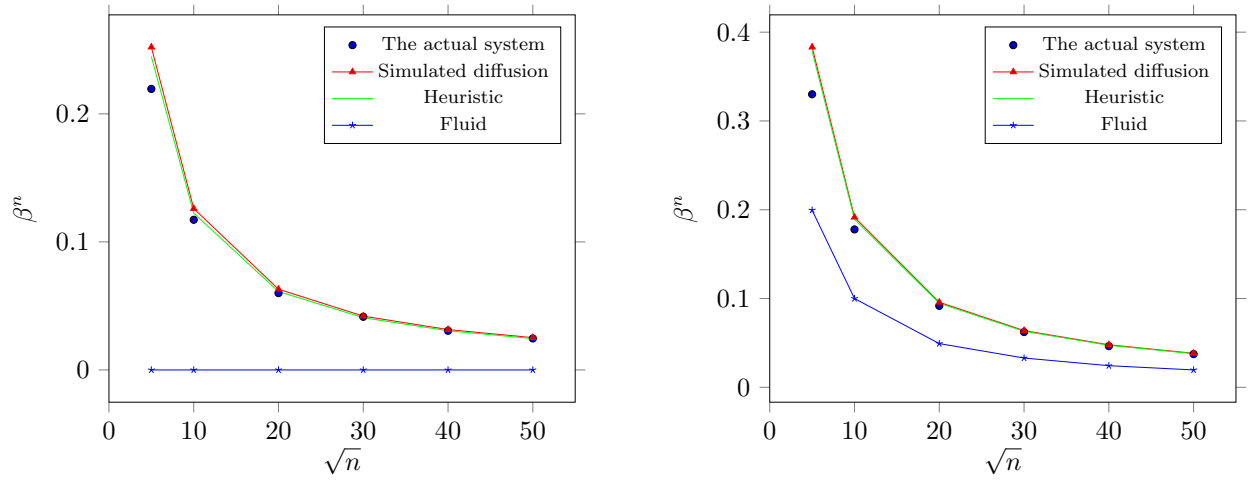
By (16), the delay probability of the nth system can be approximated by $\frac{C_\beta}{\sqrt{n}}$, the throughput rate can be approximated by $p\mu\mathbb{E}[Z_2^n(\infty)/n] = p\mu(1 + C_2/\sqrt{n})$. Thus, the accuracy of the estimation of the system performance is reflected by the coefficients C_β and C_2.

6.2.2. Accuracy of the heuristic To show how the above heuristic approximates the delay probability and throughput rate of the diffusion scaled processes as well as the actual system, we conduct a numerical experiment. We simulate large systems to obtain the diffusion limits and compare them with the heuristic ones. For $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 1$, $\frac{1}{\mu_s} = 0.0001$, $\theta = 0.4$, $\phi = 0.5$ and $\frac{1}{\mu_t} = \infty$, Figures 7 and 8 compare the diffusion scaled delay probabilities with the heuristic ones. As expected, our heuristic mimics the performance of the simulated diffusion limits well, especially when the systems are under or critically loaded. Note that, in the network, all the channels are pooled to serve both licensed and unlicensed users, while the heuristic estimates the coefficients pretending that the licensed users occupy a fixed number of channels. When the system is under or critically loaded, the heuristic works well as long as the channels are well allocated to the two types of users, as shown in Figure 7(a) and Figure 8. The impact of decoupling the channels is higher when the system is over loaded, as shown in Figure 7(b).



(a) under loaded: $\lambda_1 = 0.05$ and $\lambda_2 = 0.75$ (b) over loaded: $\lambda_1 = 0.2$ and $\lambda_2 = 1.0$

Figure 7 Under and over load: the delay probabilities as a function of n when $\lambda_1^n = \lambda_1 n$ and $\lambda_2^n = \lambda_2 n$



(a) $\tilde{\lambda}_1 = 0.5$ and $\tilde{\lambda}_2 = -0.5$ (b) $\tilde{\lambda}_1 = 1.0$ and $\tilde{\lambda}_2 = -0.5$

Figure 8 Critical load: the delay probabilities as a function of $\sqrt{n}$ when $\lambda_1^n = \tilde{\lambda}_1\sqrt{n}$ and $\lambda_2^n = p\mu n + \tilde{\lambda}_2\sqrt{n}$

6.2.3. Comparison to the Fluid Approximation Figures 7 and 8 also plot the delay probabilities under the fluid scaled processes and of the actual systems. As one can see, the fluid approximation always underestimates β, while the diffusion approximation always overestimates it. Furthermore, the fluid approximation outperforms the diffusion approximation when the system is over loaded, and the converse is true when the system is under or critically loaded. Thus, neither method is uniformly more accurate than the other. However, further comparisons reveal the following.

1. The fluid scaling leads to analytical closed-form approximations, while analysis under the diffusion scaling involves solving the steady states of multi-dimensional diffusion processes, which is known to be an open question in most cases.
2. The closed-form approximations under the fluid scaling reveal important operational insights (in Section 4) that are not obvious under the diffusion approximation.
3. The diffusion approximation may not be feasible when a system is overloaded or the number of licensed users is comparable to or more than that of unlicensed ones, while the fluid approximation can be used under any load level with any ratio between the licensed and unlicensed users. Such a drawback may further limit the diffusion approximation to be adopted to systems with time varying or random arrivals.

7. Conclusions and Future Research

Opportunistic access of licensed spectrum by unlicensed users is widely considered as a way to alleviate artificial scarcity of radio spectrum by increasing the spectrum utilization. However, it may reduce the service quality for licensed users due to potential interference from unlicensed

users. While much research on spectrum sharing has been conducted by researchers in electrical engineering with the main focus on the technological issues, the operational aspects have not been adequately addressed through analytical work.

In this paper, we model a shared network consisting of both licensed users and unlicensed users as a multi-class, many-server queueing system. The distinctive features of our model are that the service requirement of an unlicensed user can be fulfilled even after multiple interruptions and the unlicensed users waiting in the queue are required to sense channel availability periodically while waiting. These features complicate system dynamics and lead to quite different insights from those derived from most service systems. We show that the sensing frequency of the unlicensed users waiting in the queue does not affect system performance from the operational perspective and its decision should be based on technological concerns. When the system is under or critically loaded, there is no need to restrict the service session of unlicensed users. Otherwise, limiting the transmission of the unlicensed users is necessary only when the system load is above a threshold. Thus, it is possible to improve spectrum utilization while guaranteeing a very high service level, as expected by licensed users in practice, and spectrum sharing can potentially be a socially optimal solution to alleviating spectrum scarcity.

Spectrum sharing, if feasible, is especially beneficial for systems with a smaller portion or a large number of licensed users with shorter service times. Our study sheds light on the implementation of spectrum sharing and opens the door for new applications of existing queueing theory in wireless communication networks, which may lead to the development of new methodologies.

Our study also provides some rich research opportunities. For instance, the arrival rates of the users may be uncertain in practice. Our preliminary result shows that higher variance will always hurt system performance if the system is expected to be under or critically loaded. However, if the system is expected to be overloaded, it seems that increasing the variability up to a certain level will actually improve the throughput rate. Thus, research needs to be done to investigate the impact of uncertain arrival rates on system performance.

In reality, users' behavior in data transmission can be more complicated than those in the network in Figure 1. For instance, unlicensed users who have to abandon the system earlier may reenter the system later, while licensed users may abandon the system if no idle channel is available upon arrival. Also, sensing may not be perfect, e.g., a false alarm can occur, in which case a spectrum opportunity is overlooked by an unlicensed user. It will be interesting to incorporate these elements into the model and examine how they change the system performance and operational decisions.

The insights revealed in the above research may also pave the way for studying other important business issues in wireless communications such as contract design and pricing in shared networks. For instance, how should a spectrum owner set the prices and decide the service quality to both

licensed and unlicensed users in a shared network? Should unlicensed users be charged a fixed and/or usage based fee? Since unlicensed users may belong to different service providers, should a spectrum owner run an auction to select the service providers and settle the prices?

Acknowledgments

The authors acknowledge the advice of Professor Qian Zhang from the Department of Computer Science at Hong Kong University of Science and Technology, and encouragement of the area editor, the associate editor, and the two anonymous reviewers. This work was supported by the Hong Kong Research Grants Council under Grants 622713, 16500615, and 16501015, and Hong Kong Polytechnic University under Grants 252019/16E and 1-ZE5G.

References

Akyildiz, I. F., W.-Y. Lee, M. C. Vuran, and S. Mohanty (2006). Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer Networks 50*(13), 2127 – 2159.

Artalejo, J. and A. Gómez-Corral (2008). *Retrial Queueing Systems: A Computational Approach*. Springer.

Atar, R., A. Mandelbaum, and M. I. Reiman (2004). Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. *Ann. Appl. Probab. 14*(3), 1084–1134.

Bansal, G., J. Hossain, and V. Bhargava (2008). Optimal and suboptimal power allocation schemes for ofdm-based cognitive radio systems. *IEEE Transactions on Wireless Communications 7*(11), 4710–4718.

Biglieri, E., A. Goldsmith, L. Greenstein, N. Mandayam, and H. Poor (2013). *Principles of Cognitive Radio*. Cambridge University Press.

Brosh, I. (1969). Preemptive priority assignment in multichannel systems. *Operations Research 17*(3), 526–535.

Buzen, J. P. and A. B. Bondi (1983). The response times of priority classes under preemptive resume in M/M/m queues. *Operations Research 31*(3), 456–465.

Capar, F., I. Martoyo, T. Weiss, and F. Jondral (2002). Comparison of bandwidth utilization for controlled and uncontrolled channel assignment in a spectrum pooling system. In *IEEE 55th Vehicular Technology Conference*, Volume 3, pp. 1069–1073.

de Véricourt, F. and O. B. Jennings (2008). Dimensioning large-scale membership services. *Oper. Res. 56*(1), 173–187.

de Véricourt, F. and Y.-P. Zhou (2005). Managing response time in a call-routing problem with service failure. *Oper. Res. 53*(6), 968–981.

Devroye, N., P. Mitran, and V. Tarokh (2006). Achievable rates in cognitive radio channels. *IEEE Transactions on Information Theory 52*(5), 1813–1827.

El Gamal, A., J. Mammen, B. Prabhakar, and D. Shah (2006). Optimal throughput-delay scaling in wireless networks - part i: the fluid model. *IEEE Transactions on Information Theory 52*(6), 2568–2592.

Falin, G. I. and J. G. C. Templeton (1997). *Retrial queues*, Volume 75. CRC Press.

Filippov, A. F. (1988). *Differential Equations with Discontinuous Righthand Sides*, Volume 18 of *Mathematics and Its Applications*. Springer Netherlands.

Goldsmith, A., S. Jafar, I. Maric, and S. Srinivasa (2009). Breaking spectrum gridlock with cognitive radios: An information theoretic perspective. *Proceedings of the IEEE 97*(5), 894–914.

Gong, S., P. Wang, and L. Duan (2015, June). Distributed power control with robust protection for PUs in cognitive radio networks. *IEEE Transactions on Wireless Communications 14*(6), 3247–3258.

Gupta, P. and P. Kumar (2000). The capacity of wireless networks. *IEEE Transactions on Information Theory 46*(2), 388–404.

Gurvich, I. and O. Perry (2012). Overflow networks: Approximations and implications to call center outsourcing. *Operations Research 60*(4), 996–1009.

Gurvich, I. and W. Whitt (2009). Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res. 34*(2), 363–396.

Halfin, S. and W. Whitt (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res. 29*(3), 567–588.

Hossain, E., D. Niyato, and Z. Han (2009). *Dynamic Spectrum Access and Management in Cognitive Radio Networks*. Cambridge: Cambridge University Press.

Huang, S., X. Liu, and Z. Ding (2008). Opportunistic spectrum access in cognitive radio networks. In *Proc. of The 27th Conf. on Comp. Commun. (IEEE INFOCOM 2008)*, pp. 2101–2109.

Hunt, P. J. and T. G. Kurtz (1994). Large loss networks. *Stochastic Processes and their Applications 53*(2), 363–378.

Jaiswal, N. K. (1968). *Priority Queues*. Mathematics in Science and Engineering. New York and London: Elsevier Science.

Jia, J., Q. Zhang, and X. Shen (2008, Jan). HC-MAC: A hardware-constrained cognitive mac for efficient spectrum management. *IEEE Journal on Selected Areas in Communications 26*(1), 106–117.

Kurtz, T. G. (1992). Averaging for martingale problems and stochastic approximation. In *Applied Stochastic Analysis*, Volume 177 of *Lecture Notes in Control and Information Sciences*, pp. 186–209. Springer, Berlin.

Liu, K. R. and B. Wang (2010). *Cognitive radio networking and security: A game-theoretic view*. Cambridge University Press.

Luo, J. and J. Zhang (2013). Staffing and control of instant messaging contact centers. *Oper. Res. 61*(2), 328–343.

Maglaras, C. and A. Zeevi (2004). Diffusion approximations for a multiclass markovian service system with "guaranteed" and "best-effort" service levels. *Mathematics of Operations Research 29*(4), 786–813.

Maglaras, C. and A. Zeevi (2005). Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research 53*(2), 242–262.

Mandelbaum, A., W. Massey, M. Reiman, A. Stolyar, and B. Rider (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems 21*(2-4), 149–171.

Mishra, S., A. Sahai, and R. Brodersen (2006). Cooperative sensing among cognitive radios. In *Proc. of IEEE International Conference on Communications*, Volume 4, pp. 1658–1663.

Mitola, J. and J. Maguire, G.Q. (1999). Cognitive radio: making software radios more personal. *Personal Communications, IEEE 6*(4), 13–18.

Pang, G. and O. Perry (2015). A logarithmic safety staffing rule for contact centers with call blending. *Mgt. Sci. 61*(1), 73–91.

Perry, O. and W. Whitt (2011a). A fluid approximation for service systems responding to unexpected overloads. *Oper. Res. 59*(5), 1159–1170.

Perry, O. and W. Whitt (2011b). An ODE for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems 1*, 17–66.

Perry, O. and W. Whitt (2013). A fluid limit for an overloaded X model via an averaging principle. *Math. Oper. Res. 38*(2), 294–349.

Prasad, R., S. Dixit, R. Van Nee, and T. Ojanpera (Eds.) (2010). *Globalization of Mobile and Wireless Communications: Today and in 2020*. Signals and Communication Technology. Springer.

Puhalskii, A. A. and M. I. Reiman (2000). The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. in Appl. Probab. 32*(2), 564–595.

Teschl, G. (2012). *Ordinary Differential Equations and Dynamical Systems*. Graduate studies in mathematics. American Mathematical Society.

van de Beek, J., J. Riihijarvi, A. Achtzehn, and P. Mahonen (2012, Feb). TV White Space in Europe. *IEEE Transactions on Mobile Computing 11*(2), 178–188.

Wang, J., O. Baron, and A. Scheller-Wolf (2015). M/M/c queue with two priority classes. *Operations Research 63*(3), 733–749.

White, H. and L. S. Christie (1958). Queuing with preemptive priorities or with breakdown. *Operations Research 6*(1), 79–95.

Whitt, W. (2002). *Stochastic-process limits*. Springer Series in Operations Research. New York: Springer-Verlag.

Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Oper. Res. 54*(1), 37–54.

Yang, T. and J. Templeton (1987). A survey on retrial queues. *Queueing Syst. 2*(3), 201–233.

Yom-Tov, G. and A. Mandelbaum (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management 16*(2), 283–299.

Zhan, D. and A. R. Ward (2014). Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management 16*(2), 220–237.

Zhang, Q., J. Jia, and J. Zhang (2009, February). Cooperative relay to improve diversity in cognitive radio networks. *IEEE Communications Magazine 47*(2), 111–117.

Zhao, Q., S. Geirhofer, L. Tong, and B. Sadler (2008). Opportunistic spectrum access via periodic channel sensing. *IEEE Transactions on Signal Processing 56*(2), 785–796.

Brief Bios of Authors

Shining Wu is an assistant professor in Logistics and Maritime Studies at the Hong Kong Polytechnic University. His research interests include supply chain management, strategic consumer behavior, queueing theory and its applications, operational issues in spectrum sharing, and data-driven optimization for queueing systems.

Jiheng Zhang is an associate professor in Industrial Engineering and Decision Analytics at the Hong Kong University of Science and Technology. His research interests are in performance evaluation and optimal control via asymptotic analysis of queueing systems arising from applications in manufacturing and services.

Rachel Q. Zhang is a professor in Industrial Engineering and Decision Analytics at the Hong Kong University of Science and Technology. Her research interests include supply chain and inventory management, stochastic analysis of service operations, and the interface of finance and operations. Her research has appeared in such journals as Operations Research, Management Science, Interfaces, IIE Transactions, Naval Research Logistics, and Advances in Applied Probability.

Appendix A: Analysis of the Fluid Model

A.1. Proof of Theorem 1

We first provide the detailed background information about the theorem which will be used to prove the existence and uniqueness of the solution of the fluid model.

A.1.1. Theorem 3 in §10 of Filippov (1988)

Here, we translate the conditions in part 3 of §10 of Filippov (1988) for applying the theorem to our problem. Suppose that G is an n-dimensional domain, either open or closed, in $\mathbb{R}^n$.

1. Let $s(x)$, $x \in G \subseteq \mathbb{R}^n$, be a continuously differentiable function, and $S = \{x : s(x) = 0\}$ be a smooth surface that separates the domain G into $G^- = \{x : s(x) < 0\}$ and $G^+ = \{x : s(x) > 0\}$. Furthermore, the gradient $\nabla s(x) \neq 0$ on S.
2. Let $u(t,x)$ be a function on $\mathbb{R} \times G$ that is continuous up to the boundary of G^- and G^+ but is discontinuous on S. Let $u^-(t,x)$ and $u^+(t,x)$ be the limiting values of $u(t,x)$, in approaching $x \in S$ from domains G^- and G^+, respectively. Let $U(t,x)$ be an interval with the end points $u^-(t,x)$ and $u^+(t,x)$. Furthermore, $\frac{\partial u(t,x)}{\partial x_i}$, $i = 1, \cdots, n$, are continuous up to the boundary of G^- and G^+.
3. Let $f(t,x,u)$ be a continuous function from $\mathbb{R} \times G \times \mathbb{R}$ to $\mathbb{R}^n$ with continuous $\frac{\partial f}{\partial x_i}$ $(i = 1, \cdots, n)$ and $\frac{\partial f}{\partial u}$. Denote $f^-(t,x)$ and $f^+(t,x)$ to be the limiting values of $f(t,x,u(t,x))$, in approaching $x \in S$ from domains G^- and G^+, respectively. Let f_N, f_N^-, f_N^+ be projections of the vectors f, f^-, f^+ onto the normal to S, e.g., $f_N(t,x,u) = \frac{\nabla s(x) \cdot f(t,x,u)}{|\nabla s(x)|}$.
4. If $x \in S$ and $f_N^-(t,x) \cdot f_N^+(t,x) \leq 0$, then $u(t,x) \in U(t,x)$ and $f_N(t,x,u(t,x)) = 0$.

THEOREM 3 **(Theorem 3 in §10 of Filippov (1988))**. *Suppose that the differential equation*

$$\frac{dx}{dt} = f(t,x,u(t,x)) \tag{21}$$

whose elements are described in 1-4 above satisfies the following conditions

$$S \in C^2; \quad f, \frac{\partial f}{\partial u} \in C^1; \quad u^-(t,x), u^+(t,x) \in C^1;$$
$$\frac{\partial f_N(t,x,u)}{\partial u} \neq 0 \quad \textit{for all } u \in U(t,x).$$

If for each $t \in (a,b)$ at least one of the inequalities $f_N^- > 0$ or $f_N^+ < 0$ (possibly, different inequalities for different t and x) is valid at each point $x \in S$ then for $a < t < b$ in the domain G a solution with the initial data $x(t_0) = x_0 \in G$ exists and right uniqueness holds for (21).

A.1.2. Existence and Uniqueness of the Solution of the Fluid Model

We partition the four-dimensional set defined by Equation (3) into the following three regions.

$$\begin{aligned}\mathbb{S}_I &= \{(z_1, q_1, z_2, q_2) : q_1 > 0\},\\ \mathbb{S}_{II} &= \{(z_1, q_1, z_2, q_2) : z_1 + z_2 = 1,\ q_1 = 0\},\\ \mathbb{S}_{III} &= \{(z_1, q_1, z_2, q_2) : z_1 + z_2 < 1\}.\end{aligned}$$

Note that $\mathbb{S}_{II}$ is the intersection between $\mathbb{S}_I$ and $\mathbb{S}_{III}$. Thus, a solution cannot transit between $\mathbb{S}_I$ and $\mathbb{S}_{III}$ without visiting $\mathbb{S}_{II}$.

LEMMA 1. *The right hand sides of* (4)–(7) *are locally Lipschitz continuous within each region.*

Proof. Since $\beta(t)$ and $\alpha(t)$ are constant in $\mathbb{S}_I$ and $\mathbb{S}_{III}$, they are Lipschitz continuous in $\mathbb{S}_I$ and $\mathbb{S}_{III}$, respectively. Since $z_1(t) + z_2(t) = 1$ in $\mathbb{S}_{II}$, there exists $\delta > 0$ such that the denominator in (8), which can be written as $[\lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + p\mu z_2(t)] + [\lambda_1 + \lambda_2 + \theta q_2(t)](1-p)\mu z_2(t)$, is strictly greater than δ. Thus, $\beta(t)$ and $\alpha(t)$ are locally Lipschitz continuous in $\mathbb{S}_{II}$ since their derivatives or directional derivatives with respect to $x(t)$ are locally bounded, for example,

$$\begin{aligned}\left|\frac{\partial \beta(t)}{\partial q_2(t)}\right| &\le \left|\frac{[\mu_1 z_1(t) + p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t) - \lambda_1]}{\{[\lambda_1 + \lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + \mu z_2(t)] - \lambda_1[\mu_1 z_1(t) + p\mu z_2(t)]\}^2}\right|\\ &\le \left|\frac{[\mu_1 z_1(t) + p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t) - \lambda_1]}{\delta^2}\right| < \infty.\end{aligned}$$

□

Note that Lemma 1 guarantees the existence and uniqueness of the local solution evolving within each of the three regions. Since the right hand sides of (4)–(7) are only piecewise continuous in the whole state space and $\mathbb{S}_{II}$ is a "surface" of discontinuity that separates $\mathbb{S}_I$ from $\mathbb{S}_{III}$, we need to establish that the solution has a unique way transiting between regions. Specifically, we need to examine the behavior of the system when it approaches/crosses/deviates from the surface of discontinuity and rule out the possibilities that a solution starting from a point on the surface can evolve in more than one way. Below, Lemma 2 will first narrow down the possible evolutions by analyzing the values of (4)–(7) in the both-sided neighborhood of $\mathbb{S}_{II}$. Then, Theorem 4 will invoke Theorem 3 which considers the limiting values of the right hand sides of (4)–(7) as a solution enters the surface of discontinuity from both sides to establish the existence and uniqueness of the solution.

LEMMA 2. *Any local solution that starts from a point in* $\mathbb{S}_{II}$ *will either enter* $\mathbb{S}_I$ *or stay in* $\mathbb{S}_{II} \cup \mathbb{S}_{III}$.

Proof. Suppose that $x(\tau) \in \mathbb{S}_{II}$ at some time τ, i.e., $z_1(\tau) + z_2(\tau) = 1$ and $q_1(\tau) = 0$. In this case, $\beta(\tau)$ is given by (8). If we let

$$\zeta(t) = \frac{[\lambda_1 + \lambda_2 + \theta q_2(t) - \mu_1 z_1(t) - p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t)]}{[\lambda_1 + \lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + \mu z_2(t)] - \lambda_1[\mu_1 z_1(t) + p\mu z_2(t)]},$$

$\beta(\tau) = \min\{[\zeta(\tau)]^+, 1\}$. We now discuss the solution in a small time interval after τ for different values of $\zeta(\tau)$.

(i) If $\zeta(\tau) > 1$, i.e., $\mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] < \lambda_1$, then $\beta(\tau) = 1$ and there exists $\tilde{\tau} > \tau$ such that $\mu_1 z_1(t) + \mu[1 - z_1(t)] < \lambda_1$ for all $t \in [\tau, \tilde{\tau})$ by the continuity of $z_1(t)$ in t. Hence, $\beta(t) = 1$ and $q_1'(t) = \lambda_1 - \mu_1 z_1(t) - \mu z_2(t) > 0$ for all $t \in [\tau, \tilde{\tau})$. Thus, $q_1(t) > 0$ for all $t \in [\tau, \tilde{\tau})$ and any local solution, if exists, must enter $\mathbb{S}_I$.

(ii) If $\zeta(\tau) \leq 1$, we show by the following cases (a) and (b) that there exists $\tilde{\tau} > \tau$ such that $\mu_1 z_1(t) + \mu[1 - z_1(t)] > \lambda_1$ for $t \in (\tau, \tilde{\tau})$. Then, by (5), $q_1'(t) \leq 0$ and hence $q_1(t) = 0$ for all $t \in (\tau, \tilde{\tau})$. That is, a local solution, if exists, will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ for $t \in [\tau, \tilde{\tau})$.

(a) If $\zeta(\tau) = 1$, i.e., $\mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] = \lambda_1$, then $\mu < \lambda_1 < \mu_1$, $z_1(\tau) = \frac{\lambda_1 - \mu}{\mu_1 - \mu} \in \left(0, \frac{\lambda_1}{\mu_1}\right)$, $\beta(\tau) = \alpha(\tau) = 1$ and $z_1'(\tau) = \mu z_2(\tau) > \mu(1 - \frac{\lambda_1}{\mu_1}) > 0$. Since the right hand side of (4) is continuous within $\mathbb{S}_{II}$, there exists $\delta > 0$ such that $z_1'(t) > 0$ for any $x(t) \in \mathbb{S}_{II}$ and $||x(t) - x(\tau)|| < \delta$. Furthermore, since $z_1(\tau) \in \left(0, \frac{\lambda_1}{\mu_1}\right)$ and $z_2(\tau) = 1 - z_1(\tau)$ are continuous in t, there exists $\tilde{\tau} > \tau$ such that

$$z_1(t) < \frac{\lambda_1}{\mu_1}, \quad z_2(t) > 1 - \frac{\lambda_1}{\mu_1} > 0, \quad ||x(t) - x(\tau)|| < \delta$$

for all $t \in (\tau, \tilde{\tau})$. Next, we show that $z_1'(t) > 0$ and hence $z_1(t) > z_1(\tau)$ for all $t \in (\tau, \tilde{\tau})$.

- If $x(t) \in \mathbb{S}_I$, then $\beta(t) = \alpha(t) = 1$ and $z_1'(t) = \mu z_2(t) > 0$.
- If $x(t) \in \mathbb{S}_{II}$, then $z_1'(t) > 0$ since $||x(t) - x(\tau)|| < \delta$.
- If $x(t) \in \mathbb{S}_{III}$, then $\beta(t) = \alpha(t) = 0$ and $z_1'(t) = \lambda_1 - \mu_1 z_1(t) > 0$.

Thus, $\mu_1 z_1(t) + \mu[1 - z_1(t)] > \mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] = \lambda_1$ since $z_1(t) > z_1(\tau)$ and $\mu < \mu_1$.

(b) If $\zeta(\tau) < 1$, i.e., $\mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] > \lambda_1$, then $\beta(\tau) < 1$ and there exists $\tilde{\tau} > \tau$ such that $\mu_1 z_1(t) + \mu[1 - z_1(t)] > \lambda_1$ for all $t \in [\tau, \tilde{\tau})$ by the continuity of $z_1(t)$ in t.

□

Note that Lemma 2 has not shown the existence of a local solution starting from a point in $\mathbb{S}_{II}$. It only narrows down the possible evolutions of such a solution to two cases, which simplifies the proof of the existence and uniqueness in the following Theorem 4. Specifically, an explicit solution will be derived for case (i), where the uniqueness is guaranteed by the Lipschitz continuity of the ODEs in $\mathbb{S}_I$. For case (ii), although an explicit solution is almost impossible to obtain due to the non-linearity of the ODEs, we will show the existence and uniqueness simultaneously using Theorem 3.

THEOREM 4. *There exists a unique solution to the differential equations* (4)–(7).

Proof. Since the right hand sides of (4)–(7) are locally Lipschitz continuous within each region by Lemma 1, existence and uniqueness within each region follow directly by the Picard-Lindelöf

theorem (Theorem 2.2 of Teschl (2012)). If a solution transits across the regions through some point $x(\tau) \in \mathbb{S}_{II}$ at some time τ, it will either enter $\mathbb{S}_I$ or stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ by Lemma 2. We now show the existence and uniqueness of the solution for $t \in (\tau, \tilde{\tau})$, where $\tilde{\tau}$ is specified in Lemma 2.

(i) If $\zeta(\tau) > 1$, a local solution, if exists, will enter $\mathbb{S}_I$ by Lemma 2. Therefore, we only need to show the existence of a local solution in $\mathbb{S}_I$ as the uniqueness is guaranteed by the Lipschitz continuity of the ODEs in $\mathbb{S}_I$. Solving the linear ODEs (4)–(7) in $\mathbb{S}_I$, we obtain the following local solution in $\mathbb{S}_I$ for $t \in (\tau, \tilde{\tau})$.

$$
\begin{aligned}
z_2(t) &= z_2(\tau)e^{-\mu(t-\tau)}, \\
z_1(t) &= 1 - z_2(t) = 1 - [1 - z_1(\tau)]e^{-\mu(t-\tau)}, \\
q_1(t) &= (\lambda_1 - \mu_1)(t-\tau) + \frac{\mu_1 - \mu}{\mu} z_2(\tau)(1 - e^{-\mu(t-\tau)}), \\
q_2(t) &= q_2(\tau)e^{-\phi\theta(t-\tau)} + (1-\phi)\left[\frac{\lambda_2}{\phi\theta}(1 - e^{-\phi\theta(t-\tau)}) + \frac{(1-p)\mu z_2(\tau)}{\phi\theta - \mu}(e^{-\mu(t-\tau)} - e^{-\phi\theta(t-\tau)})\right].
\end{aligned}
$$

(ii) If $\zeta(\tau) \le 1$, a local solution, if exists, will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ by Lemma 2. Therefore, we only need to prove that there exists a unique local solution $x(t) = \big(z_1(t), 0, z_2(t), q_2(t)\big) \in \mathbb{S}_{II} \cup \mathbb{S}_{III}$. Note that $\mu_1 z_1(t) + \mu[1 - z_1(t)] \ge \lambda_1$ and $\zeta(t) \le 1$ for all $t \in [\tau, \tilde{\tau})$ as shown in case (ii) of Lemma 2.

To apply Theorem 3, we need to relate our setting to the four elements in section A.1.1.

(a) Let $s(x) = z_1 + z_2 - 1$, $G = \{x : q_1 = 0\}$, $S = \mathbb{S}_{II} = \{x : z_1 + z_2 = 1, q_1 = 0\}$, $G^- = \mathbb{S}_{III} = \{x : z_1 + z_2 < 1, q_1 = 0\}$ and $G^+ = \{x : z_1 + z_2 > 1, q_1 = 0\}$. Then, $s(x)$ is continuously differentiable and $\nabla s(x) = (1, 0, 1, 0)^T \neq 0$.

(b) Let $u(t,x) = \beta(t)$, if $x \in S \cup G^-$, and $u(t,x) = c > \frac{\lambda_1}{\lambda_2} + 1$ if $x \in G^+$. Then, $u(t,x) = 0$ in G^-, $u(t,x) = c$ in G^+ and $u(t,x)$ is discontinuous on S. Thus, $u^-(t,x) = 0$, $u^+(t,x) = c$ and $U(t,x) = [0, c]$ for $x \in S$.

(c) Let $f(t, x, u(t,x))$ be the right hand sides of (4)–(7) with $\beta(t)$ replaced by $u(t,x)$ and $\alpha(t)$ replaced by $\frac{\lambda_1 u(t,x)}{\mu_1 z_1 + \mu z_2}$. Then, it is obvious that $\frac{\partial f}{\partial x_i}$ and $\frac{\partial f}{\partial u}$ are continuous, and

$$
\begin{aligned}
f_N\big(t, x, u(t,x)\big) = \frac{1}{\sqrt{2}}\Bigg\{ & [1 - u(t,x)](\lambda_1 + \lambda_2 + \theta q_2) + \frac{\lambda_1 u(t,x)}{\mu_1 z_1 + \mu z_2}(\mu_1 z_1 + \mu z_2) \\
& - \mu_1 z_1 - \left[p + \frac{\lambda_1 u(t,x)}{\mu_1 z_1 + \mu z_2}(1-p)\right]\mu z_2 \Bigg\}, \\
f_N^-\big(t,x\big) = \frac{1}{\sqrt{2}}(\lambda_1 + \lambda_2 + \theta q_2 - \mu_1 z_1 - p\mu z_2), & \\
f_N^+\big(t,x\big) = \frac{1}{\sqrt{2}}\left\{\lambda_1 - (c-1)\lambda_2 - (c-1)\theta q_2 - \mu_1 z_1 - \left[p + \frac{\lambda_1 c}{\mu_1 z_1 + \mu z_2}(1-p)\right]\mu z_2\right\} < 0. &
\end{aligned}
$$

(d) If $x \in S$ and $f_N^-(t,x) \cdot f_N^+(t,x) \leq 0$, then $\lambda_1 + \lambda_2 + \theta q_2 - \mu_1 z_1 - p\mu z_2 \geq 0$ and $\zeta(t) \geq 0$. Thus, $u(t,x) = \beta(t) = \zeta(t) \in [0,1] \subseteq U(t,x)$, $z_1'(t) + z_2'(t) = 0$ and $f_N(t,x,u(t,x)) = 0$.

Given that $\frac{\partial f_N(t,x,u)}{\partial u} = \frac{1}{\sqrt{2}} \left[-(\lambda_2 + \theta q_2) - \frac{\lambda_1(1-p)\mu z_2}{\mu_1 z_1 + \mu z_2} \right] < 0$ and all the conditions in Theorem 3 hold, a unique solution can be found in $G^+ \cup \mathbb{S}_{II} \cup \mathbb{S}_{III}$. Since $f_N^+(t,x) < 0$ for any $x \in \mathbb{S}_{II}$, the solution will only be in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ starting from a point in $\mathbb{S}_{II}$. Therefore, this solution is also the unique solution to (4)–(7) in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$.

Thus, local existence and uniqueness of the solution can be guaranteed. Since the right hand sides of (4)–(7) are bounded, e.g., $|z_1'(t)| = |[1-\beta(t)]\lambda_1 + \alpha(t)[\mu_1 z_1(t) + \mu z_2(t)] - \mu_1 z_1(t)| \leq \lambda_1 + \mu_1 + \mu$, the solution will not go to infinity in a finite amount of time. Therefore, the unique local solution can be extended to the whole space as $t \to \infty$ by Theorem 2.17 of Teschl (2012). □

Based on the above proof, we can summarize the evolution of the solution. At an arbitrary moment τ, the evolution of a solution *within a small amount of time after* τ can be determined as follows.

(i) If $x(\tau) \in \mathbb{S}_I$, the solution will stay in $\mathbb{S}_I$ until it reaches the boundary of $\mathbb{S}_I$, i.e., $\mathbb{S}_{II}$, at some time. A closed form expression of the solution can be obtained by solving (4)–(7) with $\beta(t) = \alpha(t) = 1$.

(ii) If $x(\tau) \in \mathbb{S}_{II}$, its local evolution can be classified into the following cases by the value of $\zeta(\tau)$ defined in Lemma 2.
   - If $\zeta(\tau) > 1$, the solution will enter $\mathbb{S}_I$.
   - If $0 < \zeta(\tau) \leq 1$, the solution will stay in $\mathbb{S}_{II}$.
   - If $\zeta(\tau) = 0$, the solution will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$. Our proof doesn't rule out the possibility that the solution transits infinite times between $\mathbb{S}_{II}$ and $\mathbb{S}_{III}$ within a finite amount of time after τ.
   - If $\zeta(\tau) < 0$, the solution will enter $\mathbb{S}_{III}$.

(iii) If $x(\tau) \in \mathbb{S}_{III}$, the solution will stay in $\mathbb{S}_{III}$ forever or reaches the boundary of $\mathbb{S}_{III}$, i.e., $\mathbb{S}_{II}$, after some time. A closed form expression of the solution can be obtained by solving (4)–(7) with $\beta(t) = \alpha(t) = 0$.

Thus, the local evolution of a solution can be determined at every moment of time according to the above cases. Extending the process in time allows us to obtain the evolution of the global solution. For example, if the initial state $x(0)$ is in $\mathbb{S}_{III}$, the solution will first stay in $\mathbb{S}_{III}$ as in (iii). Then, depending on the system parameters and the initial state, the closed form expression will tell us whether the solution will reach the boundary $\mathbb{S}_{II}$ or not. Suppose that the solution reaches $\mathbb{S}_{II}$ at some time τ. Then, the solution will evolve according to (ii) within a small amount of time after τ, e.g., $(\tau, \tilde{\tau})$, depending on how the solution reaches $\mathbb{S}_{II}$, i.e., the value of $\zeta(\tau)$. For instance, if $\zeta(\tau) > 1$, this solution will enter $\mathbb{S}_I$ immediately after τ. That is, the solution transits

from $\mathbb{S}_{III}$ to $\mathbb{S}_I$ through a point $x(\tau) \in \mathbb{S}_{II}$ without staying in $\mathbb{S}_{II}$. Following the same process, we can determine the evolution of the solution from time $\tilde{\tau}$ until we obtain the global solution as $t \to \infty$.

A.1.3. Convergence to the Steady State

Let $x(\infty)$ denote the limit given in the Theorem. It is easy to verify that $x(\infty)$ is an invariant state, i.e., $x'(t) = 0$ for all $t \geq 0$ if $x(0) = x(\infty)$. We will show that $\lim_{t\to\infty} x(t) = x(\infty)$ for any initial state $x(0)$.

We first show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ after a finite amount of time for any initial state by the following argument. Note that it is impossible for the process $x(\cdot)$ to travel directly between $\mathbb{S}_{III}$ and $\mathbb{S}_I$ without visiting $\mathbb{S}_{II}$.

i. Suppose $x(0) \in \mathbb{S}_I$. We will show that there exists a $\tau < \infty$ such that $q_1(\tau) = 0$. In other words, the solution $x(\tau) \in \mathbb{S}_{II}$ if $x(0) \in \mathbb{S}_I$. Suppose that $x(t) \in \mathbb{S}_I$ in which case $q_1(t) > 0$ and $z_1(t) + z_2(t) = 1$ for all $t \geq 0$. Then, the differential equations (4) and (6) become $z_1'(t) = \mu z_2(t) \geq 0$ and $z_2'(t) = -\mu z_2(t)$, respectively, for all t. Thus, as t increases, $z_2(t)$ decreases while $z_1(t)$ increases at the same rate and $\lim_{t\to\infty} z_1(t) = 1$. In the meantime, the differential equation (5) becomes

$$q_1'(t) = \lambda_1 - [\mu_1 z_1(t) + \mu z_2(t)].$$

Since $\lambda_1 < \mu_1$ by Definition 1, there must exist a finite time τ and $\kappa < 0$ such that $q_1'(t) < \kappa < 0$ for all $t \geq \tau$. This implies that $q_1(\cdot)$ has to hit 0 in a finite amount of time, a contradiction.

So upon returning to $\mathbb{S}_{II}$ at τ, we must have

$$\lambda_1 - [\mu_1 z_1(\tau) + \mu z_2(\tau)] < 0. \tag{22}$$

ii. Suppose $x(0) \in \mathbb{S}_{II}$. For any t such that $x(t) \in \mathbb{S}_{II}$, substituting (8) and (9) into (5), we have

$$q_1'(t) = [\lambda_1 - \mu_1 z_1(t) - \mu z_2(t)]^+. \tag{23}$$

(a) If $\lambda_1 \leq \mu$, then $\lambda_1 - \mu_1 z_1(t) - \mu z_2(t) \leq 0$ because $z_1(t) + z_2(t) = 1$ and $q_1'(t) = 0$ by (23). This implies that the process $x(\cdot)$ will never move from $\mathbb{S}_{II}$ to $\mathbb{S}_I$.

(b) If $\lambda_1 > \mu$, then by (23) $q_1'(t) = 0$ if and only if $z_1(t) \geq z_1^\dagger := \frac{\lambda_1 - \mu}{\mu_1 - \mu}$ and it is possible that the process $x(\cdot)$ will move from $\mathbb{S}_{II}$ to $\mathbb{S}_I$. However, once the process is in $\mathbb{S}_I$, it will move back to $\mathbb{S}_{II}$ in a finite amount of time, say at time τ at which $z_1(\tau) > z_1^\dagger$ by (22). Next we show that $z_1(t) > z_1^\dagger$ for $t > \tau$ so that $x(\cdot)$ will never go back to $\mathbb{S}_I$ again. Suppose there exists a finite $\tau_1 > \tau$ such that $z_1(\tau) \leq z_1^\dagger$. Then, by the mean value theorem, there must exist some $\tau_2 \in (\tau, \tau_1)$ such that $z_1^\dagger < z_1(\tau_2) < \min\{z_1(\tau), \frac{\lambda_1}{\mu_1}\}$ and $z_1'(\tau_2) < 0$. However, for $z_1(t) > z_1^\dagger$ the differential equation (4) becomes

$$z_1'(t) = \lambda_1 - \mu_1 z_1(t), \tag{24}$$

which implies $z_1'(\tau_2) = \lambda_1 - \mu_1 z_1(\tau_2) > 0$, a contradiction. So $z_1(t) > z_1^\dagger$ and $q_1'(t) = 0$ for all $t > \tau$ and $x(\cdot)$ will not go back to $\mathbb{S}_I$ again, i.e., stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$.

In summary, if $x(0) \in \mathbb{S}_{II}$, the process $x(\cdot)$ will either stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ or visit $\mathbb{S}_I$ at most once before coming back to $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ after a finite amount of time.

iii. Suppose $x(0) \in \mathbb{S}_{III}$. If $x(\cdot)$ ever leaves $\mathbb{S}_{III}$, it will first visit $\mathbb{S}_{II}$. As we discussed above, it will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ after a finite amount of time.

Next, we derive the steady state of the fluid model by assuming that $x(t) \in \mathbb{S}_{II} \cup \mathbb{S}_{III}$ in which $q_1(t) = 0$ and the differential equation (4) becomes (24) or $z_1(t) = \frac{\lambda_1}{\mu_1} - [\frac{\lambda_1}{\mu_1} - z_1(0)]e^{-\mu_1 t}$. So

$$\lim_{t\to\infty} z_1(t) = \frac{\lambda_1}{\mu_1}. \tag{25}$$

To derive the steady state of $z_2(t)$ and $q_2(t)$, we need to consider the following three cases.

1. $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{p\mu} > 1$. We first show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{II}$ and then analyze the steady state of $z_2(t)$ and $q_2(t)$ in $\mathbb{S}_{II}$.

   (a) If $x(0) \in \mathbb{S}_{III}$, then there exists $\tau > 0$ such that $z_1(t) + z_2(t) < 1$ and $x(t) \in \mathbb{S}_{III}$ for all $t \in [0, \tau)$. Then the ODEs (4) and (6) become $z_1'(t) = \lambda_1 - \mu_1 z_1(t)$ and $z_2'(t) = \lambda_2 + \theta q_2(t) - p\mu z_2(t)$, respectively, for $t \in [0, \tau)$. Since $\lambda_1 + \lambda_2 > \lambda_1 + p\mu(1 - \frac{\lambda_1}{\mu_1}) = \lim_{t\to\infty}[\mu_1 z_1(t) + p\mu(1 - z_1(t))]$ by (25), there exist $\tau_1 \geq 0$ and an $\epsilon > 0$ such that

   $$\lambda_1 + \lambda_2 > \mu_1 z_1(t) + p\mu[1 - z_1(t)] + \epsilon \geq \mu_1 z_1(t) + p\mu z_2(t) + \epsilon \tag{26}$$

   for all $t \geq \tau_1$. This implies that $z_1'(t) + z_2'(t) = \lambda_1 + \lambda_2 + \theta q_2(t) - \mu_1 z_1(t) - p\mu z_2(t) > \epsilon$ for all $t \in [\tau_1, \infty)$. Hence $z_1(t) + z_2(t)$ will increase until it reaches 1 or $x(\cdot)$ moves to $\mathbb{S}_{II}$ after a finite amount of time.

   (b) If $x(0) \in \mathbb{S}_{II}$, the process $x(\cdot)$ will go back to $\mathbb{S}_{II}$ even if it moves to $\mathbb{S}_{III}$ as shown above. Thus, there exists a finite $\tau \geq 0$ such that $x(\tau) \in \mathbb{S}_{II}$. We next show that $x(\cdot)$ will then stay in $\mathbb{S}_{II}$ for $t \geq \tau$. By the analysis in i(a) and ii(b),

   $$\lambda_1 < \mu_1 z_1(t) + \mu[1 - z_1(t)] \tag{27}$$

   holds for all $t \geq 0$. Since (26) and (27) hold for $t \geq \tau$, we have

   $$\beta(t) = \frac{[\lambda_1 + \lambda_2 + \theta q_2(t) - \mu_1 z_1(t) - p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t)]}{[\lambda_1 + \lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + \mu z_2(t)] - \lambda_1[\mu_1 z_1(t) + p\mu z_2(t)]},$$
   $$\alpha(t) = \frac{\lambda_1 \beta(t)}{\mu_1 z_1(t) + \mu z_2(t)}.$$

   Substituting them into (4) and (6), we obtain $z_2'(t) = -\lambda_1 + \mu_1 z_1(t) = -z_1'(t)$. This implies that $z_1'(t) + z_2'(t) = 0$ for $t \geq \tau$ and $x(\cdot)$ stays in $\mathbb{S}_{II}$.

Substituting (8) and (9) into (7), we have

$$q_2'(t) = \frac{g(x(t))}{[\lambda_1+\lambda_2+\theta q_2(t)][\mu_1 z_1(t)+\mu z_2(t)] - \lambda_1[\mu_1 z_1(t)+p\mu z_2(t)]}, \tag{28}$$

where

$$\begin{aligned} g(x(t)) = {} & (1-\phi)[\lambda_1+\lambda_2+\theta q_2(t)-\mu_1 z_1(t)-p\mu z_2(t)][\mu_1 z_1(t)+\mu z_2(t)][\lambda_2+\theta q_2(t)] \\ & + (1-\phi)[\lambda_1+\lambda_2+\theta q_2(t)-\mu_1 z_1(t)-p\mu z_2(t)]\lambda_1(1-p)\mu z_2(t) \\ & - \theta q_2(t)[\lambda_1+\lambda_2+\theta q_2(t)][\mu_1 z_1(t)+\mu z_2(t)] + \theta q_2(t)\lambda_1[\mu_1 z_1(t)+p\mu z_2(t)]. \end{aligned}$$

For any given $z_1(t)$ and $z_2(t)$, $g(\cdot)$ is a concave quadratic function of $q_2(t)$ and positive at $q_2(t)=0$ by (26) and (27). Furthermore, since the denominator in (28) is positive, there exists a threshold $\hat{q}_2(z_1(t),z_2(t))$ such that $q_2'(t)>0$ if $q_2(t)<\hat{q}_2(z_1(t),z_2(t))$ and $q_2'(t)\le 0$ otherwise. Thus, there exist a C_i such that $\left|\frac{\partial \hat{q}_2(z_1(t),z_2(t))}{\partial z_i(t)}\right| < C_i$ for all $t\ge\tau$ where $i=1,2$.

We are now ready to construct a Lyapunov function to show the convergence. For any $t\ge\tau$, let $V(x(t)) = C_1\left|z_1(t)-\frac{\lambda_1}{\mu_1}\right| + C_2\left|z_2(t)-\left(1-\frac{\lambda_1}{\mu_1}\right)\right| + |q_2(t)-\hat{q}_2(z_1(t),z_2(t))|$, which is zero only at $x(\infty)=\left(\frac{\lambda_1}{\mu_1},0,1-\frac{\lambda_1}{\mu_1},\frac{1-\phi}{\theta\phi}\left[\lambda_2-p\mu\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\right)$ and positive elsewhere. Suppose $q_2(t)>\hat{q}_2(z_1(t),z_2(t))$, then $\frac{dV(x(t))}{dt} = -C_1|z_1'(t)| - C_2|z_2'(t)| + q_2'(t) - \frac{\partial \hat{q}_2(z_1(t),z_2(t))}{\partial z_1(t)}z_1'(t) - \frac{\partial \hat{q}_2(z_1(t),z_2(t))}{\partial z_2(t)}z_2'(t) < 0$. Similarly, we can show $\frac{dV(x(t))}{dt}\le 0$ when $q_2(t)\le\hat{q}_2(z_1(t),z_2(t))$. Thus, $V(x(t))$ is a Lyapunov function and hence $\lim_{t\to\infty} x(t) = x(\infty)$.

Substituting $x(\infty)$ into (8) and (9), we can obtain $\beta(\infty)$ and $\alpha(\infty)$ that satisfy (10)–(12). For example,

$$\beta = \frac{\lambda_2 - p\mu(1-\frac{\lambda_1}{\mu_1})}{\lambda_2 - p\mu(1-\frac{\lambda_1}{\mu_1}) + \phi\mu(1-\frac{\lambda_1}{\mu_1})\frac{\lambda_1+p\mu(1-\frac{\lambda_1}{\mu_1})}{\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})}}. \tag{29}$$

2. $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{p\mu}<1$. We first derive the steady state of $q_2(t)$, $\alpha(t)$ and $\beta(t)$, and then show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{III}$ before deriving the steady state of $z_2(t)$.

When $x(t)\in\mathbb{S}_{II}$, by (25)–(28), for any $\epsilon>0$, there exist $A>0$ and $\tau>0$ such that

$$q_2'(t) \le -Aq_2(t)+\epsilon. \tag{30}$$

When $x(t)\in\mathbb{S}_{III}$, the differential equation (7) is

$$q_2'(t) = -\theta q_2(t). \tag{31}$$

In either case, $\lim_{t\to\infty} q_2(t) = 0$.

Note that $\lambda_1+\lambda_2 < \lambda_1+p\mu(1-\frac{\lambda_1}{\mu_1}) = \lambda_1 + \lim_{t\to\infty}[\mu_1 z_1(t)+p\mu(1-z_1(t))]$ by (25). Thus, after a finite amount of time,

$$\lambda_1+\lambda_2+\theta q_2(t) < \mu_1 z_1(t)+p\mu[1-z_1(t)] \tag{32}$$

and $\beta(t)=\alpha(t)=0$ as long as $x(t)\in\mathbb{S}_{II}\cup\mathbb{S}_{III}$. In this case, $z_1'(t)+z_2'(t)=\lambda_1+\lambda_2+\theta q_2(t)-\mu_1 z_1(t)-p\mu z_2(t)$ after substituting $\beta(t)=\alpha(t)=0$ into (4) and (6).

We now show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{III}$ and derive the steady state of $z_2(t)$. If $x(\tau)\in\mathbb{S}_{II}$, then there exists a small $\delta>0$ such that $x(\tau+t)\in\mathbb{S}_{III}$ for all $t\in(0,\delta]$ (i.e., $x(\cdot)$ will immediately leave $\mathbb{S}_{II}$) as $z_1'(\tau+t)+z_2'(\tau+t)<0$ for $t\in[0,\delta]$ by (32). If $x(\tau)\in\mathbb{S}_{III}$, then $x(\cdot)$ will stay in $\mathbb{S}_{III}$ because $z_1(t)+z_2(t)$ can never increase to 1 by (32). Thus, no matter whether $x(0)$ is in $\mathbb{S}_{II}$ or $\mathbb{S}_{III}$, $x(t)\in\mathbb{S}_{III}$ for t large enough. Let $V(x(t))=\left|z_1(t)-\frac{\lambda_1}{\mu_1}\right|+\left|z_2(t)-\frac{\lambda_2}{p\mu}\right|+q_2(t)$, which is zero only at $x(\infty)=\left(\frac{\lambda_1}{\mu_1},0,\frac{\lambda_2}{p\mu},0\right)$ and positive elsewhere. Then, $\frac{dV(x(t))}{dt}=-|z_1'(t)|-|z_2'(t)|-\theta q_2(t)$ for $z_2(t)\le\frac{\lambda_2}{p\mu}$ and $z_2(t)\ge\frac{\lambda_2+\theta q_2(t)}{p\mu}$, and $\frac{dV(x(t))}{dt}=-|z_1'(t)|+z_2'(t)-\theta q_2(t)=-|z_1'(t)|+\lambda_2-p\mu z_2(t)$ otherwise for $x(t)\in\mathbb{S}_{III}$. $\frac{dV(x(t))}{dt}=0$ only when $x(t)=\left(\frac{\lambda_1}{\mu_1},0,\frac{\lambda_2}{p\mu},0\right)$ and $\frac{dV(x(t))}{dt}<0$ otherwise. Thus, $V(x(t))$ is a Lyapunov function and hence $\lim_{t\to\infty} z_2(t)=\frac{\lambda_2}{p\mu}$.

3. $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{p\mu}=1$. Note that (30) and (31) also hold in this case. Thus, $\lim_{t\to\infty} q_2(t)=0$. For $z_2(t)$, note that

$$\limsup_{t\to\infty} z_2(t)\le 1-\lim_{t\to\infty} z_1(t)=1-\frac{\lambda_1}{\mu_1}=\frac{\lambda_2}{p\mu}$$

since $z_2(t)\le 1-z_1(t)$. On the other hand,

$$\begin{aligned} z_2'(t) &= \begin{cases} \lambda_2+\theta q_2(t)-p\mu z_2(t), & \text{if } z_2(t)<1-z_1(t),\\ \lambda_2+\theta q_2(t)-p\mu z_2(t), & \text{if } \lambda_1+\lambda_2+\theta q_2(t)\le\mu_1 z_1(t)+p\mu z_2(t),\\ \frac{\mu_1}{p\mu}\big(\lambda_2-p\mu z_2(t)\big), & \text{otherwise,}\end{cases}\\ &\ge \min\left\{1,\frac{\mu_1}{p\mu}\right\}\cdot\big(\lambda_2-p\mu z_2(t)\big). \end{aligned}$$

Thus, $\liminf_{t\to\infty} z_2(t)\ge\frac{\lambda_2}{p\mu}$ and $\lim_{t\to\infty} z_2(t)=1-\frac{\lambda_1}{\mu_1}$. It is obvious that $\beta=\alpha=0$.

In all the cases, the limit $TH_2=\lim_{t\to\infty} p\mu z_2(t)=p\mu z_2(\infty)$. □

A.2. Proof of Corollary 1

By Theorem 1, $\beta=0$ when the system is effectively under or critically loaded, and β has a closed-form expression

$$\frac{1}{\beta}=1+\phi\frac{\mu(1-\frac{\lambda_1}{\mu_1})\left[\lambda_1+p\mu(1-\frac{\lambda_1}{\mu_1})\right]}{\left[\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})\right]\left[\lambda_2-p\mu(1-\frac{\lambda_1}{\mu_1})\right]} \tag{33}$$

$$=1-\phi+\frac{\phi}{1-\frac{p\mu}{\lambda_2}(1-\frac{\lambda_1}{\mu_1})}\left[1+\frac{\mu_t}{\mu_2}\cdot\frac{\lambda_1}{\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})}\cdot\frac{p\mu}{\lambda_2}(1-\frac{\lambda_1}{\mu_1})\right] \tag{34}$$

when the system is effectively overloaded. When $\mu_s\le\mu_2$, it can be easily seen from (33) that $\frac{1}{\beta}$ decreases in μ_t since both μ and $p\mu$ decrease in μ_t. When $\mu_s>\mu_2$, substitute $\mu_t=\frac{\mu_s(\mu-\mu_2)}{\mu_s-\mu}$ into (33) and consider $\mu\in[\mu_2,\mu_s]$ where $\mu=\mu_2$ when $\mu_t=0$ and $\mu=\mu_s$ when $\mu_t\to\infty$. Then,

$$\frac{d\{\frac{1}{\beta}\}}{d\mu}=\frac{\phi(1-\frac{\lambda_1}{\mu_1})}{\left[\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})\right]^2\left[\lambda_2-p\mu(1-\frac{\lambda_1}{\mu_1})\right]^2}\cdot h(\mu).$$

where

$$h(\mu) = -\frac{\mu_2}{\mu_s-\mu_2}\left(\frac{\mu_s}{\mu_s-\mu_2}\lambda_1+\lambda_2\right)\left(1-\frac{\lambda_1}{\mu_1}\right)^2\mu^2 - 2\frac{\mu_2}{\mu_s-\mu_2}\lambda_1\left(1-\frac{\lambda_1}{\mu_1}\right)\left[\lambda_2-\frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\mu + \lambda_1\left[\lambda_1+\frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\left[\lambda_2-\frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right].$$

Note that $h(\mu)$ is a concave quadratic function of μ. Also, $h(\mu)$ is decreasing in μ for $\mu\in[\mu_2,\infty)$ since the symmetric center of the concave quadratic function is below μ_2. So the sign of $h(\cdot)$ on $[\mu_2,\mu_s]$ depends on the value of

$$h(\mu_2) = \frac{\lambda_1\lambda_2}{\mu_s-\mu_2}\left[\lambda_1+\mu_2\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\left\{-\mu_2\left[1+\frac{\mu_2}{\lambda_1}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]+\mu_s\left[1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\right\},$$
$$h(\mu_s) = \lambda_1^2\lambda_2 - \frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\left[\lambda_1+\mu_s\left(1-\frac{\lambda_1}{\mu_1}\right)\right](\lambda_1+\lambda_2).$$

First, $h(\mu_2)>0$ if and only if

$$-\mu_2\left[1+\frac{\mu_2}{\lambda_1}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]+\mu_s\left[1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]>0,$$

i.e., $1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)>0$ and $\mu_s>\hat{\mu}_s$, where

$$\hat{\mu}_s = \frac{1+\frac{\mu_2}{\lambda_1}\left(1-\frac{\lambda_1}{\mu_1}\right)}{1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)}\mu_2.$$

Second, $h(\mu_s)\geq 0$ if and only if $-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})^2\mu_2\mu_s^2+\lambda_1\left[\lambda_1\lambda_2-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})\mu_2\right]\mu_s-\lambda_1^2\lambda_2\mu_2\geq 0$. If we treat the left hand size of this inequality as a quadratic function of μ_s, then it holds for some $\mu_s\in[\mu_s^\dagger,\mu_s^\ddagger]$ if and only if its discriminant is non-negative, i.e.,

$$\lambda_1\leq\lambda_2\left[1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}-2\sqrt{1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}}\right],$$

which is equivalent to $\lambda_1>\frac{3\mu_1\mu_2}{\mu_1+3\mu_2}$ and $\lambda_2\geq\hat{\lambda}_2$ where

$$\hat{\lambda}_2 = \frac{\lambda_1}{1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}-2\sqrt{1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}}}.$$

Furthermore, the values of $\mu_s^\dagger\leq\mu_s^\ddagger$ can be calculated by the quadratic formula as

$$\frac{\lambda_1\lambda_2-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})\mu_2\pm\sqrt{\left[\lambda_1\lambda_2-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})\mu_2\right]^2-4(\lambda_1+\lambda_2)\lambda_2(1-\frac{\lambda_1}{\mu_1})^2\mu_2^2}}{2(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})^2\mu_2}\lambda_1.$$

Due to the monotonicity of $h(\cdot)$ on $[\mu_2,\mu_s]$, we know $\hat{\mu}_s<\mu_s^\dagger\leq\mu_s^\ddagger$ when $(\mu_s^\dagger,\mu_s^\ddagger)$ exists. Now we are ready to discuss the sign of $h(\cdot)$.

1. If $\lambda_2 \le \mu_2\big(1-\frac{\lambda_1}{\mu_1}\big)$ or $\frac{1}{\mu_s} \ge \frac{1}{\hat{\mu}_s}$, then $h(\mu_2) \le 0$ and hence $h(\mu) \le 0$ or $\frac{d\{\frac{1}{\beta}\}}{d\mu} \le 0$ for all $\mu \in [\mu_2, \mu_s]$, which implies that β decreases in $\frac{1}{\mu_t}$.
2. If $\lambda_2 > \mu_2\big(1-\frac{\lambda_1}{\mu_1}\big)$ and $\frac{1}{\mu_s} < \frac{1}{\hat{\mu}_s}$, then $h(\mu_2) > 0$. It remains to discuss the sign of $h(\mu_s)$.
   - If $\lambda_1 > \frac{3\mu_1\mu_2}{\mu_1+3\mu_2}$, $\lambda_2 \ge \hat{\lambda}_2$ and $\frac{1}{\mu_s^\ddagger} \le \frac{1}{\mu_s} \le \frac{1}{\mu_s^\dagger}$, then $h(\mu_s) \ge 0$ and $h(\mu) \ge 0$ for all $\mu \in [\mu_2, \mu_s]$, which implies that β always increases in $\frac{1}{\mu_t}$ and hence $\frac{1}{\hat{\mu}_t} = 0$.
   - Otherwise, $h(\mu_s) < 0$ and $h(\mu)$ is first positive and then negative as μ increases from μ_2 to μ_s. This implies that β first decreases and then increases in $\frac{1}{\mu_t}$, and $\frac{1}{\hat{\mu}_t} > 0$.

□

A.3. Proof of Proposition 1

Suppose that the feasible region of Problem (13) is nonempty as the parameters change in all three cases. Since TH_2 is increasing in $\frac{1}{\mu_t}$, the optimization problem reduces to one of finding the largest $\frac{1}{\mu_t}$ that satisfies the delay constraint $\beta \le \eta$. By Corollary 1, β either monotonically decreases in $\frac{1}{\mu_t}$ (as in Figure 2(a)–(b)), in which case $\frac{1}{\mu_t^*} = \infty$, or first decreases and then increases in $\frac{1}{\mu_t}$ (as in Figure 2(c)–(d)). In the latter case, if η is large, $\frac{1}{\mu_t} = \infty$ is feasible and hence optimal. Otherwise, the line $\beta = \eta$ crosses the β curve at most twice or touches its lowest point and $\frac{1}{\mu_t^*}$ is finite and equal to the larger intersection, which lies in the increasing part of the curve.

By (33) and (34), we can easily see that β increases in the cases (2) and (3) in this proposition for a given $1/\mu_t$, i.e., the curves in Figure 2 move upwards as the parameters change in the cases (2) and (3), and hence $\frac{1}{\mu_t^*}$ will either remain as ∞ or decrease as long as the feasible region is still feasible. For the case (1), while keeping $\lambda_1 + \lambda_2 = C$,

$$\frac{d\{\frac{1}{\beta}\}}{d\lambda_1} = -\left[\mu - \frac{\lambda_1(\mu-\mu_2)}{\mu_2}\right] \cdot \frac{\phi(1-\frac{\lambda_1}{\mu_1})\left[\lambda_1 + p\mu(1-\frac{\lambda_1}{\mu_1})\right]}{\left[\lambda_1 + \mu(1-\frac{\lambda_1}{\mu_1})\right]^2\left[\lambda_2 - p\mu(1-\frac{\lambda_1}{\mu_1})\right]} - \frac{d\{\frac{1}{\beta}\}}{d\mu} \le -\frac{d\{\frac{1}{\beta}\}}{d\mu}.$$

This implies that $\frac{d\{\frac{1}{\beta}\}}{d\lambda_1} \le 0$ if $\frac{d\{\frac{1}{\beta}\}}{d\mu} \ge 0$. Thus, the increasing part, where $\frac{d\beta}{d\{\frac{1}{\mu_t}\}} \ge 0$, of the β curve move upwards as λ_1 increases, and hence $\frac{1}{\mu_t^*}$ will either remain as ∞ or decrease as λ_1 increases.

Since $TH_2 = p\mu\big(1-\frac{\lambda_1}{\mu_1}\big) = \frac{\mu_2\mu_s}{\mu_t+\mu_s}\big(1-\frac{\lambda_1}{\mu_1}\big)$, it is easy to see that TH_2^* decreases in cases (1) and (2). In case (3), since both μ_2 and μ_t^* increases, the change of TH_2^* is not known. □

Appendix B: Analysis of the Underlying Stochastic Process

Note that, to obtain the system dynamics, we need to keep track of the status of the unlicensed users in service, i.e., in transmission or sensing, as the actual length of a service session is a phase-type rather than exponential. Although we are able to obtain the system dynamics and the fluid approximation when the length of a service session is a phase-type, in this paper we will only present the system dynamics and all the subsequent analysis as if the length of a service session

were exponential with the same mean for the following reasons. (1) Do not burden the reader with heavy notation and tedious mathematical expressions with only a single phase in each service session. As a result, the dynamics and subsequent analysis are much easier to understand and more intuitive. (2) The fluid approximation with a phase-type (two exponential phases) service session can be obtained following similar arguments. (3) The fluid approximations with an exponential or phase-type service session lead to exactly the same steady-state performance as the rates at which the unlicensed users leave and enter service are the same in both cases.

B.1. System Dynamics

In addition to notation introduced in Section 3, let

$S_i^n(t) =$ total number of type i users who have completed their transmission by t,

$D_2^n(t) =$ total number of service sessions completed by the unlicensed users by t,

$C_2^n(t) =$ total number of times the unlicensed users in the orbit queue have performed sensing by t.

It is easy to see that $S_1^n(t)$, $D_2^n(t)$, and $C_2^n(t)$ are random-time-changed Poisson processes with the rates $\mu_1 Z_1^n(t)$, $\mu Z_2^n(t)$, and $\theta Q_2^n(t)$, respectively. Since an unlicensed user will leave the system at the end of a service session with probability p, $S_2^n(t)$ is a "thinned" Poisson process of $D_2^n(t)$ with a time-varying rate $p\mu Z_2^n(t)$. Next, we derive the dynamics of $Z_i^n(t)$ and $Q_i^n(t)$ for $i = 1, 2$.

Note that the number of licensed users in service increases whenever an arriving licensed user sees an idle channel or a waiting licensed user sees a licensed user completing service or an unlicensed user finishing a session, and decreases whenever a licensed user completes his service. Likewise, the queue length of the licensed users increases whenever an arriving licensed user sees a busy system and decreases whenever a waiting licensed user sees a service or session completion. Thus, we have

$$\begin{aligned} Z_1^n(t) =& Z_1^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)>0\}} d\Lambda_1^n(s) + \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} d[S_1^n(s) + D_2^n(s)] - S_1^n(t), \\ Q_1^n(t) =& Q_1^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)=0\}} d\Lambda_1^n(s) - \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} d[S_1^n(s) + D_2^n(s)]. \end{aligned}$$

The dynamics of the unlicensed users is more complex as they may go back and forth between in service and waiting. The number of unlicensed users in service increases whenever a new arrival or waiting unlicensed user sees an idle channel and decreases whenever an unlicensed user finishes his transmission or is interrupted. The number of unlicensed users in the orbit queue increases whenever an arriving unlicensed user sees a busy system or an unlicensed user is interrupted but is willing to wait and decreases whenever a waiting unlicensed user enters service or abandons the system. Then,

$$Z_2^n(t) = Z_2^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)>0\}} d[\Lambda_2^n(s) + C_2^n(s)] - \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} d[D_2^n(s) - S_2^n(s)] - S_2^n(t),$$

$$Q_2^n(t) = Q_2^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)=0\}}[1 - B_\Lambda^n(s)]d\Lambda_2^n(s) + \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}}[1 - B_D^n(s)]d[D_2^n(s) - S_2^n(s)]$$
$$- \int_0^t \left[\mathbf{1}_{\{I^n(s)>0\}} + \mathbf{1}_{\{I^n(s)=0\}} B_C^n(s)\right] dC_2^n(s),$$

where $B_\Lambda^n(s)$, $B_C^n(s)$ and $B_D^n(s)$ are Bernoulli random variables with parameter ϕ at any s.

B.2. Proof of Theorem 2

Step 1: Martingale Representation. Let

$$\bar{M}_{\Lambda,i} = \bar{\Lambda}_i^n(t) - \bar{\lambda}_i^n t, \qquad \bar{M}_{S,1}^n = \bar{S}_1^n(t) - \int_0^t \mu_1 \bar{Z}_1^n(s)ds, \quad \bar{M}_{S,2}^n = \bar{S}_2^n(t) - \int_0^t p\mu \bar{Z}_2^n(s)ds,$$
$$\bar{M}_{C,2}^n = \bar{C}_2^n(t) - \int_0^t \theta \bar{Q}_2^n(s)ds, \quad \bar{M}_{D,2}^n = \bar{D}_2^n(t) - \int_0^t \mu \bar{Z}_2^n(s)ds$$

be the martingales corresponding to the processes. Recall $m^n(t)$ defined in (15). Then, we can rewrite the system dynamics as

$$\bar{Z}_1^n(t) = \bar{Z}_1^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)<0\}} d\bar{M}_{\Lambda,1}^n(s) + \int_0^t \mathbf{1}_{\{m^n(s)>0\}} d[\bar{M}_{S,1}^n(s) + \bar{M}_{D,2}^n(s)] - \bar{M}_{S,1}^n(t)$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)<0\}} \bar{\lambda}_1^n ds + \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [\mu_1 \bar{Z}_1^n(s) + \mu \bar{Z}_2^n(s)]ds - \int_0^t \mu_1 \bar{Z}_1^n(s)ds,$$
$$\bar{Q}_1^n(t) = \bar{Q}_1^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} d\bar{M}_{\Lambda,1}^n(s) - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} d[\bar{M}_{S,1}^n(s) + \bar{M}_{D,2}^n(s)]$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} \bar{\lambda}_1^n ds - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [\mu_1 \bar{Z}_1^n(s) + \mu \bar{Z}_2^n(s)]ds,$$
$$\bar{Z}_2^n(t) = \bar{Z}_2^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)<0\}} d[\bar{M}_{\Lambda,2}^n(s) + \bar{M}_{C,2}^n(s)] - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} d[\bar{M}_{D,2}^n(s) - \bar{M}_{S,2}^n(s)] - \bar{M}_{S,2}^n(t)$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)<0\}} \left[\bar{\lambda}_2^n + \theta \bar{Q}_2^n(s)\right] ds - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [\mu \bar{Z}_2^n(s) - p\mu \bar{Z}_2^n(s)]ds - \int_0^t p\mu \bar{Z}_2^n(s)ds,$$
$$\bar{Q}_2^n(t) = \bar{Q}_2^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} [1 - B_\Lambda^n(s)] d\bar{M}_{\Lambda,2}^n(s) + \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [1 - B_D^n(s)] d[\bar{M}_{D,2}^n(s) - \bar{M}_{S,2}^n(s)]$$
$$- \int_0^t \left[\mathbf{1}_{\{m^n(s)<0\}} + \mathbf{1}_{\{m^n(s)\geq 0\}} B_C^n(s)\right] d\bar{M}_{C,2}^n(s) + \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} (1-\phi)\bar{\lambda}_2^n ds$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)>0\}} (1-\phi)[\mu \bar{Z}_2^n(s) - p\mu \bar{Z}_2^n(s)]ds - \int_0^t \left[\mathbf{1}_{\{m^n(s)<0\}} + \mathbf{1}_{\{m^n(s)\geq 0\}} \phi\right] \theta \bar{Q}_2^n(s)ds.$$

The dynamics of the process depends on the state of $m^n(t) \in \mathbb{Z}$. We compactify $\mathbb{Z}$ by letting $\bar{\mathbb{Z}} = \mathbb{Z} \cup \{\pm\infty\}$ (e.g., Perry and Whitt 2013) and denote by $\mathbb{M}$ the space of all measures ν on $[0,\infty) \times \bar{\mathbb{Z}}$ satisfying $\nu([0,t] \times \bar{\mathbb{Z}}) = t$. Consider the random measure $\nu^n(\cdot) \in \mathbb{M}$ defined by

$$\nu^n((0,t) \times \Gamma) = \int_0^t \mathbf{1}_{\{m^n(u)\in\Gamma\}} du \tag{35}$$

for all $t \in (0,\infty)$ and measurable $\Gamma \subset \bar{\mathbb{Z}}$.

Step 2: Tightness. Let $\mathbb{D}_{\mathbb{R}^4}[0,\infty)$ be the space of all right-continuous $\mathbb{R}^4$-valued functions with left limits defined on the real line. We show that the sequence $\{\bar{X}^n(\cdot),\nu^n\}$ is relatively compact in $\mathbb{D}_{\mathbb{R}^4}[0,\infty)\times\mathbb{M}$ by showing that both $\{\bar{X}^n(\cdot)\}$ and $\{\nu^n\}$ are relatively compact.

$\{\nu^n\}$ is relatively compact due to the compactness of $\mathbb{M}$, which follows from the compactness of $\bar{\mathbb{Z}}$ by Prohorov's theorem (cf. Theorem 11.6.1 in Whitt 2002). $\{\bar{X}^n(\cdot)\}$ is relatively compact in $\mathbb{D}_{\mathbb{R}^4}[0,\infty)$ if it satisfies the conditions (6.3) and (6.4) of Theorem 11.6.3 in Whitt (2002). For any $\epsilon>0$, there exists a $c>0$ such that

$$\mathbb{P}(|\bar{X}^n(0)|>c)<\epsilon,\ \text{for all } n\geq 1,$$

since $\bar{X}^n(0)\Rightarrow x(0)$. Thus, the initial states are stochastically bounded and hence condition (6.3) is satisfied.

To show that condition (6.4) is satisfied, for any $\delta>0$, we define the modulus of continuity for a function $y(\cdot)$ as

$$w(y(\cdot),\delta,T)=\sup_{|t-s|\leq\delta,\ s,t\in[0,T]}|y(t)-y(s)|,$$

and show that, for any $\epsilon,\eta,T>0$, there exists a δ such that

$$\mathbb{P}\left(w(\bar{X}^n(\cdot),\delta,T)>\epsilon\right)<\eta, \tag{36}$$

for all n large enough. To do so, we decompose the oscillations of the process $X^n(t)$. Take the component $Q_2^n(t)$ for example,

$$\begin{aligned}|\bar{Q}_2^n(t)-\bar{Q}_2^n(s)|\leq&|\bar{M}_{\Lambda,2}^n(t)-\bar{M}_{\Lambda,2}^n(s)|+|\bar{M}_{D,2}^n(t)-\bar{M}_{D,2}^n(s)|+|\bar{M}_{C,2}^n(t)-\bar{M}_{C,2}^n(s)|\\&+\int_s^t(1-\phi)\left[\bar{\lambda}_2^n+(1-p)\mu\right]du+\int_s^t\theta\bar{Q}_2^n(u)du.\end{aligned}$$

Since the fourth term on the right hand side is deterministic and uniformly continuous, there exists a $\delta'>0$ such that it is less than $\frac{\epsilon}{5}$, i.e., $\mathbb{P}\left(w\left(\int_0^t(1-\phi)\left[\bar{\lambda}_2^n+(1-p)\mu\right]du,\delta',T\right)>\frac{\epsilon}{5}\right)=0$. Furthermore, since $\bar{M}_{\Lambda,2}^n$, $\bar{M}_{D,2}^n$ and $\bar{M}_{C,2}^n$ are square-integrable martingales, they weakly converge to 0 as $n\to\infty$ by Doob's inequality and hence their oscillations can also be controlled, e.g., $\mathbb{P}\left(w\left(\bar{M}_{\Lambda,2}^n(\cdot),\delta',T\right)>\frac{\epsilon}{5}\right)<\frac{\eta}{5}$ for large enough n. For the last term, we can bound the process $\bar{Q}_2^n(t)$ by a stable and bounded auxiliary one with simple dynamics. Thus, there exists a constant c such that $\mathbb{P}\left(\sup_{t\in[0,T]}\{\bar{Q}_2^n(t)\}>c\right)\leq\frac{\eta}{5}$ for all large n. Let $\delta=\min\left\{\frac{\epsilon}{5\theta c},\delta'\right\}$. Then,

$$\mathbb{P}\left(\sup_{|t-s|\leq\delta,\ s,t\in[0,T]}\left\{\int_s^t\theta\bar{Q}_2^n(u)du\right\}>\frac{\epsilon}{5}\right)\leq\frac{\eta}{5}$$

and

$$\mathbb{P}\left(w\left(\bar{Q}_2^n(t),\delta,T\right)>\epsilon\right)\leq\frac{\eta}{5}+\frac{\eta}{5}+\frac{\eta}{5}+0+\frac{\eta}{5}<\eta$$

for large enough n. Following a similar procedure, we can control the oscillations of $\bar{Z}_1^n(t)$ and $\bar{Z}_2^n(t)$, which implies (36) and condition (6.4) are satisfied. By Theorem 11.6.3 of Whitt (2002), $\{\bar{X}^n(\cdot)\}$ is relatively compact.

Step 3: The Limiting Process. Since $\{\bar{X}^n(\cdot), \nu^n\}$ is relatively compact, there exists a convergent subsequence whose limit is denoted by $\{x(\cdot), \nu\}$. Then, by the continuous mapping theorem, the subsequence satisfies

$$z_1(t) = z_1(0) + \lambda_1 \nu([0,t] \times \bar{\mathbb{Z}}^-) + \int_{[0,t]\times\bar{\mathbb{Z}}^+} [\mu_1 z_1(s) + \mu z_2(s)] \nu(ds \times dy) - \int_0^t \mu_1 z_1(s) ds, \tag{37}$$

$$q_1(t) = q_1(0) + \lambda_1 \nu([0,t] \times \bar{\mathbb{N}}) - \int_{[0,t]\times\bar{\mathbb{Z}}^+} [\mu_1 z_1(s) + \mu z_2(s)] \nu(ds \times dy) - \int_0^t \mu_1 z_1(s) ds, \tag{38}$$

$$\begin{aligned} z_2(t) = z_2(0) &+ \int_{[0,t]\times\bar{\mathbb{Z}}^-} [\lambda_2 + \theta q_2(s)] \nu(ds \times dy) - \int_{[0,t]\times\bar{\mathbb{Z}}^+} (1-p)\mu z_2(s) \nu(ds \times dy) \\ &- \int_0^t p\mu z_2(s) ds, \end{aligned} \tag{39}$$

$$\begin{aligned} q_2(t) = q_2(0) &+ \int_{[0,t]\times\bar{\mathbb{N}}} (1-\phi)[\lambda_2 + \theta q_2(s)] \nu(ds \times dy) + \int_{[0,t]\times\bar{\mathbb{Z}}^+} (1-\phi)(1-p)\mu z_2(s) \nu(ds \times dy) \\ &- \int_0^t \theta q_2(s) \nu(ds \times dy), \end{aligned} \tag{40}$$

where $\bar{\mathbb{N}} = \{0, 1, 2, \cdots, +\infty\}$, $\bar{\mathbb{Z}}^+ = \{1, 2, \cdots, +\infty\}$ and $\bar{\mathbb{Z}}^- = \{-1, -2, \cdots, -\infty\}$.

Kurtz (1992) shows in Lemma 1.4 that the limit measure $\nu(\cdot)$ can be separated into a product form. That is, for any Borel set $\Gamma_1 \subset [0, \infty)$ and $\Gamma_2 \subset \bar{\mathbb{Z}}$,

$$\nu(\Gamma_1 \times \Gamma_2) = \int_{\Gamma_1} \pi_s(\Gamma_2) ds, \tag{41}$$

where π_s is a probability measure on $\bar{\mathbb{Z}}$. Next, we complete the proof of Theorem 2 by deriving the expression of $\pi_s(\cdot)$. Let $\{m(\cdot|x) : x = (z_1, z_2, q_2) \in \mathbb{R}_+^3\}$ be a family of continuous-time Markov chains with transition rates dependent on x as follows:

$$m(\cdot|x) \to \begin{cases} m(\cdot|x) + 1, \text{ at the rate } \mathbf{1}_{\{m(\cdot|x)<0\}}(\lambda_1 + \lambda_2 + \theta q_2) + \mathbf{1}_{\{m(\cdot|x)\geq 0\}}\lambda_1, \\ m(\cdot|x) - 1, \text{ at the rate } \mathbf{1}_{\{m(\cdot|x)\leq 0\}}(\mu_1 z_1 + p\mu z_2) + \mathbf{1}_{\{m(\cdot|x)>0\}}(\mu_1 z_1 + \mu z_2). \end{cases}$$

We now show that π_s is the stationary distribution of $m(\cdot|x(s))$ for $s \in (0, \infty)$.

For any bounded continuous function f on $\bar{\mathbb{Z}}$,

$$\begin{aligned} \frac{f(m^n(t))}{n} =& \frac{f(m^n(0))}{n} + \int_0^t [f(m^n(s)+1) - f(m^n(s))] \left\{ d\bar{M}_{\Lambda,1}^n(s) + \mathbf{1}_{\{m^n(s)<0\}} d[\bar{M}_{\Lambda,2}^n(s) + \bar{M}_{C,2}^n(s)] \right\} \\ &+ \int_0^t [f(m^n(s)-1) - f(m^n(s))] \left\{ \bar{M}_{S,1}^n(s) + \mathbf{1}_{\{m^n(s)\leq 0\}} d\bar{M}_{S,2}^n(s) + \mathbf{1}_{\{m^n(s)>0\}} d\bar{M}_{D,2}^n \right\} \\ &+ \int_0^t [f(m^n(s)+1) - f(m^n(s))] \left\{ \bar{\lambda}_1^n + \mathbf{1}_{\{m^n(s)<0\}} [\bar{\lambda}_2^n + \theta \bar{Q}_2^n(s)] \right\} ds \\ &+ \int_0^t [f(m^n(s)-1) - f(m^n(s))] \left\{ \mu_1 \bar{Z}_1^n(s) + \mathbf{1}_{\{m^n(s)\leq 0\}} p\mu \bar{Z}_2^n(s) + \mathbf{1}_{\{m^n(s)>0\}} \mu \bar{Z}_2^n(s) \right\} ds. \end{aligned}$$

As $n \to \infty$, the martingale parts (the second and third terms) converge to zero by Doob's inequality and $\frac{f(m^n(t))-f(m^n(0))}{n} \to 0$ since f is bounded. Therefore, the sum of the last two terms should also converge to zero, which, by the continuous mapping theorem, leads to

$$\int_{[0,t]\times\bar{\mathbb{Z}}} [f(y+1)-f(y)]\left\{\mathbf{1}_{\{y<0\}}[\lambda_1+\lambda_2+\theta q_2(s)]+\mathbf{1}_{\{y\geq 0\}}\lambda_1\right\}\nu(ds\times dy)$$
$$+\int_{[0,t]\times\bar{\mathbb{Z}}} [f(y-1)-f(y)]\left\{\mathbf{1}_{\{y\leq 0\}}[\mu_1 z_1(s)+p\mu z_2(s)]+\mathbf{1}_{\{y>0\}}[\mu_1 z_1(s)+\mu z_2(s)]\right\}\nu(ds\times dy)=0,$$

for any t. Hence, by (41),

$$\int_{\bar{\mathbb{Z}}} [f(y+1)-f(y)]\left\{\mathbf{1}_{\{y<0\}}[\lambda_1+\lambda_2+\theta q_2(s)]+\mathbf{1}_{\{y\geq 0\}}\lambda_1\right\}$$
$$+[f(y-1)-f(y)]\left\{\mathbf{1}_{\{y\leq 0\}}[\mu_1 z_1(s)+p\mu z_2(s)]+\mathbf{1}_{\{y>0\}}[\mu_1 z_1(s)+\mu z_2(s)]\right\}\pi_s(dy)=0$$

for almost all s and it follows from Proposition 4.9.2 of Ethier and Kurtz (1986) that π_s is the stationary (invariant) measure for $m(\cdot|x(s))$. Thus, the steady-state probability can be obtained as follows:

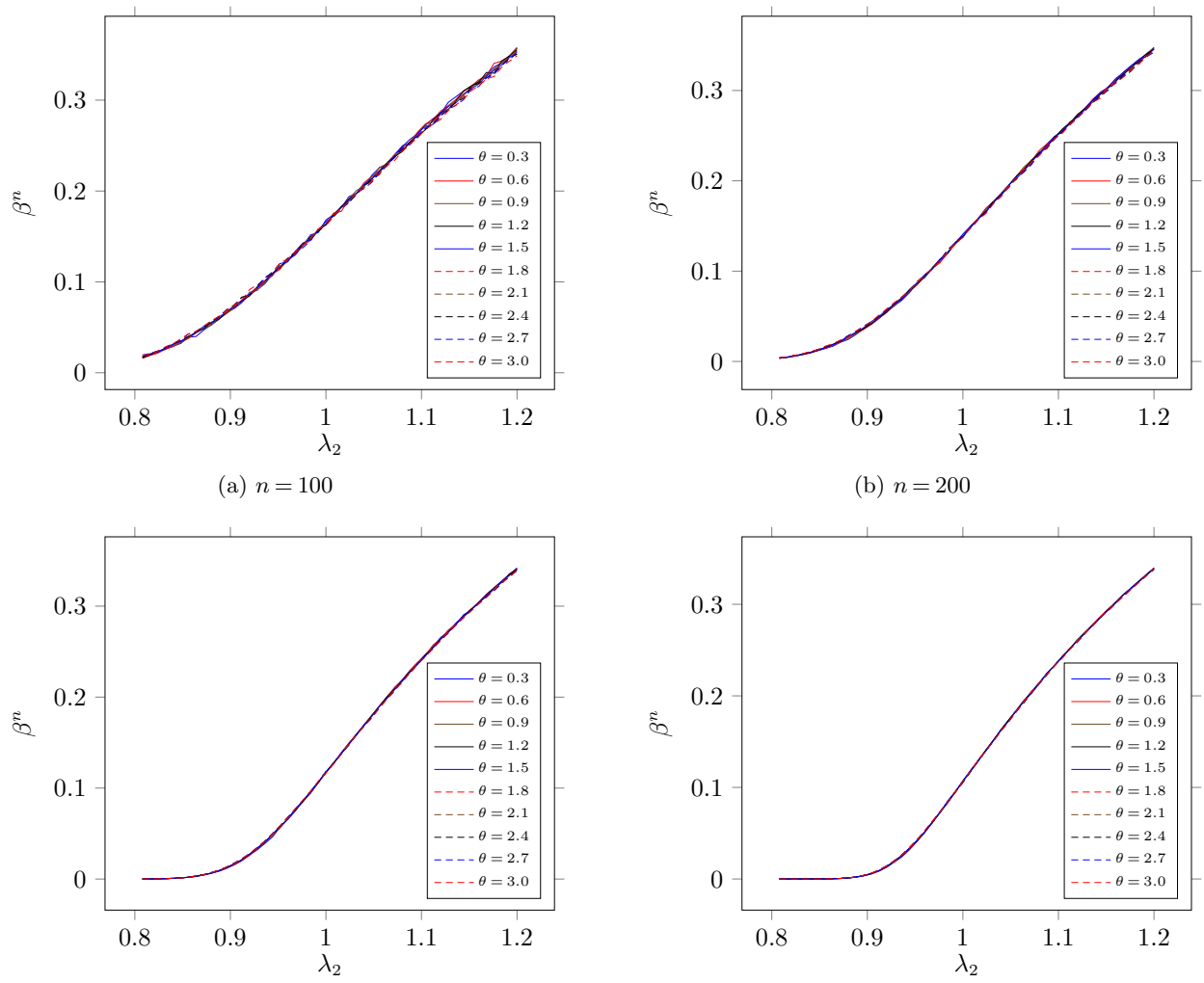
- For $q_1(s)>0$, $m(\cdot|x(s))=\infty$, $\pi_s(\bar{\mathbb{N}})=\pi_s(\bar{\mathbb{Z}}^+)=1$ and $\pi_s(\bar{\mathbb{Z}}^-)=0$.
- For $z_1(s)+z_2(s)<1$, $m(\cdot|x(s))=-\infty$, $\pi_s(\bar{\mathbb{N}})=\pi_s(\bar{\mathbb{Z}}^+)=0$ and $\pi_s(\bar{\mathbb{Z}}^-)=1$.
- For $q_1(s)=0$ and $z_1(s)+z_2(s)=1$, $\pi_s(\bar{\mathbb{N}}) = \min\left\{\left(\frac{[\lambda_1+\lambda_2+\theta q_2(s)-\mu_1 z_1(s)-p\mu z_2(s)][\mu_1 z_1(s)+\mu z_2(s)]}{[\lambda_1+\lambda_2+\theta q_2(s)][\mu_1 z_1(s)+\mu z_2(s)]-\lambda_1[\mu_1 z_1(s)+p\mu z_2(s)]}\right)^+, 1\right\}$, $\pi_s(\bar{\mathbb{Z}}^-) = 1-\pi_s(\bar{\mathbb{N}})$ and $\pi_s(\bar{\mathbb{Z}}^+) = \min\left\{\frac{\lambda_1}{\mu_1 z_1(s)+\mu z_2(s)}\pi_s(\bar{\mathbb{N}}), 1\right\}$.

By (35) and (41), $\pi_s(\bar{\mathbb{N}}) = \lim_{n\to\infty}\mathbb{P}(I^n(s)=0)$ and $\pi_s(\bar{\mathbb{Z}}^+) = \lim_{n\to\infty}\mathbb{P}(Q_1^n(s)>0)$. If we let $\beta(s):=\pi_s(\bar{\mathbb{N}})$ and $\alpha(s):=\pi_s(\bar{\mathbb{Z}}^+)$, then $\beta(s)$ and $\alpha(s)$ represent the instantaneous probability that an arriving licensed user is delayed and the probability that an unlicensed user has to release the channel after a service session, respectively. By substituting them into (37)–(40) and taking the derivative with respect to t, we can easily show that the limit $x(t)$ satisfies the differential equations (4)–(7).

Appendix C: A Numerical Study on the Impact of the Sensing Frequency

We simulate the delay probability and throughput rate for $n \in \{100, 200, 500, 1000\}$, $\lambda_1 \in \{0.02, 0.05\}, \lambda_2 \in [0.8, 1.2]$, $\frac{1}{\mu_t} \in \{\infty, 0.5, 0.25, 0.125\}$. For each combination, the system performance is almost identical when we vary $\theta \in \{0.3, 0.6, \cdots, 3.0\}$. This shows that the performance of unscaled systems is indeed insensitive to the sensing frequency and our fluid limits represent the actual systems accurately.

For illustration purposes, we consider systems with $\frac{1}{\mu_1}=\frac{1}{\mu_2}=1$, $\frac{1}{\mu_s}=0.0001$, $\phi=0.5$ and $\lambda_i^n = n\lambda_i$. We plot the delay probability as a function of λ_2 for (1) different n when $\lambda_1=0.05$ and $\frac{1}{\mu_t}=0.5$ in Figure 9, and (2) different $\frac{1}{\mu_t}$ when $n=500$ and $\lambda_1=0.02$ in Figure 10. As one can see, the delay probability curves are almost identical for $\theta \in \{0.3, 0.6, \cdots, 3.0\}$.



(a) $n = 100$ (b) $n = 200$

(c) $n = 500$ (d) $n = 1000$

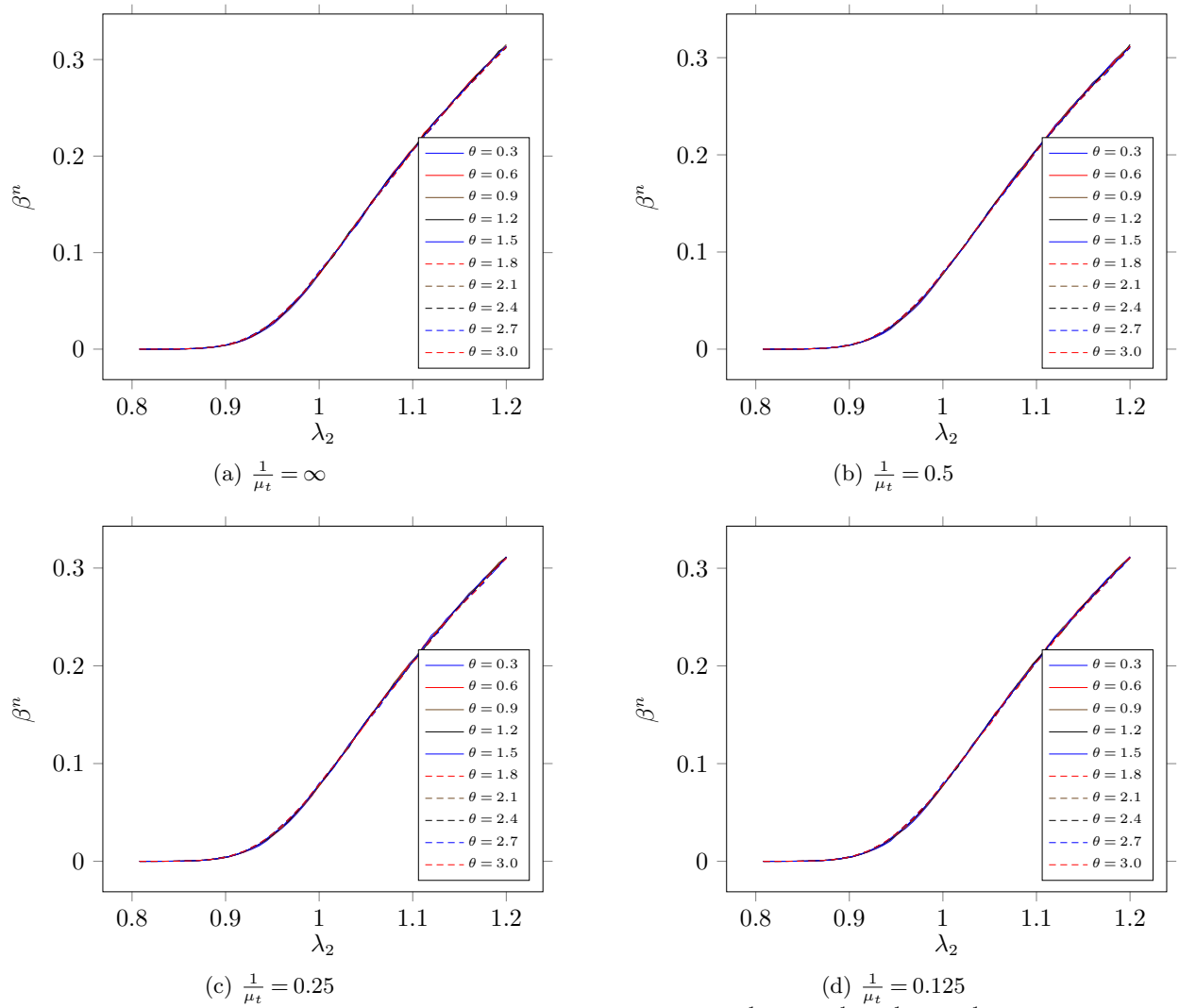
Figure 9 The delay probability as a function of λ_2 for different θ and n when $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 1$, $\frac{1}{\mu_s} = 0.0001$, $\phi = 0.5$, $\lambda_i^n = n\lambda_i$, $\lambda_1 = 0.05$ and $\frac{1}{\mu_t} = 0.5$

References for Appendices

Ethier, S. N. and T. G. Kurtz (1986). *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

Filippov, A. F. (1988). *Differential Equations with Discontinuous Righthand Sides*, Volume 18 of *Mathematics and Its Applications*. Springer Netherlands.

Teschl, G. (2012). *Ordinary Differential Equations and Dynamical Systems*. Graduate studies in mathematics. American Mathematical Society.



(a) $\frac{1}{\mu_t} = \infty$ (b) $\frac{1}{\mu_t} = 0.5$

(c) $\frac{1}{\mu_t} = 0.25$ (d) $\frac{1}{\mu_t} = 0.125$

Figure 10 The delay probability as a function of λ_2 for different θ and $\frac{1}{\mu_t}$ when $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 1$, $\frac{1}{\mu_s} = 0.0001$, $\phi = 0.5$, $\lambda_i^n = n\lambda_i$, $n = 500$ and $\lambda_1 = 0.02$