OPERATIONS WITH SUBSCRIBERS

A DISSERTATION
SUBMITTED TO THE GRADUATE SCHOOL OF BUSINESS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ramandeep Singh Randhawa
August 2006

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Sunil Kumar (Principal Advisor)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

J. Michael Harrison

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Peter W. Glynn

Approved for the University Committee on Graduate Studies.

Preface

This thesis studies a monopolistic firm that offers reusable products, or a service, to price and quality-of-service sensitive customers – a rental firm can be thought of as the canonical example. Customers' perception of quality is determined by their likelihood of obtaining the product or service immediately upon request. We model the rental firm as a loss system with multiple servers used by pay-per-use customers, as well as by a pool of subscribers. Though a Poisson process is widely used to model pay-per-use customers, it is inappropriate to model the subscribers, especially as a non-negligible fraction of the subscriber pool could be renters at any time. We propose a Markovian On-Off-Hold model of subscriber requests for the product. An added benefit of the proposed request model is that retrials by subscribers denied service are implicitly taken into account. We analyze the system when the load on the system, and consequently, the number of servers are large. We obtain diffusion approximations in this asymptotic regime, which we use to compute an estimate for the corresponding invariant distribution. Though the limiting diffusion process depends on the Markovian assumptions for the subscribers, the invariant distribution is insensitive to the distributional assumptions, as long as the Off and Hold times are distributed identically, though independently.

We then study the firms' alternatives of offering either a subscription option or a pay-per-use option from a profit-maximizing perspective. In a large market setting, under the assumption of exponential demand, using the diffusion approximations obtained earlier we show that using the subscription option is more profitable for the firm. Further, via a numerical study, we show that this assumption is not essential for the result to hold. However, we show that it is not necessarily true that the

subscription option dominates the pay-per-use option on quality-of-service. The firm is able to manage the trade-off between price and quality-of-service better in the subscription option. Moreover, we show that the social welfare and the consumer surplus can also be higher in the subscription option, indicating that both the firm and the consumers can benefit from the subscription option.

Acknowledgments

As I sat down to write my thesis, a wave of nostalgia got the better of me. I realized that I have been at Stanford for *five* years and that these years have just breezed by. I can honestly say that I have had a great time at Stanford and thoroughly enjoyed (almost) every moment of it. It gives me immense pleasure to acknowledge all those who helped make my stay and education at Stanford so enjoyable.

I would like to begin by thanking my parents for their unwavering love, support and blessings. I could not have completed this thesis without you. Thank you!

A lot of credit for an enjoyable PhD goes to the advisor. Sunil Kumar has been absolutely great. He has been a wonderful mentor and role-model. His acumen and willingness to discuss mathematical issues at any depth are truly remarkable. He has always been there when I needed advice, be it research or any other matter.

I would like to thank Nick Bambos, Darrell Duffie, Christian Gromoll, and Larry Wein for graciously agreeing to be on my Orals Committee. It was a great pleasure to defend in front of them. Many thanks are due to Peter Glynn for teaching me so much in the areas of stochastic processes, stochastic calculus, and simulation. I will always be thankful to Mike Harrison for his help with my presentations.

I will always be thankful to Kiran Seth, for introducing me to the field of *Operations Research*, and to Sandeep Juneja for helping me enormously with my first steps on the 'research road'. Had it not been for these two professors, I may not even have come to Stanford. They made me visualize my goal in life and showed me the path towards its fulfillment.

Having amazing friends at Stanford enhanced the experience a great deal. I will always cherish the many laughs I have had with Tomer Yahalom, the pondering on

research issues over games of pool with Achal Bassamboo, the exciting drives with Gautam Tandon. Having two great friends as room mates for the initial part of my stay at Stanford, Ankur Jain and Parag Jain, made my transition to the US a lot easier. I was fortunate to have my Uncle and Aunt, Ivninder and Jagdeep Sidhu, in the neighborhood. Their love and warmth always made me feel at home, even in this distant land.

I am thankful to Stanford University for giving me the opportunity to meet my fiancée Smrity, whose love and affection have been a pillar of strength for me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many firms provide reusable products or services to customers who generate requests for these products or services in some random fashion. Another source of variability in this system is the duration of use of the product by the customer. A canonical example is a rental firm. The firm cannot always accommodate all such requests because the number of products stocked or the number of available servers is limited. A measure of quality-of-service as seen by a customer in such a firm is the likelihood of obtaining the product immediately upon request. In order to manage the variability in the presence of capacity constraints, the firm needs to set the capacity level, i.e., the number of products stocked, and the price so as to optimally extract profit.

Access to such rental firms is typically of two kinds: subscription or pay-per-use, with either option having an equal ease of implementation. For example, in the case of DVD rentals, Netflix has emerged as a major player and it provides only a subscription option. On the other hand, Blockbuster mostly provides a pay-per-use option. Recently, with the introduction of Freedom Pass, Blockbuster has also made a foray into the subscription option. Of course, one would expect a firm to do better in face of competition by locking in customers via subscription. In this thesis, We study subscription from a different angle. We study whether the firm is able to handle the inherent variability in the system better if it offers a subscription option rather than a pay-per-use option. That is, we study whether there is any *operational* benefit to the firm of offering subscription.

These rental firms can be modeled as multi-server loss systems, where customer requests that are not served on request are lost. The study of loss systems has a long history going back to Erlang (Erlang (1917), Gross and Harris (1998)). Most of these models assume Poisson, or at best renewal, arrival streams, independent of the number of servers being utilized. While this assumption is justifiable for pay-per-use customers, it is not appropriate for modeling subscribers. We build a model that better describers the behavior of subscribers by associating each subscriber with a Markov chain having three states: On, Off, and Hold. A subscriber spends an exponentially distributed amount of time with mean $\frac{1}{\lambda}$ in the Off state, after which she requests for a server. If no servers are available she transits to the Hold state. Otherwise a server is assigned to her and she transits to the On state. We do not allow a server to be assigned to more than one subscriber at any time. In the Hold state, she retries to obtain a server after an exponentially distributed amount of time with mean $\frac{1}{\nu}$ until a server is available. Once a server is assigned to her, she transits to the On state. In the On state, she uses the server for an exponenitally distributed amount of time with mean $\frac{1}{\mu}$, after which she releases the server and transits to the Off state. All the times are independent and identically distributed and independent of each other. Figure 1.1 illustrates the transitions, while a precise formulation using Poisson processes is provided in (2.1-2.2). This subscriber model is related to the classical Engset model (see Kleinrock (1975)) used in telecommunications. In fact, for the special case in which the Hold and Off state are indistinguishable, i.e., when the rates at which a subscriber tries to obtain a server from the Hold and Off states are equal, the subscribers can be characterized by a two state Markov chain; this is identical to the Engset model.

In this thesis, we consider a monopolistic firm serving price and quality sensitive customers. The firm has the option of offering a subscription or a pay-per-use option. We assume that subscribers pay a subscription fee per unit time, independent of the usage, while the pay-per-use customers pay a price each time they use the product. In each case, the firm statically sets price and capacity levels and observes demand in response. The firm's objective is to maximize the difference between the revenues it obtains from the customers and the cost of maintaining the capacity. We assume

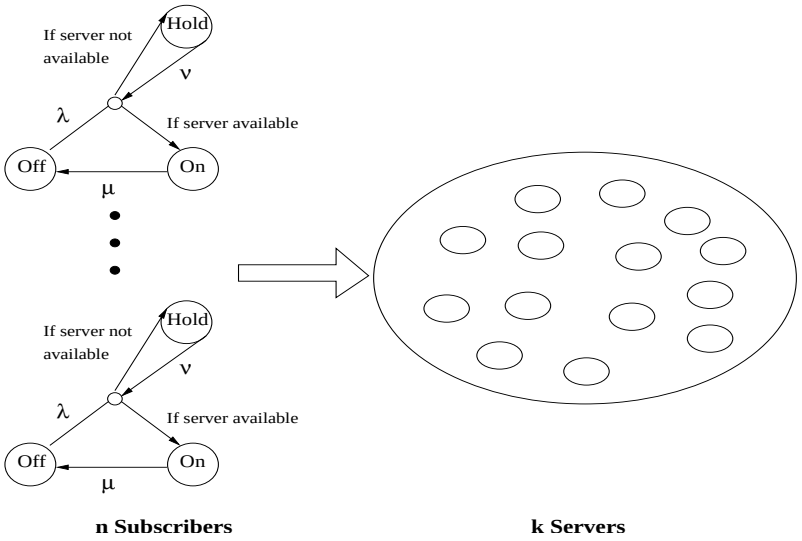


Figure 1.1: The system with n subscribers and k servers.

that the customer demand, that is, the number of subscribers joining the system or the rate at which pay-per-use customers show up at the firm, depends on both the prices the firm sets as well as the equilibrium likelihood of obtaining the product immediately upon request. The reader can think of this equilibrium as having been reached via repeated interactions with the customers, and hence this likelihood is common knowledge in the market. Noting the difficulty of an exact analysis, we use an asymptotic characterization to solve the firm's profit maximization problem. The natural asymptotic regime that we consider is one where the market for potential customers is large. To this end, we asymptotically analyze the general rental system with both subscribers and pay-per-use customers, who are modeled using a Poisson arrival stream. This analysis encompasses both these individual systems and is also of theoretical interest for its novelty. It also facilitates in quantifying the differences between the subscriber model and that with a Poisson arrival stream.

We study the asymptotic behavior of this system as the number of subscribers n grows without bound. We set the rate of the Poisson stream as $\lambda_p = \lambda_1 n + \lambda_2\sqrt{n}$ and choose the number of servers to be of the form $k = k_1 n + k_2\sqrt{n}$ for some k_1 and k_2, akin to Halfin and Whitt (1981). We shall see in the economic analysis that this is indeed the right order of magnitude for the Poisson arrival rate and the number of

servers if our aspiration is to get within $O(\sqrt{n})$ of the optimal profit. When k_1 is set at a level that corresponds to the nominal utilization of the system, we derive diffusion limits for the underlying system. In particular, we prove that a scaled and centered version of the number-in-system process converges to a reflected affine-drift diffusion process. When k_1 is set at a higher level, the limits obtained are similar to those for the critically loaded case, except that asymptotically this system does not behave as a loss system, i.e., the resulting diffusion is an affine-drift diffusion process without reflection, and thus we do not treat this case in detail. (The asymptotic results for this system are quite similar to those in Mandelbaum and Pats (1998).) The limits obtained when k_1 is lower than the critical level are uninformative, as well as trivial. For the critically loaded case, we extend the process level convergence to steady-state convergence by proving the convergence of the corresponding invariant distribution. When the Off and Hold states are indistinguishable for the subscribers, we are able to completely characterize the limiting invariant distribution as a truncated two-dimension Gaussian distribution.

We use the asymptotic limits thus derived to perform a comparison between the subscription and pay-per-use options. In comparing the two options, the first question that one might ask is the following. Is it not true that locking customers into the subscription option leads to a reduction in the inherent system variability, resulting in higher quality-of-service at any choice of capacity or price? We show that this is not true. In fact, neither of these options dominates the other in terms of quality-of-service. This motivates us to compare profits in the two options. For the case of exponential demand we prove that the subscription option is indeed better for the firm. Through a numerical study we show that this result is robust with respect to the choice of demand function. Although the subscription option does not necessarily dominate the pay-per-use option on quality-of-service, the firm is able to manage the trade-off between price and quality-of-service better in the subscription option. Moreover, we show that the social welfare and the consumer surplus can also be higher in the subscription option, indicating that both the firm and the consumers can benefit from the subscription option. However this need not always be the case, with the firm sometimes profiting from the subscription option at the customers'

expense.

Organization. Chapter 2 develops the basic subscriber model and its asymptotic limits, and then compares the subscriber model with an exogenous arrival model. Chapter 3 contains the main theoretical results of this thesis. It provides the asymptotic limits for a general rental system that has subscribers and pay-per-use customers. Chapter 4 uses the asymptotic results derived thus far to perform an economic analysis that demonstrates that the subscriptions option is operationally better than the pay-per-use option for a firm. Finally, Chapter 5 discusses some points that are not addressed in the rest of the thesis, and are worth not closing without.

1.1 Literature Review

The literature for limit theory of closed queueing systems is relatively small with single server stations studied in Harrison and Williams (1996), and Kumar (2000). Krichagina and Puhalskii (1997) and Kogan, Lipster and Smorodinskii (1986) study queueing systems that allow stations to have state-dependent service rates, but infinite buffers. A very recent related paper is de Véricourt and Jennings (2005) that develops limit theory for a multi-server queueing system serving a pool of subscribers. This work models the system as a queue, where requests which are not accepted immediately are queued. Using the probability of queueing as the relevant performance metric, the authors use the limit to size capacity. Critically loaded loss systems with customers arriving as a Poisson process are a subject of study in papers such as Hunt and Kelly (1989), Reiman (1991), and Puhalskii and Reiman (1998).

In Chapter 4, we use the approach in Mendelson and Whang (1990) to build a micro-economic framework around our service system and study optimal pricing and capacity sizing at the system equilibrium. We characterize the system equilibrium as in Armony and Maglaras (2004) and Whitt (2003). These papers deal with multi-server queueing systems with congestion sensitive demand but without economic considerations and study the system equilibrium behavior using asymptotic methods. Perhaps the most related to our work is a recent paper Maglaras and Zeevi

(2003) that studies pricing and capacity selection in a multi-server queueing system with Poisson arrivals. This paper uses the Halfin-Whitt (see Halfin and Whitt (1981)) asymptotic results to develop estimates for congestion levels to compute optimal price levels. Another application of optimal pricing and capacity sizing in an asymptotic regime is in Plambeck and Ward (2005), where the authors study static pricing, capacity selection and dynamic scheduling using the traditional heavy traffic assumptions in an assemble-to-order setting. Optimal capacity sizing under fixed, exogenous demand is studied in Borst, Mandelbaum and Reiman (2004), where the authors use asymptotic methods to compute the optimal staffing level in a call center with the trade-off being between the agents' cost and the quality of service provided.

Though we only deal with static pricing, work in Paschalidis and Tsitsiklis (2000) and Gallego and van Ryzin (1994) suggests that this is not a major limitation. Paschalidis and Tsitsiklis (2000) studies revenue management in the context of Internet service provision using a Markov decision process framework where customers are only assumed to be price sensitive. Although they focus on dynamic pricing, an important conclusion of their study is that static pricing rules can achieve near optimal performance. Gallego and van Ryzin (1994) derives a similar insight in the "classical" context of selling a set of goods within a finite time horizon.

1.2 Mathematical Preliminaries

All the random quantities in this thesis are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In this thesis, we shall consider stochastic processes that lie in $D_{\mathbb{R}^m}[0,\infty)$, the space of right continuous functions having left limits with values in $\mathbb{R}^m$. Although the norm used on this space is the Skorohod J_1, we will restrict ourselves to the uniform norm because all the limits we consider will have continuous sample paths almost surely (a.s.). For $X \in D_{\mathbb{R}^m}[0,\infty)$ and for $T \geq 0$, we denote $\| X \|_T \equiv \sup_{t \leq T} \max_{i=1,\ldots,m} |X_i^n(t)|$. We shall use the convention that for any function $f : \mathbb{R}_+ \to \mathbb{R}_+$, $\| X(f(\cdot)) \|_T = \sup_{t \leq T} \max_{i=1,2,3} |X_i(f(t))|$.

We shall say that for random elements of $D_{\mathbb{R}^m}[o,\infty)$, (X^n, X), $X^n \to X$ if $\| X^n - X \|_T \to 0$, a.s. for each $T > 0$. For a collection of probability measures

P_n and P defined on $(S, \mathcal{S})$, where S is a general metric space and $\mathcal{S}$ its Borel σ-field, we say that $P_n \Rightarrow P$, i.e., P_n weakly converges to P, if and only if $\int_S f dP_n \to \int_S f dP$ for all bounded, continuous real-valued functions on S. Further, if X_n and X are random variables defined on this space such that P_n and P are the distributions of X_n and X respectively, then $P_n \Rightarrow P$ is equivalent to $X_n \Rightarrow X$. Note that unless stated otherwise, all convergence results in this thesis take place as the index $n \to \infty$.

For any given stochastic process $X \in D_{\mathbb{R}^m}[0, \infty)$ and any measurable function $f : \mathbb{R}^m \to \mathbb{R}$, $\mathbb{E}_x f(X(t)) \equiv \mathbb{E}[f(X(t))|X(0) = x]$, and if π denotes a probability measure, then $\mathbb{E}_\pi f(X(t))$ refers to the expectation of $f(X(t))$ conditioned on $X(0)$ being distributed according to π.

We use the following strong approximation result from Kurtz (1978), which follows from Komlós, Major and Tusnady (1975), and will be used extensively in the proofs.

Lemma 1. *(Kurtz (1978), Lemma 3.1) A standard (rate 1) Poisson process $N(t)$ can be realized on the same probability space as a standard Brownian motion $B(t)$ in such a way that the positive random variable X given by*

$$X \equiv \sup_{t \geq 0} \frac{|N(t) - t - B(t)|}{\log(2 \vee t)} \tag{1.1}$$

satisfies $\mathbb{E}e^{\theta X} < \infty$, $\forall \theta > 0$ sufficiently small. In particular, $\mathbb{E}X < \infty$.

For any given Poisson process N, $\bar{N}$ denotes its centered version, namely $\bar{N}(t) = N(t) - t$, $t \geq 0$. $C^{i,j,k}$ is the space of functions defined on $\mathbb{R}^3$ that are i, j and k times continuously differentiable with respect to the first, second and third argument respectively. We use the convention that for a vector $a \in \mathbb{R}^n$, a_i represents its i^{th} component and a' denotes its transpose. Finally, we say $f(n) = O(g(n))$ if there are positive constants c and k, such that $0 \leq f(n) \leq cg(n)$ for all $n \geq k$.

Chapter 2

The Subscriber Model

This chapter introduces the subscriber model and analyzes it asymptotically as the number of subscribers grows without bound. We shall observe that the asymptotic limits are as tractable as those of a system with exogenous arrivals according to a Poisson stream, though quite different. Most of the results in this chapter are special cases of results in the following chapter and are stated here largely for the reader to get a feel for the asymptotic regime, as well as a sense of the differences in the two models.

This chapter is organized as follows: Section 2.1 deals with the subscriber model with general retrial rates, charactering the asymptotic limits, Section 2.2 focuses on the exogenous arrival model, and finally Section 2.3 performs a comparison of both the models. Appendix A contains the proofs of results in this chapter.

2.1 Subscriber Model with Retrials

We begin by introducing the subscriber model and carrying out the asymptotic analysis in the regime where the number of subscribers and the number of servers increase without bound.

We consider a system with n subscribers and k servers, where a server cannot be assigned to more than one subscriber at any time. Each subscriber has an underlying continuous time Markov chain $J(\cdot)$, where $J(t) \in \{\text{On, Off, Hold}\}$ for $t \geq 0$. A

subscriber spends an exponentially distributed amount of time with mean $\frac{1}{\lambda}$ in the Off state, after which she requests for a server. If no servers are available she transits to the Hold state, otherwise a server is assigned to her and she transits to the On state. In the Hold state, she retries to obtain a server after every exponentially distributed amount of time with mean $\frac{1}{\nu}$ until a server is available. Once a server is assigned to her, she transits to the On state. In the On state, she uses the server for an exponenitally distributed amount of time with mean $\frac{1}{\mu}$, after which she returns the server and transits to the Off state. We assume that all the times are independent and identically distributed and independent of each other.

This system is a special case of that discussed in Chapter 3, where in addition to the subscribers, an external arrival stream of customers is also present. However, the novelty of this subscriber model and the interesting nature of its diffusion limits motivate us to present the asymptotic results for this special case before doing so for the general system.

We shall begin by introducing some notation, let $Q^n \in D_{\mathbb{R}^2}[0,\infty)$ such that $Q_1^n(t)$ and $Q_2^n(t)$ represent the number of servers in use at time t and the number of subscribers in the Hold state respectively, when there are a total of n subscribers, i.e., $Q_1^n(t) = \sum_{i=1}^n 1_{\{\text{Subscriber } i \text{ in On state at } t\}}$ and $Q_2^n(t) = \sum_{i=1}^n 1_{\{\text{Subscriber } i \text{ in Hold state at } t\}}$. A precise definition of our system is as follows [1]:

$$\begin{aligned} Q_1^n(t) = Q_1^n(0) &+ N^a\left(\int_0^t 1_{\{Q_1^n(u)<k\}}(n - Q_1^n(u) - Q_2^n(u))\lambda du\right) \\ &+ N^r\left(\int_0^t 1_{\{Q_1^n(u)<k\}}\nu Q_2^n(u)du\right) - N^d\left(\int_0^t \mu Q_1^n(u)du\right), \end{aligned} \tag{2.1}$$

$$\begin{aligned} Q_2^n(t) = Q_2^n(0) &+ N^a\left(\int_0^t (n - Q_1^n(u) - Q_2^n(u))\lambda du\right) \\ &- N^a\left(\int_0^t 1_{\{Q_1^n(u)<k\}}(n - Q_1^n(u) - Q_2^n(u))\lambda du\right) \\ &- N^r\left(\int_0^t 1_{\{Q_1^n(u)<k\}}\nu Q_2^n(u)du\right), \end{aligned} \tag{2.2}$$

for $t \geq 0$, where $N^a(\cdot)$, $N^d(\cdot)$ and $N^r(\cdot)$ are three independent, one-dimensional unit rate Poisson processes.

As we are interested in asymptotic results, we consider capacity levels of the form $k^n = k_1 n + k_2\sqrt{n}$ for some $k_1 \in \mathbb{R}_+$ and $k_2 \in \mathbb{R}$. Define

$$m = \frac{\lambda\mu}{\lambda+\mu}$$

and the centered and scaled process

$$\hat{Q}^n(\cdot) = \frac{Q^n(\cdot) - (k^n, 0)'}{\sqrt{n}} \leq 0.$$

Further defining $q^n(\cdot) = \frac{Q^n(\cdot)}{n}$ and $\bar{q}(\cdot) = (\bar{q}_1(\cdot), \bar{q}_2(\cdot))'$ with $\bar{q}_1(\cdot) \equiv \min\left(k_1, \frac{\lambda}{\lambda+\mu}\right)$ and $\bar{q}_2(\cdot) \equiv \frac{\lambda}{\lambda+\mu} - \bar{q}_1(\cdot)$, we have the following asymptotic results for this system.

Proposition 1. *If $q^n(0) \to \bar{q}(0)$ a.s., then*

(a) $q^n \to \bar{q}$.

(b) If $k_1 = \frac{\lambda}{\lambda+\mu}$ and $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$, then $\hat{Q}^n \Rightarrow \hat{Q}$, where

$$\begin{aligned}\hat{Q}_1(t) = \hat{Q}_1(0) &- \int_0^t \Big((\lambda+\mu)(\hat{Q}_1(u) + k_2) + (\lambda-\nu)\hat{Q}_2(u)\Big)\, du \\ &+ \sqrt{2m}B(t) - Y(t)\end{aligned} \tag{2.3}$$

$$\hat{Q}_2(t) = \hat{Q}_2(0) - \int_0^t \nu\hat{Q}_2(u)du + Y(t), \tag{2.4}$$

where B is a standard Brownian motion and Y is the non-negative, non decreasing process such that $\int_0^t \hat{Q}_1(u)dY(u) = 0$, $\forall t \geq 0$, and $Y(0) = 0$.

(c) The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \Rightarrow \hat{\pi}$, where $\hat{\pi}$ is the unique invariant distribution of the diffusion process given by (2.3-2.4).

This result is a special case of Lemma 4 and Theorem 1 in Section 3.1, and hence its proof is omitted. All the asymptotic results in the thesis are structured in a manner similar to Proposition 1. We first prove the fluid level convergence as in (a), followed by weak convergence to diffusion limits as in (b). In doing so we first establish that the equations describing the diffusion limit (2.3-2.4) have a unique

strong solution. We also establish that the limiting diffusions have unique stationary distributions and establish the convergence of the invariant distributions of $\hat{Q}^n$ to the invariant distribution of the limiting diffusions as in (c). Finally, where possible, we characterize this limiting invariant distribution.

The diffusion process given by (2.3-2.4) is quite interesting. Note that there is a single one-dimensional Brownian motion driving this diffusion and the only stochasticity in the process $\hat{Q}_2(\cdot)$ arises through the local time of this Brownian motion at the boundary. This process is fairly complicated, and it renders a further investigation into its invariant distribution futile. (The reader is directed to Section 3.1 page 21 for an illustration of the difficulties that arise in the computation of the invariant distribution of this process.) However, we shall see that for the case $\nu = \lambda$, where the Off and Hold states are indistinguishable, we obtain a far more tractable process for which we will be able to compute the invariant distribution.

2.1.1 Special case $\nu = \lambda$

When $\nu = \lambda$, the Off and Hold state are indistinguishable, which implies that we do not need to keep track of the number of subscribers in the Hold state. This allows us to focus on a one dimensional process $Q^n(\cdot) \in D_{\mathbb{R}}[0, \infty)$ that refers to the number of servers in use by the subscribers. We obtain the following characterization of $Q^n(\cdot)$.

$$Q^n(t) = Q^n(0) + N^a\left(\int_0^t 1_{\{Q^n(u)<k^n\}}(n - Q^n(u))\lambda\, du\right) - N^d\left(\int_0^t \mu Q^n(u)\, du\right).$$

Defining $\bar{q}(\cdot) = \min\left(k_1, \frac{\lambda}{\lambda+\mu}\right)$, we have the following asymptotic results for this system.

Proposition 2. *If $q^n(0) \to \bar{q}(0)$ a.s., then*

(a) $q^n \to \bar{q}$.

(b) If $k_1 = \frac{\lambda}{\lambda+\mu}$ and $\hat{Q}^n(0) \Rightarrow \hat{Q}^n(0)$, then $\hat{Q}^n \Rightarrow \hat{Q}$, where $\hat{Q}(\cdot)$ is a reflected

affine-drift diffusion process with an upper reflecting barrier at 0. *That is,*

$$\hat{Q}(t) = \hat{Q}(0) - (\lambda + \mu) \int_0^t (\hat{Q}(t) + k_2)dt + \sqrt{2m}\ B(t) - Y(t), \tag{2.5}$$

where B *is a standard Brownian motion, and* Y *is the non-negative, non-decreasing process such that* $\int_0^t \hat{Q}(u)dY(u) = 0$, $\forall t \geq 0$ *and* $Y(0) = 0$.

(c) The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \to \hat{\pi}$, *where* $\hat{\pi}$ *is the unique invariant distribution of the diffusion process given by* (2.5).

The limiting invariant distribution can be characterized using Proposition 1 in Ward and Glynn (2003) as follows.

Proposition 3. *The density corresponding to the invariant distribution of the diffusion process given by* (2.5) *is*

$$\hat{p}(x) = \frac{\exp\left(-\frac{1}{2m}(\lambda+\mu)(x+k_2)^2\right)}{\int_{-\infty}^0 \exp\left(-\frac{1}{2m}(\lambda+\mu)(z+k_2)^2\right) dz}, \quad x \leq 0.$$

At this point, we shall take a step back and try to better understand the differences between the subscriber model and one in which the customers arrive exogenously according to a Poisson process. To do so, we build a diffusion model for the system with customers arriving according to a Poisson process and compare the "attempt" processes, that is, the processes according to which customers attempt to obtain service, and the denial rates, that is, the steady-state rate at which customers' attempts are denied, in the two systems.

2.2 Exogenous Arrival Model

We consider a system with a large number of customers that arrive according to a Poisson process. If an arriving customer does not find an available server, she leaves the system and is considered lost, otherwise she uses the server for an exponentially distributed time interval, after which she leaves the system and does not return. This is the standard $M/M/k/k$ loss model. As we are interested in a comparison

of this model with the subscriber model, we shall set the arrival rate of this Poisson process such that the nominal loads are equal in the two systems. The nominal load is defined as the number of servers that will be in use in steady-state in the system when the system has infinite capacity. For a subscription based system, this can be computed to be $n\frac{\lambda}{\lambda+\mu}$. Hence, we shall set the arrival rate of the Poisson process at $\Lambda^n = \frac{\lambda}{\lambda+\mu}\mu n + \lambda_2\sqrt{n} = mn + \lambda_2\sqrt{n}$, where $\lambda_2 \in \mathbb{R}$.

As before we consider capacity levels of the form $k^n = k_1 n + k_2\sqrt{n}$ and denote the scaled process by $q^n(\cdot)$, the centered and scaled process by $\hat{Q}^n(\cdot)$. Also, define $\bar{q}(\cdot) = \min\left(k_1, \frac{\lambda}{\lambda+\mu}\right)$. We now obtain the following asymptotic results for this system.

Proposition 4. *If $q^n(0) \to \bar{q}(0)$ a.s., then*

(a) $q^n \to \bar{q}$.

(b) If $k_1 = \frac{\lambda}{\lambda+\mu}$ and $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$, then $\hat{Q}^n \Rightarrow \hat{Q}$, where $\hat{Q}(\cdot)$ is a reflected affine drift diffusion process such that

$$\hat{Q}(t) = \hat{Q}(0) + \lambda_2 t - \mu\int_0^t (\hat{Q}(s) + k_2)ds + \sqrt{2m}B(t) - Y(t) \tag{2.6}$$

where B is a standard Brownian motion, and Y is the non-negative, non-decreasing process such that $\int_0^t \hat{Q}(u)dY(u) = 0$, $\forall t \geq 0$ and $Y(0) = 0$.

(c) The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \to \hat{\pi}$, where $\hat{\pi}$ is the unique invariant distribution of the diffusion process given by (2.6).

The limiting invariant distribution can then be characterized using Proposition 1 in Ward and Glynn (2003) as follows.

Proposition 5. *The density corresponding to the invariant distribution of the diffusion process given by* (2.6) *is*

$$\hat{p}(x) = \frac{\exp\left(-\frac{1}{2}\frac{(x+k_2-\lambda_2/\mu)^2\mu}{\lambda_1}\right)}{\int_{-\infty}^0 \exp\left(-\frac{1}{2}\frac{(z+k_2-\lambda_2/\mu)^2\mu}{\lambda_1}\right)dz}, \quad x \leq 0.$$

We now perform a descriptive comparison of the two models. We shall compare the asymptotic attempt processes and the denial rates in the two systems. If $k_1 = \frac{\lambda}{\lambda+\mu}$, then at the fluid scale the attempt processes and the denial rates in both systems are identical, which motivates us to study whether these similarities exist at a finer level as well.

2.3 Comparison of the Two Models

We shall now compare the attempt processes and the denial rates in the two systems. For convenience, we shall set $\nu = \lambda$ for the subscriber model.

2.3.1 Comparison of attempt processes

The attempt process of the subscribers occurs according to a non-homogeneous Poisson process, i.e. the number of attempts until time t, $A^n(t) = N(\lambda \int_0^t (n - Q^n(s))ds)$, with $N(\cdot)$ being a Poisson process with unit rate. We have the following asymptotic characterization of the attempt process.

Lemma 2. *If f* $k_1 = \frac{\lambda}{\lambda+\mu}$ *and* $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$*, then*

$$\sqrt{n}\left(A^n(\cdot)/n - m\cdot\right) \Rightarrow -\lambda \int_0^{\cdot} (\hat{Q}(s) + k_2)ds + B\left(m\cdot\right), \tag{2.7}$$

where $\hat{Q}(\cdot)$ *is the reflected affine drift diffusion process given by* (2.5)*.*

Denoting the attempt process of the exogenous Poisson arrival process by $\tilde{A}^n(\cdot)$, we get an analog of the above lemma, i.e. $\sqrt{n}\left(\tilde{A}^n(\cdot)/n - m\cdot\right) \Rightarrow B\left(m\cdot\right)$. Comparing this with (2.7), we observe that the attempt process in the proposed model is indeed different from the Poisson arrival process at the diffusion scale, even though they are identical at the fluid scale $\left(\lim_{n\to\infty} \frac{\|A^n(\cdot) - \tilde{A}^n(\cdot)\|_T}{n} = 0\right)$. The state dependence of the drift highlights the key difference in the two attempt processes and also indicates the manner in which retrials are incorporated.

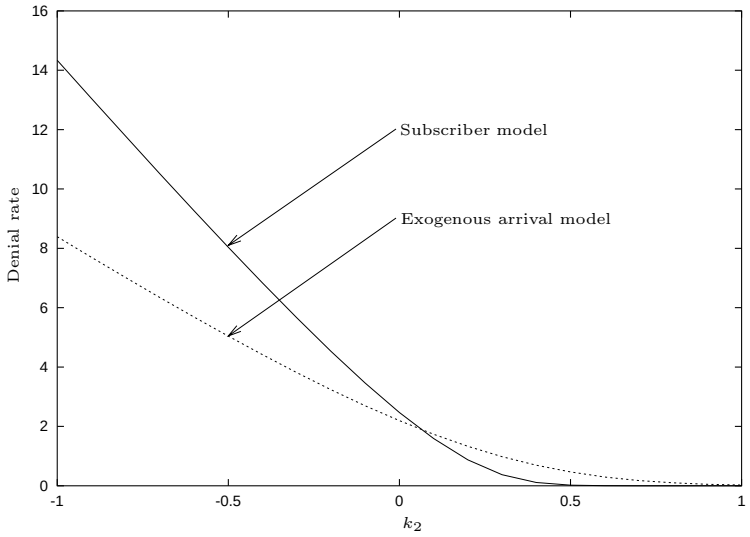


Figure 2.1: Comparison of denial rates in the two models.

2.3.2 Comparison of the denial rates

To further illustrate the difference in these two systems, we compare the limiting denial rates, or the steady-state rate at which customer attempts are denied. We shall perform the comparison for the critically loaded case as the nominal loads in both systems are equal for this case. Specifically, we set the number of servers in each system at $\frac{\lambda}{\lambda+\mu}n + k_2\sqrt{n}$ and then vary k_2. The denial rate as a function of the number of servers, $D^n(k^n)$, can be asymptotically characterized as follows.

Proposition 6. *(a) The denial rate for the system with subscribers converges as*

$$\lim_{n\to\infty} \frac{D^n(k^n)}{\sqrt{n}} = \sqrt{\lambda\mu}\; h\left(-k_2\sqrt{\frac{\lambda+\mu}{m}}\right).$$

(b) The denial rate for the system with exogenous arrivals converges as

$$\lim_{n\to\infty} \frac{D^n(k^n)}{\sqrt{n}} = \sqrt{\mu m}\; h\left(-\left(k_2 - \frac{\lambda_2}{\mu}\right)\sqrt{\frac{\mu}{m}}\right).$$

We shall now choose $\lambda = \mu = 2$ for our computations and compare the limiting denial rates. Figure 2.1 plots the limiting denial rates as a function of k_2. We observe

that if k_2 is small, the denial rate is lower in the exogenous arrival model, however as k_2 increases the subscriber model has a lower denial rate, which demonstrates the difference in the two models. Figure 2.1 can be explained loosely as follows. As the number of subscribers that are in the On state increases, the number of subscribers attempting to obtain a server decreases. Consequently, when the capacity level is high, the denial rates seen by the subscribers is smaller than that seen by the exogenous stream whose attempt rate is independent of the state of the system. When the capacity level is low, the number of subscribers in the On state is small due to server unavailability, and consequently the number of subscribers attempting increases.

Notes

[1]Lemma 3 in Section 3.1 proves the existence of a unique Q that satisfies a relation similar to that in (2.1-2.2). The proof can be adapted to prove the same for this case.

Chapter 3

A General Rental System Model

The focus of this chapter is a general rental system with two types of customers. The first type comprises subscribers, while the other type constitutes a Poisson arrival stream. Though most of the asymptotic results required for the economic analysis in Chapter 4 have been provided in Chapter 2, we shall still study this combined system for its novelty, namely, a loss system with this structure has not been studied in the literature. Further, the asymptotic analysis of this combined system allows us to answer economic questions in more general settings where both subscription and pay-per-use options are offered by the firm (see Section 4.4). We shall explicitly compute the asymptotic limits for this system as the number of servers and the load on the system grow without bound. We shall prove the process level convergence as well as that of the invariant distributions. The main result of this chapter is Theorem 1, which encompasses the results in Chapter 2.

This chapter is organized as follows. In Section 3.1, we develop the general model and state the asymptotic results. Section 3.2 contains a proof of the main diffusion level convergence, while Section 3.3 contains a proof of the convergence of the invariant distribution. All the results that are stated but not proved in this chapter are proved in Appendix B.

3.1 Asymptotic Results

Consider the system with a total of n subscribers as before and an exogenous stream of customers that arrive according to a Poisson process with rate λ_p^n. Let $Q^n \in D_{\mathbb{R}^3}[0,\infty)$, where $Q_1^n(\cdot), Q_2^n(\cdot)$, and $Q_3^n(\cdot)$ are the processes that denote the number of servers in use by the subscribers, the number of servers in use by the exogenous customers, and the number of subscribers in the Hold state respectively. We obtain the following characterization of Q^n.

$$\begin{aligned}
Q_1^n(t) =& Q_1^n(0) + N_1^a\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}(n - Q_1^n(u) - Q_3^n(u))\lambda\, du\right) \\
&- N_1^d\left(\int_0^t \mu Q_1^n(u)\, du\right) + N^r\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\nu Q_3^n(u)\, du\right) \\
Q_2^n(t) =& Q_2^n(0) + N_2^a\left(\lambda_p^n\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\, du\right) - N_2^d\left(\int_0^t \mu Q_2^n(u)\, du\right) \\
Q_3^n(t) =& Q_3^n(0) - N^r\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\nu Q_3^n(u)\, du\right) \\
&+ N_1^a\left(\int_0^t (n - Q_1^n(u) - Q_3^n(u))\lambda\, du\right) \\
&- N_1^a\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}(n - Q_1^n(u) - Q_3^n(u))\lambda\, du\right)
\end{aligned}$$

for $t \geq 0$, where $N_i^a(\cdot)$ for $i = 1, 2$, $N_j^d(\cdot)$ for $j = 1, 2$, and $N^r(\cdot)$ are five independent unit rate Poisson processes. For $(x, y, z) \in \mathbb{R}_+^3$, let $\lambda^n(x, y, z) = ((n - x - z)\lambda,\ \lambda_p^n,\ 0)'$ and $\mu(x, y, z) = (\mu x,\ \mu y,\ 0)'$. Then, using the notation $N(x, y, z) = (N_1(x), N_2(y), N_3(z))'$ and defining $N_3^a(t) = N_3^d(t) = 0$ for $t \geq 0$, the above relation can be re-expressed as

$$\begin{aligned}
Q^n(t) =& Q^n(0) + N^a\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\lambda^n(Q^n(u))du\right) - N^d\left(\int_0^t \mu(Q^n(u))\, du\right) \\
&+ N^r\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\nu Q_3^n(u)\, du\right)(1, 0, -1)' \\
&+ (0, 0, \Delta N_1^a(Q_1^n, Q_2^n, Q_3^n)(t))',
\end{aligned} \tag{3.1}$$

where for $A \in D_{\mathbb{R}}[0,\infty)$,

$$\begin{aligned}\Delta A(Q_1^n, Q_2^n, Q_3^n)(\cdot) \equiv & A\left(\int_0^{\cdot}(n - Q_1^n(u) - Q_3^n(u))\lambda\, du\right) \\ & - A\left(\int_0^{\cdot} 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}(n - Q_1^n(u) - Q_3^n(u))\lambda\, du\right).\end{aligned}$$

The following result proves the existence of a unique Q^n that satisfies (3.1).

Lemma 3. *There exists a unique solution Q^n defined on $(\Omega, \mathcal{F}, \mathbb{P})$ to (3.1).*

Note that (3.1) can be rewritten as

$$\begin{aligned}Q^n(t) = Q^n(0) &+ \int_0^t [\lambda^n(Q^n(u)) + \nu(Q^n(u)) - \mu(Q^n(u))]\, du \\ &+ \bar{N}^a\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\lambda^n(Q^n(u))\, du\right)' - \bar{N}^d\left(\int_0^t \mu(Q^n(u))\, du\right)' \\ &+ \bar{N}^r\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\nu Q_3^n(u)\, du\right)(1,0,-1)' + \left(0,0,\Delta\bar{N}_1^a(Q_1^n, Q_2^n, Q_3^n)\right)' \\ &- \int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)=k^n\}}\tilde{\lambda}^n(Q^n(u))\, du,\end{aligned} \tag{3.2}$$

where for $(x,y,z) \in \mathbb{R}^3_+$, $\tilde{\lambda}^n(x,y,z) = ((n-x-z)\lambda + \nu z,\ \lambda_p^n,\ -(n-x-z)\lambda + \nu z)'$ and $\nu(x,y,z) = (\nu z,\ 0,\ -\nu z)'$.

To obtain meaningful asymptotic limits, we shall choose $\lambda_p^n = \lambda_1 n + \lambda_2\sqrt{n}$, where $\lambda_1, \lambda_2 \geq 0$, and the capacity level $k^n = k_1 n + k_2\sqrt{n}$ as before. Defining

$$q^n(\cdot) = \frac{Q^n(\cdot)}{n} \quad \text{and} \tag{3.3}$$

$$\bar{q}(\cdot) = \left(\frac{\lambda}{\lambda+\mu}, \frac{\lambda_1}{\mu}, 0\right)', \tag{3.4}$$

we state the asymptotic results for this system corresponding to the fluid limit.

Lemma 4. *(a) If $k_1 \geq \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$ and $q^n(0) \to \bar{q}(0)$ a.s., then $q^n \to \bar{q}$.*

(b) If $k_1 < \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$ and $q^n(0) \to \bar{\bar{q}}(0)$ a.s., where $\bar{\bar{q}}_1(\cdot) = \frac{-b-\sqrt{b^2-4ac}}{2a}$ with $a =$

$\nu + \frac{\mu}{\lambda}(\nu - \lambda)$,

$$b = -\left(\nu(k_1 + 1) + \lambda_1 + \frac{\mu}{\lambda}(\nu - \lambda)k_1\right),$$

and $c = k_1\nu$, $\bar{\bar{q}}_2(\cdot) = k_1 - \bar{\bar{q}}_1(\cdot)$ *and* $\bar{\bar{q}}_3(\cdot) = 1 - \frac{\lambda+\mu}{\lambda}\bar{\bar{q}}_1$, *then* $q^n \to \bar{\bar{q}}$.

A point worth noting about this result is that we require $q^n(0)$ to converge to the fluid limit as $n \to \infty$. If this condition does not hold, i.e., if there exists a sequence along which $q^n(0)$ converges to a point different from the fluid limit, then along this sequence $q^n(t)$ shall converge to the fluid limit only as both $n \to \infty$ and $t \to \infty$.

We now wish to study the diffusion limits for this system. To this effect, define $\hat{Q}^n(\cdot) = \sqrt{n}(q^n(\cdot) - \bar{q}(\cdot)) - (k_2, 0, 0)'$. Note that if $k_1 < \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$, the correction process simply does not exists, the limit instantaneously achieves the point $\bar{\bar{q}}$ and stays there. If $k_1 > \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$, this case is identical to an infinite server setting and is treated in Mandelbaum and Pats (1998). We now state the diffusion result for the case $k_1 = \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$.

Theorem 1. *(a) If* $k_1 = \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$, $q^n(0) \to \bar{q}(0)$ *a.s., and* $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$, *then* $\hat{Q}^n \Rightarrow \hat{Q}$, *where*

$$\begin{aligned}\hat{Q}_1(t) = \hat{Q}_1(0) &+ \int_0^t [(\nu - \lambda)\hat{Q}_3(u) - (\lambda + \mu)(\hat{Q}_1(u) + k_2)]du \\ &+ \sqrt{2m}B_1(t) - mY(t),\end{aligned} \tag{3.5}$$

$$\hat{Q}_2(t) = \hat{Q}_2(0) + \int_0^t (\lambda_2 - \mu\hat{Q}_2(u))du + \sqrt{2\lambda_1}B_2(t) - \lambda_1 Y(t), \tag{3.6}$$

$$\hat{Q}_3(t) = \hat{Q}_3(0) - \int_0^t \nu\hat{Q}_3(u))du + mY(t), \tag{3.7}$$

where B_1 *and* B_2 *are two independent standard Brownian motions, and* Y *is the non-negative, non-decreasing process such that* $\int_0^t (\hat{Q}_1(u) + \hat{Q}_2(u))dY(u) = 0$, $\forall t \geq 0$, *and* $Y(0) = 0$.

(b) The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \Rightarrow \hat{\pi}$, *where* $\hat{\pi}$ *is the unique invariant distribution of the diffusion process* $\hat{Q}(\cdot)$.

The following result proves that the diffusion process given by (3.5 -3.7) is well defined.

Lemma 5. *(3.5 -3.7) has a unique strong solution.*

Noting the convergence in Theorem 1(b), it would be quite useful if we can characterize the invariant distribution $\hat{\pi}$. However, we shall now demonstrate using a non-rigorous argument the difficulty in estimating $\hat{\pi}$. Let us assume this invariant distribution has a density $\hat{p} \in C^{2,2,1}$. Denote the state space of the diffusion process (3.5-3.7) by S, i.e., $S = \{(x, y, z) : x+y \leq 0, z \geq 0\}$, and its boundary by $\partial S = \{(x, y, z) : x+y = 0\}$. Let $L = ((\nu-\lambda)z-(\lambda+\mu)(x+k_2))\frac{\partial}{\partial x}+(\lambda_2-\mu y)\frac{\partial}{\partial y}-\nu z\frac{\partial}{\partial z}+m\frac{\partial^2}{\partial x^2}+\lambda_1\frac{\partial^2}{\partial y^2}$ denote the generator of the diffusion process; its domain is $C^{2,2,1}$. Pick any $f \in C^{2,2,1}$ with compact support such that $m\frac{\partial f}{\partial x}+\lambda_1\frac{\partial f}{\partial y} = 0$ on ∂S. Then, applying Ito's lemma to f for this diffusion process and taking expectations with respect to the invariant distribution, we obtain the condition

$$\int_S \hat{p}(v)Lf(v)dv = 0. \tag{3.8}$$

Assuming sufficient regularity conditions on $\hat{p}$, repeated use of integration by parts leads us to contend that $\hat{p}$ should solve the following partial differential equation (p.d.e):

$$m\frac{\partial^2\hat{p}}{\partial x^2}+\lambda_1\frac{\partial^2\hat{p}}{\partial y^2}+(\lambda+\mu)(x+k_2)\frac{\partial\hat{p}}{\partial x}-(\lambda_2-\mu y)\frac{\partial\hat{p}}{\partial y}+\nu z\frac{\partial\hat{p}}{\partial z}+(\lambda+2\mu+\nu)\hat{p} = 0,$$

for $(x, y, z) \in S \setminus \partial S$, with the boundary condition

$$-(\lambda x+(\lambda+\mu)k_2-\lambda_2)\hat{p}-m\frac{\partial\hat{p}}{\partial x}-\lambda_1\frac{\partial\hat{p}}{\partial y} = 0, \quad \text{for } (x, y, z) \in \partial S.$$

We are unable to solve this p.d.e, and thus cannot provide a better characterization of the invariant distribution. We shall see in the following that for the case $\nu = \lambda$, we can actually compute the invariant distribution of the limiting diffusion process.

3.1.1 Special case, $\nu = \lambda$

When $\nu = \lambda$, the Off and Hold state are indistinguishable, which implies that we do not need to keep track of the number of subscribers in the Hold state. This allows us

to focus on a two dimensional process $Q^n \in D_{\mathbb{R}^2}[0,\infty)$, where $Q_1^n(\cdot)$ and $Q_2^n(\cdot)$ denote the number of servers in use by the subscribers and the exogenous customer stream respectively. For $(x,y) \in \mathbb{R}_+^2$ let $\lambda^n(x,y) = ((n-x)\lambda,\ \lambda_p^n)'$ and $\mu(x,y) = (\mu x,\ \mu y)'$. Then, we obtain the following characterization of Q^n:

$$Q^n(t) = Q^n(0) + N^a\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\lambda(Q^n(u))\,du\right) - N^d\left(\int_0^t \mu(Q^n(u))\,du\right)$$

for $t \geq 0$. Defining $q^n(\cdot) = \frac{Q^n(\cdot)}{n}$, $\bar{q}(\cdot) = \left(\frac{\lambda}{\lambda+\mu}, \frac{\lambda_1}{\mu}\right)'$, and $\hat{Q}^n(\cdot) = \sqrt{n}(q^n(\cdot) - \bar{q}(\cdot)) - (k_2, 0)'$, the corresponding asymptotic results for this system are as follows.

Proposition 7. *If $k_1 = \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$ and $q^n(0) \to \bar{q}(0)$ a.s., then*

(a) $q^n \to \bar{q}$.

(b) If $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$, then $\hat{Q}^n \Rightarrow \hat{Q}$, where

$$\hat{Q}_1(t) = \hat{Q}_1(0) - \int_0^t (\lambda+\mu)(\hat{Q}_1(u) + k_2)du + \sqrt{2m}B_1(t) - mY(t) \tag{3.9}$$

$$\hat{Q}_2(t) = \hat{Q}_2(0) + \int_0^t (\lambda_2 - \mu\hat{Q}_2(u))du + \sqrt{2\lambda_1}B_2(t) - \lambda_1 Y(t), \tag{3.10}$$

where B_1 and B_2 are two independent standard Brownian motions, and Y is the non-negative, non-decreasing process such that $\int_0^t (\hat{Q}_1(u) + \hat{Q}_2(u))dY(u) = 0$, $\forall t \geq 0$, and $Y(0) = 0$.

(c) The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \Rightarrow \hat{\pi}$, where $\hat{\pi}$ is the unique invariant distribution of the diffusion process $\hat{Q}(\cdot)$.

This result is another special case of Lemma 4 and Theorem 1, and hence its proof is omitted.

We shall now characterize the invariant distribution of this diffusion process.

Proposition 8. *(a) If $\lambda_1 > 0$, the density corresponding to the invariant distribution of the diffusion process (3.9-3.10) is*

$$\hat{p}(x,y) = \begin{cases} \dfrac{\exp\left(-\frac{1}{2}\left(\frac{(x+k_2)^2(\lambda+\mu)}{m} + \frac{(y-\frac{\lambda_2}{\mu})^2\mu}{\lambda_1}\right)\right)}{\int_{-\infty}^{\infty}\int_{-\infty}^{-x}\exp\left(-\frac{1}{2}\left(\frac{(x+k_2)^2(\lambda+\mu)}{m} + \frac{(y-\frac{\lambda_2}{\mu})^2\mu}{\lambda_1}\right)\right)dy\,dx}, & \text{if } x+y \le 0, \\ 0, & \text{else.} \end{cases} \tag{3.11}$$

(b) If $\lambda_1 = 0$, the density corresponding to the invariant distribution of the diffusion process (3.9-3.10) is

$$\hat{p}(x,y) = \begin{cases} \dfrac{\exp\left(-\frac{1}{2m}(\lambda+\mu)(x+k_2)^2\right)}{\int_{-\infty}^{-\frac{\lambda_2}{\mu}}\exp\left(-\frac{1}{2m}(\lambda+\mu)(x+k_2)^2\right)dx}, & \text{if } y = \frac{\lambda_2}{\mu},\ x \le -\frac{\lambda_2}{\mu}, \\ 0, & \text{else.} \end{cases} \tag{3.12}$$

Proof. We shall prove (a); the proof of (b) follows in a similar manner. Denote the state space of the diffusion process (3.9-3.10) by S, i.e., $S = \{(x,y) : x+y \le 0\}$, and its boundary by $\partial S = \{(x,y) : x+y = 0\}$. The generator of the diffusion process (3.9-3.10) is given by $L = -(\lambda+\mu)(x+k_2)\frac{\partial}{\partial x} + (\lambda_2 - \mu y)\frac{\partial}{\partial y} + m\frac{\partial^2}{\partial x^2} + \lambda_1\frac{\partial^2}{\partial y^2}$ with the domain $C^{2,2}$. Pick any $f \in C^{2,2}$ with compact support such that $m\frac{\partial f}{\partial x} + \lambda_1\frac{\partial f}{\partial y} = 0$ on ∂S. It can be verified that $\hat{p}$ defined in (3.11) satisfies $\int_S \hat{p}(v)Lf(v)dv = 0$. Using $\hat{\pi}$ to denote the probability measure corresponding to the density $\hat{p}$, we mimic Proposition 1 in Ward and Glynn (2003) to verify that $\hat{\pi}$ is the invariant measure as follows: using Ito's formula we obtain $\mathbb{E}_{\hat{\pi}}[f(\hat{Q}(t)] = \mathbb{E}_{\hat{\pi}}[f(\hat{Q}(0)]$ for all $t \ge 0$, which implies that $\hat{\pi}$ must be an invariant measure. The result then follows by uniqueness of the invariant distribution. ■

Insensitivity of the steady-state distribution of the number-in-system process in the original system

It is well known that the steady-state distribution of the number-in-system in the M/G/k/k loss system is independent of the service time distribution (see Gross and Harris (1998), pages 245–247). It is also known that the steady-state distribution of

the number in system for the On-Off source model is independent of the distribution of On and Off times (see Cohen (1957)). We will now focus on the system with both subscribers and an exogenous pay-per-use stream and show that the steady-state distribution of the number-in-system of the pre-limit process is insensitive to the distribution of On times, Off times and the service times, i.e., it depends only on the means of these distributions, when Hold times are identically distributed as Off times. This insensitivity result does not hold when the Hold time and Off time distributions are different and when the exogenous stream does not arrive as a Poisson process.

Before stating the result, we shall introduce some notation. Let F and G denote the Off time and On time distributions respectively, with $\frac{1}{\lambda}$ and $\frac{1}{\mu}$ denoting their respective means. We shall assume that these distributions have densities. Note that G is also the distribution of service times for the Poisson arrivals. Let the total number of subscribers in the system be n, the arrival rate of the Poisson stream be λ_p^n, and the number of servers be k. Let $J^s(t)$ be a set whose elements represent the subscribers in the On state at time t, where each subscriber is numbered from 1 to n and $U = \{1, 2, \ldots, n\}$ denotes the set of all subscribers. Let $J^p(t)$ denote the number of customers from the Poisson stream in service at time t. Let $R^s(t) \in \mathbb{R}_+^n$ denote the residual times for the subscribers at time t, i.e., for subscriber $u \in J(t)$, $R_u^s(t)$ denotes the amount of On time remaining, while for subscriber $u \notin J(t)$, it denotes the amount of Off time remaining. Let $R^p(t)$ denote the residual times for the customers in service from the Poisson stream. It can be verified that $\{X(t) = (J^s(t), R^s(t), J^p(t), R^p(t)) : t \geq 0\}$ is a Markov process. The number-in-system process is given by $\{(|J^s(t)|, J^p(t)) : t \geq 0\}$, where for a set A, $|A|$ denotes its cardinality. As this process is not Markovian, its invariant distribution is not well defined. Thus, we shall work with the the steady-state probability of the number-in-system process, which is given by $\mathbb{P}((|J^s(\infty)|, J^p(\infty)) = (i, j))$ for $i, j \geq 0$. We are now ready to state the main insensitivity result.

Proposition 9. *The steady-state distribution of* $\{(|J^s(t)|, J^p(t)) : t \geq 0\}$*, the number-in-system process, depends on the distributions* F *and* G *only through their means.*

Proof. We shall begin by computing the steady-state distribution of the Markov process $X(\cdot)$. Let $\{X(t) : t \geq 0\}$ be defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Further, let

$$\pi^1(j^s, x^s) = \pi^s(|j^s|) \prod_{u \in j^s} \lambda \bar{G}(x_u^s) \prod_{u \notin j^s} \mu \bar{F}(x_u^s), \quad \text{and}$$
$$\pi^2(j^p, x^p) = \pi^p(j^p) \prod_{i=1}^{j^p} \mu \bar{G}(x_i^p),$$

where π^s denotes the invariant distribution for the stand-alone subscriber system with Off and On times exponentially distributed, π^p denotes the invariant distribution for the stand-alone Poisson arrival stream with service times exponentially distributed, and for a distribution H, $\bar{H} = 1 - H$. Further, we shall say a state (j^s, x^s, j^p, x^p) is feasible if $(j^s, x^s, j^p, x^p) \geq 0$, $x_i^p = 0$ for $i > j^p$ and $j^s + j^p \leq k$.

If the following result holds, then Proposition 9 follows by integrating (3.13) over the appropriate sets.

Lemma 6. *If, for an appropriate constant $C > 0$,*

$$\hat{\pi}(j^s, x^s, j^p, x^p) = \begin{cases} C(n - |j^s|)!|j^s|!(j^p!)\pi^1(j^s, x^s), \pi^2(j^p, x^p), & \textit{if state is feasible,} \\ 0, & \textit{else,} \end{cases} \tag{3.13}$$

then $\hat{\pi}(j^s, x^s, j^p, x^p)$ is the density of the steady-state distribution π of $\{X(t); t \geq 0\}$, i.e., $\pi(A) = \int_A \hat{\pi}(j^s, x^s, j^p, x^p) \prod_{\ell=1}^n dx_\ell^s \prod_{m=1}^{j^p} dx_m^p$ for any set $A \in \mathcal{F}$.

■

This insensitivity result proves that to asymptotically characterize the invariant distribution for such systems, it is sufficient to analyze a Markovian model. This allows us to circumvent the issue of obtaining the process limits for these systems, which are known to be measure-valued and fairly complicated, and provides a further justification for using Markovian models.

3.2 Convergence to Diffusion Limits

This section is concerned with the proof of Theorem 1(a). Note that Propositions 1 and 7 are special cases of Theorem 1. Proposition 1 follows from Theorem 1 by setting $\lambda_1 = \lambda_2 = 0$, while Proposition 7 follows by setting $Q_3^n(\cdot) = 0$, as the system is equivalent to one in which subscribers turn Off immediately when denied service.

Before beginning with the proof, we introduce the following notation

$$\kappa^n = \frac{k^n}{n}, \tag{3.14}$$

$$\theta^n(x,y,z) = \lambda^n(x,y,z) + \nu^n(x,y,z) - \mu^n(x,y,z), \tag{3.15}$$

$$\begin{aligned} M^{a,n}(t) &= \bar{N}^a\left(\int_0^t 1_{\{Q_1^n(u)+Q_2(u)<k^n\}}\lambda^n(Q^n(u))du\right) \\ &\quad + \bar{N}^r\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\nu Q_3^n(u)\,du\right)(1,0,0)' \end{aligned} \tag{3.16}$$

$$\begin{aligned} M^{d,n}(t) &= \bar{N}^d\left(\int_0^t \mu^n(Q^n(u))\,du\right) \\ &\quad + \bar{N}^r\left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\nu Q_3^n(u)\,du\right)(0,0,1), \end{aligned} \tag{3.17}$$

$$\alpha^n = \frac{1}{n}(M^{a,n} - M^{d,n}), \tag{3.18}$$

$$\delta^n = \frac{1}{n}\left(0,0,\Delta\bar{N}_1^a(Q_1^n,Q_2^n,Q_3^n)\right)', \tag{3.19}$$

$$S^n = \left\{(x,y,z)\in\mathbb{R}^3 | x,y,z\geq 0,\ x+y\leq k^n\right\}, \tag{3.20}$$

where S^n denotes the state space of the process Q^n, and for $(x,y,z)\in\mathbb{R}^3$, $\lambda^n(x,y,z) = ((n-x-z)\lambda, \lambda_1 n + \lambda_2\sqrt{n}, 0)'$, $\mu^n(x,y,z) = (\mu x, \mu y, 0)'$ and $\nu^n(x,y,z) = (\nu z, 0, -\nu z)'$.

Rearranging terms in (3.2), we obtain the following characterization.

Lemma 7. *For $t\geq 0$, $q^n(t)$ can be written as*

$$q^n(t) = X^n(t) + \int_0^t R^n(u)\,dY^n(u), \tag{3.21}$$

where $X^n, R^n \in D_{\mathbb{R}^3}[0,\infty)$ *and* $Y^n \in D_{\mathbb{R}}[0,\infty)$ *are defined as*

$$X^n(t) \equiv q^n(0) + \frac{1}{n}\int_0^t \theta^n(nq^n(u))\,du + \alpha^n(t) + \delta^n(t), \tag{3.22}$$

$$R^n(t) \equiv -((1-q_1^n(t)-q_3^n(t))\lambda + \nu q_3^n(t), \lambda_1 + \frac{\lambda_2}{\sqrt{n}}, -(1-q_1^n(t)-q_3^n(t))\lambda + \nu q_3^n(t))', \tag{3.23}$$

$$Y^n(t) \equiv \int_0^t 1_{\{q_1^n(u)+q_2^n(u)=\kappa^n\}}\,du. \tag{3.24}$$

In addition, we have

$$\hat{n}' \int_0^t R^n(u)\,dY^n(u) = -\sup_{0\le s\le t}\left(X_1^n(s) + X_2^n(s) - \kappa^n\right)^+,$$

where $\hat{n} = (1,1,0)'$.

We shall use the notation $\Phi^n_{\kappa^n}(X^n)(t)$ to denote $X^n(t) + \int_0^t R^n(u)\,dY^n(u)$ as in (3.22-3.24). A point worth noting about $R^n(t)$, which can be thought of as a reflection direction to keep the underlying process within the feasible region, is its dependence on the state at time t, as well as on n. This dependence complicates establishing regularity properties, such as Lipschitz continuity, on the reflection map $\Phi^n_{\kappa^n}$. To circumvent this issue, we introduce an intermediate system which is defined on the same probability space and differs from the original system only in the reflection direction. For this intermediate system, we choose the reflection direction

$$\tilde{R} \equiv \lim_{n\to\infty} R^n(t) = -\left(m, \lambda_1, -m\right)'. \tag{3.25}$$

We shall prove that this intermediate system is indistinguishable from the system under consideration at both the fluid and diffusion scale. It will then be sufficient to derive the diffusion limit for the intermediate system alone. However, to prove the equivalence of these two systems, we shall require the existence of a weak limit for the intermediate system. Thus, we shall first obtain the diffusion limit for the intermediate system, and then establish that this intermediate system is indistinguishable from the system under consideration at the fluid and diffusion scale, which will complete

the proof.

We shall now define the intermediate system. We shall use the same notation for the intermediate system as we did for the original system, albeit with a modifier "~" to differentiate between the two, i.e., $\tilde{Q}^n \in D_{\mathbb{R}^3}[0,\infty)$ represents the number-in-system process for this system. Before defining the intermediate system, we introduce a mapping $\tilde{\Phi}_a : D_{\mathbb{R}^3}[0,\infty) \to D_{\mathbb{R}^3}[0,\infty)$ such that

$$\tilde{\Phi}_a(X) \equiv X + \tilde{R}\tilde{Y}, \tag{3.26}$$

where $\tilde{R}$ is given by (3.25) and for $t \geq 0$

$$\tilde{Y}(t) = -\frac{\sup_{0\leq s\leq t}\left(X_1(s) + X_2(s) - a\right)^+}{\hat{n}'\tilde{R}}$$

with $\hat{n} = (1,1,0)'$. Note that the mapping $\tilde{\Phi}_a$ is independent of n.

We then define the following dynamics for the scaled number-in-system process for the intermediate system $\tilde{q}^n(\cdot) \equiv \frac{\tilde{Q}^n(\cdot)}{n}$:

$$\tilde{q}^n(t) = \tilde{\Phi}_{\kappa^n}(\tilde{X}^n)(t) = \tilde{X}^n(t) + \tilde{R}\tilde{Y}^n(t),$$

where

$$\tilde{X}^n(t) \;=\; \tilde{q}^n(0) + \frac{1}{n}\int_0^t \theta^n(n\tilde{q}^n(u))du + \tilde{\alpha}^n(t) + \tilde{\delta}^n(t), \tag{3.27}$$

$\tilde{Y}^n(t) = -\frac{\sup_{0\leq s\leq t}\left(\tilde{X}_1^n(s)+\tilde{X}_2^n(s)-\kappa^n\right)^+}{\hat{n}'\tilde{R}}$, and we have the following analog to (3.15-3.19).

$$\theta^n(x,y,z) = \lambda^n(x,y,z) + \nu^n(x,y,z) - \mu^n(x,y,z), \tag{3.28}$$

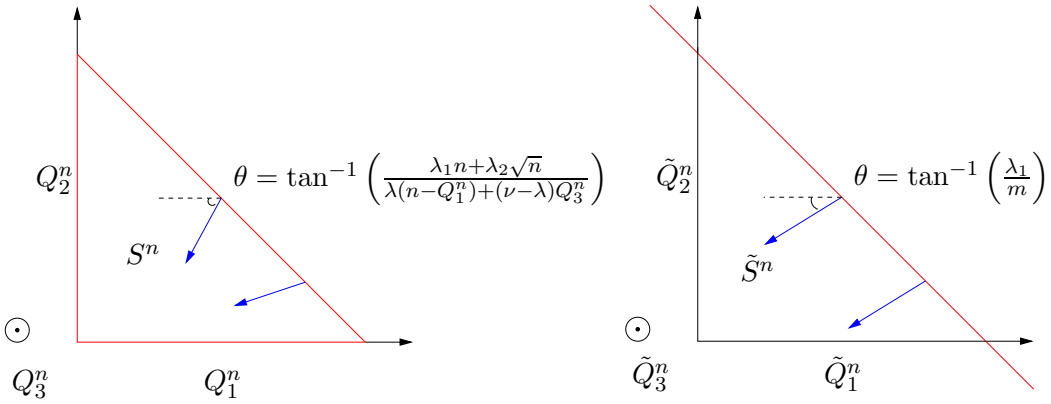
$$\begin{aligned}\tilde{M}^{a,n}(t) = {} & \bar{N}^a\left(\int_0^t 1_{\{\tilde{Q}_1^n(u)+\tilde{Q}_2^n(u)<k^n\}}\lambda^n(\tilde{Q}^n(u))du\right) \\ & + \bar{N}^r\left(\int_0^t 1_{\{\tilde{Q}_1^n(u)+\tilde{Q}_2^n(u)<k^n\}}\nu\tilde{Q}_3^n(u)^+\,du\right)(1,0,0)'\end{aligned} \tag{3.29}$$

$$\begin{aligned}\tilde{M}^{d,n}(t) &= \bar{N}^d\left(\int_0^t \mu^n(\tilde{Q}^n(u)^+)\,du\right)\\ &\quad + \bar{N}^r\left(\int_0^t 1_{\{\tilde{Q}_1^n(u)+\tilde{Q}_2^n(u)<k^n\}}\nu\tilde{Q}_3^n(u)^+\,du\right)(0,0,1)',\end{aligned} \tag{3.30}$$

$$\tilde{\alpha}^n = \frac{1}{n}(M^{a,n} - M^{d,n}), \tag{3.31}$$

$$\tilde{\delta}^n = \frac{1}{n}\left(0,0,\Delta\bar{N}_1^a(\tilde{Q}_1^n,\tilde{Q}_2^n,\tilde{Q}_3^n)\right)'. \tag{3.32}$$

The state-space of the $\tilde{Q}^n(\cdot)$ process is $\tilde{S}^n = \{(x,y,z) : x+y \le k^n, x+z \le n\}$. Note that as the components of $\tilde{Q}^n$ can take negative values as well, this intermediate system has no "real" interpretation. Figure 3.1 provides a two-dimensional cross-section of the state space along with the reflection direction of the number-in-system process in the two systems.



(a) Original system (b) Intermediate system

Figure 3.1: Reflection directions in the two systems.

We now obtain the asymptotic limits for the intermediate system. Defining $Q^{*n}(\cdot) = \sqrt{n}(\tilde{q}^n(\cdot) - \bar{q}(\cdot)) - (k_2,0,0)'$, where $\bar{q}(\cdot) = \left(\frac{\lambda}{\lambda+\mu}, \frac{\lambda_1}{\mu}, 0\right)'$, we have the following asymptotic results.

Proposition 10. *If* $k_1 = \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$ *and* $\tilde{q}^n(0) \to \bar{q}(0)$ *a.s., then*

(a) $\tilde{q}^n \to \bar{q}$.

(b) If $Q^{*n}(0) \Rightarrow Q^*(0)$, then $Q^{*n} \Rightarrow Q^*$, where

$$Q^*(t) = \tilde{\Phi}_0(Z^*(t)) \tag{3.33}$$

$$Z_1^*(t) = Q_1^*(0) + \int_0^t [(\nu - \lambda)Q_3^*(u) - (\lambda + \mu)(Q_1^*(u) + k_2)]\, du + \sqrt{2m}B_1(t) \tag{3.34}$$

$$Z_2^*(t) = Q_2^*(0) + \int_0^t (\lambda_2 - \mu Q_2^*(u))du + \sqrt{2\lambda_1}B_2(t) \tag{3.35}$$

$$Z_3^*(t) = Q_3^*(0) - \int_0^t \nu Q_3^*(u)du, \tag{3.36}$$

where B_1 and B_2 are two independent standard Brownian motions.

Proof. Part (a) follows in an identical manner as Lemma 4(a). We shall prove Part (b) on similar lines as Theorem 7.2 in Mandelbaum and Pats (1998). First, we shall rewrite Q^{*n} in the following manner.

$$Q^{*n} = \sqrt{n}\left(\tilde{\Phi}_{\kappa^n}\left(\bar{q} + \frac{\tilde{Z}^n + (k_2, 0, 0)'}{\sqrt{n}}\right) - \tilde{\Phi}_{\kappa^n}(\bar{q})\right) - (k_2, 0, 0)',$$

where

$$\tilde{Z}^n(t) = Q^{*n}(0) + \tilde{D}_\theta^n(t) + \tilde{M}^n(t) + \sqrt{n}\tilde{\delta}^n, \tag{3.37}$$

$$\tilde{D}_\theta^n(t) = \sqrt{n}\int_0^t \left(\theta(\tilde{q}^n(u)) - \theta(\bar{q}(u)) + \left(0, \frac{\lambda_2}{\sqrt{n}}\right)\right) du, \tag{3.38}$$

$$\tilde{M}^n = \sqrt{n}\tilde{\alpha}^n, \tag{3.39}$$

where $\theta(x, y, z) = (-(\lambda + \mu)x + (\nu - \lambda)z, -\mu y, -\nu z)'$ for $(x, y, z) \in \mathbb{R}^3$.

Using the fact that for any vector $z \in \mathbb{R}^3$, $\Phi_{a+z_1+z_2}(X + z) = \Phi_a(X) + z$ and for any $b \in \mathbb{R}_+$, $b\Phi_0(X) = \Phi_0(bX)$, we can rewrite Q^{*n} as

$$Q^{*n} = \tilde{\Phi}_0\left(\tilde{Z}^n\right). \tag{3.40}$$

We now state a few results that will be used to complete the proof. The first result

proves that the mapping $\tilde{\Phi}_a$ defined in (3.26) is Lipschitz continuous. Lemma 9 proves the convergence of the processes $\tilde{M}^n$ and $\sqrt{n}\tilde{\delta}^n$. Lemmas 10-12 prove the compact containment, tightness, and weak convergence of the relevant processes.

Lemma 8. *The mapping $\tilde{\Phi}_a$ is Lipschitz continuous for any $a \in \mathbb{R}$, i.e, there exists $K \geq 0$ such that for every $T > 0$ and $Z^1,\ Z^2 \in D_{\mathbb{R}^3}[0,\infty)$, $\| \tilde{\Phi}_a(Z^1) - \tilde{\Phi}_a(Z^2) \|_T \leq K\| Z^1 - Z^2 \|_T$.*

Lemma 9. *The following convergences hold.*

(a) $\tilde{M}^n \Rightarrow W$, where $W_1(\cdot) = \sqrt{2m}B_1(\cdot)$, $W_2(\cdot) = \sqrt{2\lambda_1}B_2(\cdot)$, and $W_3(\cdot) = 0$.

(b) $\sqrt{n}\tilde{\delta}^n \Rightarrow 0$.

Lemma 10. *The sequence $\{Q^{*n}\}$ adheres to the compact containment condition*

$$\lim_{\ell \uparrow \infty} \limsup_n \mathbb{P}\{\| Q^{*n} \|_T > \ell\} = 0.$$

Lemma 11. *The sequence $(Q^{*n}, \tilde{Z}^n, \tilde{D}^n_\theta, \tilde{M}^n, \sqrt{n}\tilde{\delta}^n)$ is C-tight.*

Lemma 12. *Let $(Q^*, Z^*, \tilde{D}_\theta, W, 0)$ be any weak limit of $(Q^{*n}, \tilde{Z}^n, \tilde{D}^n_\theta, \tilde{M}^n, \sqrt{n}\tilde{\delta}^n)$. Then, Z^* satisfies (3.33-3.36).*

The proof of Part (b) thus follows. ■

We shall now prove that the intermediate system is asymptotically equivalent to the original system at the $\sqrt{n}$-scale, which implies that both these systems have identical diffusion limits. Although, this equivalence result holds in far more generality, we shall prove it only for the case we are interested in.

Lemma 13. *For any $T > 0$, if $k_1 = \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$, $q^n \to \bar{q}$, $\tilde{q}^n \to \bar{q}$ and $|\hat{Q}^n(0) - Q^{*n}(0)| \Rightarrow 0$, then $\| \hat{Q}^n - Q^{*n} \|_T \Rightarrow 0$.*

This, along with Proposition 10, completes the proof of Theorem 1(a).

3.3 Convergence of Invariant Distributions

The subject of this section is the proof of Theorem 1(b). We shall prove the convergence of the invariant distributions, as well as uniqueness of the invariant distribution of the limiting diffusion process. The proof employs a Lyapunov function argument as in Gamarnik and Zeevi (2006). However, as the processes we consider have state-dependant drift and Gamarnik and Zeevi (2006) prove the limit interchange for Jackson networks (with state-independent drift), their main results are not directly applicable here. We shall begin with the following definitions for a Markov chain $(Q(t) : t \geq 0)$ with a complete, metrizable state space $\mathcal{S}$ as in Gamarnik and Zeevi (2006).

Definition 1. *A function $f : \mathcal{S} \to \mathbb{R}_+$ is said to be a Lyapunov function with drift size parameter $-\gamma$, where $\gamma > 0$, drift time parameter $t_0 > 0$, and exception parameter K, if*

$$\sup_{\{x \in \mathcal{S}: f(x) > K\}} \{\mathbb{E}_x f(Q(t_0)) - f(x)\} \leq -\gamma.$$

A function $f : \mathcal{S} \to \mathbb{R}_+$ is said to be a geometric Lyapunov function with a geometric drift size $0 < \gamma < 1$, drift time $t_0 > 0$, and exception parameter K, if

$$\sup_{\{x \in \mathcal{S}: f(x) > K\}} \left\{(f(x))^{-1} \mathbb{E}_x f(Q(t_0))\right\} \leq \gamma.$$

Define $\phi(t) = \sup_{x \in \mathcal{S}} (f(x))^{-1} \mathbb{E}_x f(Q(t))$ and for any given $\beta > 0$,

$$L_1(\beta, t) \equiv \sup_{x \in \mathcal{S}} \mathbb{E}_x \left[\exp(\beta(f(Q(t)) - f(x)))\right]$$
$$L_2(\beta, t) \equiv \sup_{x \in \mathcal{S}} \mathbb{E}_x \left[(f(Q(t)) - f(x))^2 \exp(\beta(f(Q(t)) - f(x))^+)\right],$$

for $t \geq 0$.

Let $\hat{q}^n = n\,(\bar{q} - q^n) - (k_2\sqrt{n}, 0, 0)'$. Recall that $q^n = Q^n/n$ and $\bar{q}(\cdot) = (\lambda/(\lambda + \mu), \lambda_1/\mu, 0)'$. Differing from Gamarnik and Zeevi (2006), we shall prove that the function $f(z) = e'|z|$, where $e = (1, 1, 1)'$, is a Lyapunov function, where $|z| = (|z_1|, |z_2|, |z_3|)'$. The fact that the centered and scaled process can take negative

values leads us to this choice. We shall use $\mathcal{S} = \mathbb{R}^2 \times \mathbb{R}_+$ equipped with the sup norm.

Lemma 14. *For all sufficiently large n, the function $f(z) = e'|z|$ is a Lyapunov function with drift size parameter $-\sqrt{n}$, drift time parameter t_0 and exception parameter $c_0\sqrt{n}$. In addition, the following hold.*

$$\limsup_{n\to\infty} L_1(\beta_0/\sqrt{n}, t_0) < \infty, \tag{3.41}$$

$$\limsup_{n\to\infty} \frac{1}{n} L_2(\beta_0/\sqrt{n}, t_0) < \infty. \tag{3.42}$$

We now establish the required tightness of the invariant distributions $\{\hat{\pi}^n\}$. Note that the existence and uniqueness of these invariant distributions follows from the fact that their corresponding Markov chains are irreducible and are defined on a finite state-space, and thus are positive-recurrent.

Proposition 11. *There exist constants C_1, c_1 such that for all sufficiently large n, the sequence of invariant distributions $\hat{\pi}^n$ satisfies*

$$P_{\hat{\pi}^n}(n^{-\frac{1}{2}} e'|\hat{q}^n(0)| > s) \leq C_1 \exp(-c_1 s),$$

for all $s > 0$.

Proof. Using Lemma 14 for $\beta_n = c\beta_0 n^{-\frac{1}{2}}$ where $0 < c < 1$ is a constant, we obtain

$$\limsup_{n\to\infty} n^{-\frac{1}{2}} \beta_n L_2(\beta_n, t_0) = \limsup_{n\to\infty} n^{-1} c\beta_0 L_2(\beta_n, t_0) \leq 1$$

for c chosen sufficiently small. The following result from Gamarnik and Zeevi (2006) (Theorem 6) for Markov chains defined on a complete, metrizable state space $\mathcal{S}$ will be useful in proving this result.

Lemma 15. *Suppose the Markov process $Q(\cdot)$ possesses an invariant distribution π, and suppose f is a Lyapunov function with parameters γ, t_0, K. Assume in addition that there exists $\beta > 0$ such that*

$$\beta L_2(\beta, t_0) \leq \gamma. \tag{3.43}$$

Then, $e^{\beta f(\cdot)}$ is a geometric Lyapunov function with geometric drift size parameter $(1-\gamma\beta/2)$, drift time parameter t_0, and exception parameter $e^{\beta K}$. Consequently, for every $s > K$

$$P_\pi(f(Q(0)) > s) \leq (1-\gamma\beta/2)^{-1} L_1(\beta, t_0) e^{-\beta(s-K)}.$$

Now, applying Lemma 15 and using $\beta_n = c\beta_0 n^{-\frac{1}{2}}$, we obtain for every $s > 0$

$$\begin{aligned} P_{\hat{\pi}^n}\left(e'|\hat{q}^n(0)|/\sqrt{n} > s\right) &\leq \frac{L_1(\beta_n, t_0)}{1-c\beta_0/2} \exp\left(-\beta_n(sn^{\frac{1}{2}} - c_0 n^{\frac{1}{2}})\right) \\ &= \frac{L_1(\beta_n, t_0)}{1-c\beta_0/2} \exp\left(-c\beta_0(s-c_0)\right). \end{aligned}$$

(3.41) implies that there exists a sufficiently large $n_0 > 0$ and a constant $c_2 > 0$ such that $L_1(\beta_n, t_0) < c_2$ for $n > n_0$. Then for all $s > 0$ and all $n > n_0$

$$P_{\hat{\pi}^n}\left(e'|\hat{q}^n(0)|/\sqrt{n} > s\right) \leq \frac{c_2}{1-c\beta_0/2} \exp\left(-c\beta_0(s-c_0)\right).$$

■

A consequence of this proposition is the following result.

Corollary 1. *Let $\frac{\hat{q}^n(0)}{\sqrt{n}}$ be distributed according to $\hat{\pi}^n$. Then the sequence of random vectors $\{\hat{q}^n(0)/\sqrt{n}\}$ is tight.*

As any limit point of a converging sub-sequence of $\{\hat{\pi}^n\}$ must be invariant for the limiting diffusion process $\hat{Q}(\cdot)$, this result proves the existence of invariant distributions for the diffusion process. We shall now prove that the diffusion process can only have a unique invariant distribution, and this shall complete the proof of Theorem 1(b).

Lemma 16. *The diffusion process $\hat{Q}(\cdot)$ given by (3.5-3.7) has a unique invariant distribution.*

Having established the desired weak convergence of the invariant distributions, we shall end this chapter by proving the following uniform integrability result that will be useful.

Proposition 12. *There exists $\delta > 0$ such that the family $\{\exp(\delta e'|\hat{q}^n(0)|/\sqrt{n})\}$ is uniformly integrable, where $\frac{\hat{q}^n(0)}{\sqrt{n}} \sim \hat{\pi}^n$ for each $n = 1, 2, \ldots$. Consequently,*

$$\mathbb{E}_{\hat{\pi}^n}[n^{-\frac{1}{2}}\hat{q}_i^n(0)] \to \mathbb{E}_{\hat{\pi}}[\hat{Q}_i(0)], \quad \text{for } i = 1, 2, 3.$$

Proof. Using Proposition 11, we have for all sufficiently large n and $\delta > 0$

$$\begin{aligned}\mathbb{E}_{\hat{\pi}^n}[\exp(n^{-\frac{1}{2}}\delta e'|\hat{q}^n(0)|)] &= \int_0^\infty \mathbb{P}(n^{-\frac{1}{2}}e'|\hat{q}^n(0)| > \delta^{-1}\log x)dx \\ &\leq 1 + \int_1^\infty C_1 \exp(-c_1\delta^{-1}\log x)dx \\ &< \infty,\end{aligned}$$

provided $c_1\delta^{-1} > 1$. Setting $0 < \delta < c_1$, we obtain the result. ∎

Chapter 4

Operational Benefits of Subscription Services

This chapter builds an economic framework around the models introduced in earlier chapters. We shall now assume that our customers are price and quality-of-service sensitive, where the customers' perception of quality is determined by their likelihood of obtaining the product or service immediately upon request. We shall study the firm's alternatives of offering either a subscription option or a pay-per-use option from a profit-maximizing perspective. In order to perform this comparison, we shall make the behavioral assumption that the customers' propensity to consume the products is the same in both options, i.e., we shall assume that the maximum load on the system (through the customers' requests for service) is identical for both the options. We shall show that in a large market setting using the subscription option is more profitable for the firm. However, we shall show that it is not necessarily true that the subscription option dominates the pay-per-use option on quality-of-service. The firm is able to manage the trade-off between price and quality-of-service better in the subscription option. Moreover, we shall show that the social welfare and the consumer surplus can also be higher in the subscription option, indicating that both the firm and the consumers can benefit from the subscription option.

Our approach for solving the firm's optimization problem for either option (or for the combined case) is as follows. Denoting the size of the market, that is, the number

of potential subscribers or the potential arrival rate of pay-per-use customers, by n, we first solve a nominal optimization problem at the $O(n)$ level. (As usual $O(n)$ denotes any quantity such that $\frac{O(n)}{n}$ is bounded as $n \to \infty$.) In such a system the natural scale at which stochastic variability manifests itself is $O(\sqrt{n})$. Any refinement to the nominal problem that takes into account the stochastic variability in the system must be at the $O(\sqrt{n})$ level. The following example may help illustrate this scale difference. Consider a newsvendor model with demand arising from n independent, identically distributed sources. In this case, the optimal stock level comprises of the mean of the total demand, which is $O(n)$, and a safety stock against variability, which is some number of standard deviations of the total demand and is $O(\sqrt{n})$. The safety stock in the newsvendor model is exactly analogous to the refinement to the nominal solution that we propose. By applying a law of large numbers argument, Proposition 13 shows that on the $O(n)$ scale, the system operates in a deterministic regime with essentially all requests being satisfied immediately. Furthermore, using a functional central limit theorem proved in earlier work, we perform a refined analysis on the $O(\sqrt{n})$ scale to obtain prices and capacity levels that are optimal on this scale. That is, these choices result in profits that are within $o(\sqrt{n})$ of the optimal values (see Proposition 16). (As usual $o(n)$ denotes any quantity such that $\frac{o(n)}{n} \to 0$ as $n \to \infty$.) Furthermore, Proposition 15 shows that in this regime pricing and capacity sizing are equivalent levers for the firm. That is, any refinement in prices can be imitated by an equivalent refinement in capacity levels for both options. This asymptotic approach allows us to compare the profits obtained under each option up to a resolution of $o(\sqrt{n})$. Equating the nominal solutions for both options, we prove the primary result of this chapter, that is, the profit is higher for the subscription option by a quantity that is $O(\sqrt{n})$. Combining this with the size of the possible error in estimating the profit allows us to conclude that in a large enough market, the subscription option dominates.

This chapter is organized in the following manner. In Section 4.1 we analyze the subscription option. We build the customer demand function, set up the firm's optimization problem, and approximately solve it. We solve both the nominal problem on the $O(n)$ scale and its variability refinement and show that pricing and capacity sizing are equivalent levers for the firm on the $O(\sqrt{n})$ scale. In Section 4.2, we

perform an analogous analysis for the pay-per-use option. Section 4.3 compares the two options and proves the main result of this chapter, that is, the firm's profits are higher when offering a subscription option. Section 4.4 solves the profit maximization problem for the firm when it offers both options, and finally Section 4.5 discusses the conclusions and scope for future work. All the proofs of results in this chapter can be found in Appendix C.

4.1 The Subscription Option

We begin by building a demand function for the subscribers. We consider a market with a total of n *potential subscribers*. (For a detailed description of the subscriber model see Section 2.1.) We shall assume that the retrial rate of the subscribers $\nu = \lambda$; Section 4.5 discusses the general case. In accordance with Naor's original idea (see Naor (1969)), we assume that each potential subscriber has a value S for each service completed. This value is assumed to be the same for every service completed for each potential subscriber but may differ among the potential subscribers. The valuations are drawn from some distribution with the cumulative distribution function $F(\cdot)$ and the density function $f(\cdot)$. Each potential subscriber has a reservation value of $r \geq 0$, which is to be interpreted as the value per unit time she would obtain from some alternative to the current system. This reservation value is used by the potential subscriber to decide whether to join the system or not.

The subscribers' sensitivity to the quality-of-service is captured through the denial probability, that is, the steady-state probability that upon request a subscriber does not receive the product. We characterize the denial probability by

$$\gamma = \lim_{t \to \infty} \frac{\text{No. of denied attempts by time } t}{\text{No. of attempts by time } t}. \tag{4.1}$$

The higher the denial probability the longer it takes a subscriber to obtain the product. In particular, the mean time to complete a service (starting from the Off state) for each subscriber is $\frac{1}{\lambda} + \frac{\sum_{i=1}^{\infty} \gamma^i}{\nu} + \frac{1}{\mu}$. (Lemma 24 in Appendix C proves this result for $\nu = \lambda$; for general retrial rates, this relation is approximate.)

A potential subscriber joins the system only if the value she obtains per service exceeds her cost. Thus, when the system manager sets a subscription fee of p per unit time, the number of potential subscribers that join the system, $N(p,\gamma)$, is given by

$$\begin{aligned} N(p,\gamma) &= n\mathbb{P}\left(S > \left(\frac{1}{\lambda} + \frac{\sum_{i=1}^{\infty}\gamma^i}{\nu} + \frac{1}{\mu}\right)(r+p)\right) \\ &= n\bar{F}\left(\left(\frac{1}{\lambda} + \frac{\gamma}{\nu(1-\gamma)} + \frac{1}{\mu}\right)(r+p)\right), \end{aligned} \tag{4.2}$$

where $\bar{F}(\cdot)$ denotes the tail of the cumulative density function. (Note that we are approximating the number of subscribers that actually join the system, which is a random quantity, by its mean.)

In addition to the subscription fee, the system manager also decides on the capacity level k. Assuming each server costs $\$c$ per unit time, where $c > 0$, the optimization problem of the system manager can be stated as

$$\begin{aligned} \max_{(p,k)\in\mathbb{R}_+\times\mathbb{Z}_+} \quad & pN(p,\gamma) - kc \\ \text{s.t.} \quad & \gamma = d(N(p,\gamma),k), \end{aligned}$$

where $d(N(p,\gamma),k)$ is the denial probability as a function of the number of subscribers and the capacity level. Note that for any subscription fee and capacity level set, the denial probability must satisfy an equilibrium condition, which is captured via the constraint $\gamma = d(N(p,\gamma),k)$. The inability to characterize this equilibrium denial probability in a simple form renders this problem, as stated, extremely hard to solve. This motivates us to try to approximately solve it in the natural asymptotic regime of large number of potential subscribers. In particular, we shall let n grow without bound and use the superscript n to make explicit the dependence of the various parameters on n, i.e., p^n, k^n, and γ^n denote the subscription fee, number of products, and denial probability respectively, and $N^n(p^n,\gamma^n)$ denotes the number of potential

subscribers joining the system. The optimization problem can then be written as

$$\max_{(p^n,k^n)\in\mathbb{R}_+\times\mathbb{Z}_+} \Pi^n(p^n,k^n) \equiv p^n N^n(p^n,\gamma^n) - k^n c$$
$$\text{s.t. } \gamma^n = d(N^n(p^n,\gamma^n),k^n).$$

With this setup, we try to find a sequence of subscription fees and number of products (p^{n*},k^{n*}) that is optimal in the limiting regime as $n\to\infty$. We shall begin by providing a nominal solution, and then constructing a refinement that accounts for the variability. The nominal solution optimizes the objective function on the $O(n)$ scale by appealing to the law of large numbers, while use of the central limit theorem implies that the refinement is on the $O(\sqrt{n})$ scale. The $O(\sqrt{n})$ scale refinement can be understood by considering a newsvendor model with demand arising from n independent, identically distributed sources. In this case, the optimal stock level comprises of the mean of the total demand, which is $O(n)$, and a safety stock against variability, which is some number of standard deviations of the total demand and is $O(\sqrt{n})$.

A sequence (p^{n*},k^{n*}) is said to be a nominal solution to the problem if for all sequences (p^n,k^n)

$$\liminf_{n\to\infty} \frac{\Pi^n(p^{n*},k^{n*})}{n} \geq \limsup_{n\to\infty} \frac{\Pi^n(p^n,k^n)}{n}.$$

We denote the corresponding nominal profit by $\bar{\Pi}$, i.e., $\bar{\Pi} = \liminf_{n\to\infty} \frac{\Pi^n(p^{n*},k^{n*})}{n}$. The nominal solution is a deterministic solution, and thus can be refined by studying the variability in the system. To do so, we introduce the notion of asymptotic optimality at the $O(\sqrt{n})$ scale. We expect the actual profit in the system to be less than that expected in the nominal setting. This motivates us to look at the loss in profits due to variability and refine our solution to minimize these losses. That is, for any sequence (p^n,k^n), denoting $\tilde{\Pi}^n(p^n,k^n) \equiv \sqrt{n}\left(\bar{\Pi} - \frac{\Pi^n(p^n,k^n)}{n}\right)$ as the $O(\sqrt{n})$ scale loss in profit, which is the difference between the nominal profit $\bar{\Pi}$ and that obtained while using the rule (p^n,k^n), we define an asymptotically optimal sequence as follows.

Definition 2. *A sequence (p^{n*},k^{n*}) is said to be asymptotically optimal if the scaled*

loss in profit is minimized, that is, for any sequence (p^n, k^n)

$$\limsup_{n\to\infty} \tilde{\Pi}^n(p^{n*}, k^{n*}) \leq \liminf_{n\to\infty} \tilde{\Pi}^n(p^n, k^n).$$

It is clear that for a sequence to be asymptotically optimal, it must be a nominal solution.

4.1.1 Nominal solution

We shall begin by solving the nominal problem. This problem is achieved by scaling the objective function by n and letting n grow without bound. This resulting problem is in fact deterministic, and hence one expects that its solution should have $\gamma = 0$. Further, for $\gamma = 0$, the capacity level must equal the offered load in the system, which is defined as the mean number of products that will be in use in steady-state when the system has infinite capacity. Using a renewal theory argument, we can show that the probability that a subscriber will be using a product in steady-state is $\frac{\lambda}{\lambda+\mu}$. Hence, as the number of subscribers that join this system is $n\bar{F}\left(\frac{r+p^n}{m}\right)$, where $m \equiv \frac{\lambda\mu}{\lambda+\mu}$, the offered load is $n\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+p^n}{m}\right)$. This implies that if the price and capacity levels converge as $(p^n, k^n/n) \to (p, k)$, then the capacity level k must equal $\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+p}{m}\right)$. Thus, we obtain the following static optimization problem, that we call the nominal problem. (This will be proved rigorously in Proposition 13.)

$$\begin{aligned} \max_{(p,k)\in\mathbb{R}^2_+} \; & p\bar{F}\left(\frac{r+p}{m}\right) - kc \\ \text{s.t. } & k = \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+p}{m}\right). \end{aligned} \tag{4.3}$$

Noting that p cannot be extremal, we use the necessary condition for optimality, that is, the first derivative of the objective function must be zero, to obtain that the optimal subscription fee $\bar{p}$ satisfies the relation

$$\bar{F}\left(\frac{r+\bar{p}}{m}\right) = \frac{1}{m} f\left(\frac{r+\bar{p}}{m}\right)\left(\bar{p} - \frac{\lambda}{\lambda+\mu}c\right), \tag{4.4}$$

and the optimal capacity level is given by $\bar{k} = \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+\bar{p}}{m}\right)$. We then have the following result.

Proposition 13. *A sequence (p_p^n, k_p^n) is a nominal solution if $\gamma^n \to 0$ and $\left(p^n, \frac{k^n}{n}\right) \to (\bar{p}, \bar{k})$.*

4.1.2 Refining the nominal solution

We begin by defining the capacity imbalance in the system as the difference between the number of servers and the offered load scaled by n. As derived earlier, this offered load is $n\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+p^n}{m}\right)$. Thus, for any sequence (p^n, k^n), the capacity imbalance is given by

$$\theta^n = \frac{k^n}{n} - \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+p^n}{m}\right). \tag{4.5}$$

The following lemma presents a "decay" condition on the asymptotically optimal imbalance.

Lemma 17. *For any asymptotically optimal sequence (p^n, k^n), $\limsup_{n\to\infty} |\theta^n\sqrt{n}| < \infty$.*

We now focus on constructing an asymptotically optimal sequence (p^n, k^n) with the associated equilibrium denial probability γ^n. Motivated by the above lemma, we restrict our search to θ^n such that $\lim_{n\to\infty}\theta^n\sqrt{n} = \theta$, where $\theta \in \mathbb{R}$. We choose $p^n = \bar{p} + \phi^n$ and $k^n = (\bar{k} + \kappa^n)n$ such that (p^n, k^n, θ^n) satisfy (4.5) and $\phi^n, \kappa^n \to 0$. In fact we shall assume without loss of generality that $\phi^n\sqrt{n} \to \phi$ and $\kappa^n\sqrt{n} \to \kappa$. We can rewrite (4.5) using the Taylor series expansion of $\bar{F}\left(\frac{r+p^n}{m}\right)$ about $\left(\frac{r+\bar{p}}{m}\right)$ to obtain

$$\theta^n = \frac{\phi^n}{\mu} f\left(\frac{r+\bar{p}}{m}\right) + \kappa^n + o(\phi^n), \tag{4.6}$$

and hence

$$\theta = \frac{\phi}{\mu} f\left(\frac{r+\bar{p}}{m}\right) + \kappa.$$

We are now ready to asymptotically characterize the denial probability for our system.

Proposition 14. *For a sequence (p^n, k^n) such that $\theta \equiv \lim_{n\to\infty} \theta^n\sqrt{n}$ exists, $\gamma^n\sqrt{n} \to \gamma$ given by*

$$\gamma = \sqrt{\frac{\lambda+\mu}{m\bar{F}\left(\frac{r+\bar{p}}{m}\right)}}\, h\left(-\left[\theta + f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\frac{\gamma}{\lambda+\mu}\right]\sqrt{\frac{\lambda+\mu}{m\bar{F}\left(\frac{r+\bar{p}}{m}\right)}}\right). \tag{4.7}$$

Note that in the limit the denial probability solves a fixed point relation and is a function of only the limiting capacity imbalance. In the following result, we show that the profit is also a function of the capacity imbalance alone, and hence pricing and capacity sizing are equivalent levers for the firm at the $O(\sqrt{n})$ scale.

Proposition 15. *For a sequence (p^n, k^n) such that $p^n = \bar{p} + \phi^n$, $k^n = (\bar{k} + \kappa^n)n$ and $\theta \equiv \lim_{n\to\infty} \theta^n\sqrt{n}$ exists, the loss in profit function satisfies*

$$\lim_{n\to\infty} \tilde{\Pi}^n(p^n, k^n) = \theta c + \bar{p} f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\gamma/\lambda.$$

Therefore, any (p^n, k^n) that has the same limiting imbalance θ has the same asymptotic loss in profit.

This result implies that the asymptotically optimal sequence is characterized by the value of θ that minimizes $\lim_{n\to\infty} \tilde{\Pi}^n(p^n, k^n)$. Hence, the optimal asymptotic capacity imbalance θ^* solves

$$\begin{aligned} &\min_{\theta\in\mathbb{R}} \ \theta c + \bar{p} f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\gamma/\lambda \\ &\text{s.t. } (4.7). \end{aligned}$$

Denoting $z \equiv h'^{-1}\left(\frac{c\lambda m\bar{F}\left(\frac{r+\bar{p}}{m}\right)}{(r+\bar{p})(\lambda+\mu)\bar{p}f\left(\frac{r+\bar{p}}{m}\right)-(r+\bar{p})f\left(\frac{r+\bar{p}}{m}\right)c\lambda}\right)$, the first order conditions imply that the optimal denial probability and capacity imbalance are

$$\begin{aligned} \gamma^* &= \sqrt{\frac{\lambda+\mu}{m\bar{F}\left(\frac{r+\bar{p}}{m}\right)}}\, h(z) \\ \theta^* &= -z\sqrt{\frac{m\bar{F}\left(\frac{r+\bar{p}}{m}\right)}{\lambda+\mu}} - f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\frac{\gamma^*}{\lambda+\mu}. \end{aligned} \tag{4.8}$$

This optimality can be verified using the fact that h is a convex function (proved on page 12 in Zeltyn and Mandelbaum (2005)).

We now characterize all asymptotically optimal sequences in the following theorem.

Proposition 16. *Any sequence* $(\bar{p}+\phi^{n*}, (\bar{k}+\kappa^{n*})n)$ *for which* $\frac{\phi^{n*}}{\mu} f\left(\frac{r+\bar{p}}{m}\right) + \kappa^{n*} = \frac{\theta^*}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$ *is asymptotically optimal.*

The equivalences of corrections in price and capacity levels suggests that the firm could choose to keep one of the decisions at the nominal level and correct the other. This is particularly useful if adjusting one of prices or capacity levels is difficult for the firm. For example, in a DVD rental firm, one expects making changes to capacity levels to be easier than changing prices. Hence, in this scenario the firm can set prices at the nominal level and make corrections in the capacity level in accordance with the equilibrium to achieve the asymptotically optimal solution. However, if changing prices is easier than changing capacity, for example, in a car rental firm, then the firm could do the opposite. This is summarized in the following result.

Corollary 2. *The sequences* $(\bar{p}, (\bar{k}+\frac{\theta^*}{\sqrt{n}})n)$ *and* $\left(\bar{p}+\frac{\theta^*}{\sqrt{n}}\frac{\mu}{f\left(\frac{r+\bar{p}}{m}\right)}, \bar{k}n\right)$ *are asymptotically optimal.*

We now mimic the arguments of this section to solve the firm's optimization problem when offering a pay-per-use option. Readers may choose to skip this section and jump directly to the comparison of the two options in Section 4.3.

4.2 The Pay-per-use Option

We now consider the scenario where the firm offers a pay-per-use option instead of the subscription. Namely, customers now pay each time they use the service. We model the pay-per-use customers by a Poisson process with a potential rate of $\tilde{\Lambda}$. For comparability of this system with the subscription option having n potential subscribers, we shall equate the nominal load in both systems. The nominal load is defined as the number of products that will be in use in steady-state in the system

when all customers join and the system has infinite capacity. For a subscription based system, this can be computed to be $n\frac{\lambda}{\lambda+\mu}$. Hence, we shall set $\tilde{\Lambda} = n\frac{\lambda}{\lambda+\mu}\mu = nm$.

As with the subscribers, the pay-per-use customers are sensitive to both the price and the quality-of-service. In particular, the customers value each service at a random value S_p that is drawn from a distribution identical to that of the subscribers. We further assume that these customers have a reservation value of $\$r_p$ per service. When the system manager sets a pay-per-use price of p_p per use, each customer compares the expected value obtained by joining the system which is the probability of obtaining the product times the value derived from using the product, i.e., $(1-\gamma)(S_p-p_p)$, with her reservation value r_p. Note that we are making this comparison with r_p and not $(1-\gamma)r_p$. The rationale for this is that each customer that joins the system, but does not get the product upon request incurs some cost, and as a means of quantifying this cost we prohibit the customers from deriving their reservation value upon denial.

Thus, the total arrival rate of such customers Λ is given as a function of the pay-per-use price and the denial probability by

$$\begin{aligned}\Lambda(p_p,\gamma) &= nm\mathbb{P}\left((1-\gamma)(S_p-p_p) > r_p\right)\\ &= nm\bar{F}\left(\frac{r_p}{1-\gamma}+p_p\right).\end{aligned}$$

The optimization problem of the system manager can be stated as

$$\begin{aligned}&\max_{(p_p,k_p)\in\mathbb{R}^2_+} \; p_p(1-\gamma)\Lambda(p_p,\gamma) - k_p c\\ &\text{s.t. } \gamma = d(\Lambda(p_p,\gamma),k_p),\end{aligned}$$

where $d(\Lambda(p_p,\gamma),k_p)$ is the denial probability as a function of the arrival rate of the customers and the number of products. As in Section 4.1 we shall approximately solve this problem in the large market regime. We shall let n grow without bound and use the superscript n to make explicit the dependence of the various parameters on n, i.e., p_p^n, k_p^n, γ^n and $\Lambda(p_p^n,\gamma^n)$ denote the pay-per-use fee, the capacity level, the denial probability and the rate of arrival of the pay-per-use customers respectively.

The optimization problem can then be written as

$$\max_{(p_p^n,k_p^n)\in\mathbb{R}_+^2} \Pi^n(p_p^n,k_p^n) \equiv p_p^n(1-\gamma^n)\Lambda^n(p_p^n,\gamma^n) - k_p^n c$$
$$\text{s.t. } \gamma^n = d(\Lambda^n(p_p^n,\gamma^n),k_p^n).$$

With this setup, we try to find a sequence of prices and number of products (p_p^{n*},k_p^{n*}) that is optimal in the limiting regime as $n\to\infty$.

4.2.1 Nominal solution

We begin by solving the nominal problem for this system analogous to (4.3) given by

$$\max_{(p_p,k_p)\in\mathbb{R}_+^2} p_p m\bar{F}(r_p+p_p) - k_p c$$
$$\text{s.t. } k_p = \frac{\lambda}{\lambda+\mu}\bar{F}(r_p+p_p).$$

It is easy to verify using the first order conditions that the nominal price $\bar{p}_p$ satisfies the relation

$$\bar{F}(r_p+\bar{p}_p) = \left(\bar{p}_p - \frac{c}{\mu}\right) f(r_p+\bar{p}_p) \tag{4.9}$$

and the nominal capacity level $\bar{k}_p = \frac{\lambda}{\lambda+\mu}\bar{F}\left(r_p+\bar{p}_p\right)$. We then have the following result.

Proposition 17. *A sequence (p_p^n,k_p^n) is a nominal solution if $\gamma^n\to 0$ and $\left(p_p^n,\frac{k_p^n}{n}\right)\to(\bar{p}_p,\bar{k}_p)$.*

4.2.2 Refining the nominal solution

As before, we define the capacity imbalance in the system as the difference between the number of servers and the offered load in the absence of denials scaled by n. Thus, for any sequence (p_p^n,k_p^n), the capacity imbalance is given by

$$\theta^n = \frac{k_p^n}{n} - \frac{\lambda}{\lambda+\mu}\bar{F}(r_p+p_p^n). \tag{4.10}$$

We now focus on constructing an asymptotically optimal sequence (p_p^n, k_p^n) with the associated denial probability γ^n. An analog to Lemma 17 holds in this setting as well which motivates us to restrict our search to θ^n such that $\lim_{n\to\infty} \theta^n\sqrt{n} = \theta$, where $\theta \in \mathbb{R}$. We choose $p_p^n = \bar{p}_p + \phi_p^n$ and $k_p^n = (\bar{k}_p + \kappa_p^n)n$ such that (p_p^n, k_p^n, θ^n) satisfy (4.10) and $\phi_p^n, \kappa_p^n \to 0$. Further, we rewrite (4.10) using the Taylor series expansion of $\bar{F}\left(\frac{r_p}{1-\gamma^n} + p_p^n\right)$ about $(r_p + \bar{p}_p)$ to obtain

$$\theta^n = \frac{\phi_p^n m}{\mu} f\left(r_p + \bar{p}_p\right) + \kappa_p^n + o(\phi_p^n).$$

The asymptotic denial probability can now be characterized as follows.

Proposition 18. *For a sequence (p_p^n, k_p^n) such that $\theta \equiv \lim_{n\to\infty} \theta^n\sqrt{n}$ exists, $\gamma^n\sqrt{n} \to \gamma$ given by*

$$\gamma = \sqrt{\frac{\mu}{m\bar{F}\left(r_p + \bar{p}_p\right)}} h\left(-\left[\theta + f\left(r_p + \bar{p}_p\right)\frac{m}{\mu}\gamma r_p\right]\sqrt{\frac{\mu}{m\bar{F}\left(r_p + \bar{p}_p\right)}}\right). \tag{4.11}$$

We now show that as in the case of the subscription option, the profit is a function of the capacity imbalance alone, and hence pricing and capacity sizing are equivalent levers for the firm at the $O(\sqrt{n})$ scale.

Proposition 19. *For a sequence (p_p^n, k_p^n) such that $p_p^n = \bar{p}_p + \phi_p^n$, $k_p^n = (\bar{k} + \kappa_p^n)n$ and $\theta \equiv \lim_{n\to\infty} \theta^n\sqrt{n}$ exists, the loss in profit function satisfies*

$$\lim_{n\to\infty} \tilde{\Pi}^n(p_p^n, k_p^n) = \theta c + \left(f\left(r + \bar{p}_p\right) r_p + \bar{F}(r_p + \bar{p}_p)\right) m\bar{p}_p\gamma.$$

This result implies that the asymptotically optimal sequence is characterized by the value of θ that minimizes $\lim_{n\to\infty} \tilde{\Pi}^n(p_p^n, k_p^n)$. Hence, the optimal asymptotic capacity imbalance θ^* solves

$$\begin{aligned} &\min_{\theta\in\mathbb{R}} \ \theta c + \left(f\left(r + \bar{p}_p\right) r_p + \bar{F}(r_p + \bar{p}_p)\right) m\bar{p}_p\gamma \\ &\text{s.t. } (4.11). \end{aligned}$$

As before we use the first order conditions to characterize the optimal solution. Hence,

denoting $z \equiv h'^{-1}\left(\frac{c\bar{F}(r_p+\bar{p}_p)}{(f(r_p+\bar{p}_p)r_p+\bar{F}(r_p+\bar{p}_p))\bar{p}_p\mu - cr_p f(r_p+\bar{p}_p)}\right)$, the optimal denial probability and capacity imbalance are given by

$$\begin{aligned}\gamma^* &= \sqrt{\frac{\mu}{m\bar{F}\left(r_p+\bar{p}_p\right)}}\, h(z) \\ \theta^* &= -z\sqrt{\frac{m\bar{F}\left(r_p+\bar{p}_p\right)}{\mu}} - \frac{f\left(r_p+\bar{p}_p\right)r_p m}{\mu}\gamma^*.\end{aligned} \tag{4.12}$$

We now characterize all asymptotically optimal sequences in the following result.

Proposition 20. *Any sequence* $(\bar{p}_p+\phi_p^{n*}, (\bar{k}_p+\kappa_p^{n*})n)$ *for which* $\frac{\phi_p^{n*}m}{\mu}f(r_p+\bar{p}_p)+\kappa_p^n = \frac{\theta^*}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$ *is asymptotically optimal.*

Having solved the firm's optimization problem for both the subscription and the pay-per-use options, we now turn our attention to their comparison. We will also compare measures like social welfare and consumer surplus in these options.

4.3 Comparison

4.3.1 Quality-of-service: neither option dominates

Before comparing the profits in the two options, we address the issue as to whether it is obvious that locking customers into subscription must be beneficial to the firm. We compute the asymptotic quality-of-service levels (using Proposition 14 and Proposition 18) for both options at different capacity imbalance levels and compare them in Figure 4.1. We set $F(x) = e^{-x}$ for $x > 0$, $r_s = r_p = 0$ (we shall henceforth use the qualifying subscript s for terms pertaining to the subscription option in Section 4.1) and $\lambda = 5$ and $\mu = \frac{5}{4}$ (so that $m = 1$) for these computations. The figure illustrates that neither option dominates the other. In fact, at low capacity imbalance levels, the pay-per-use option offers a higher quality-of-service, while the subscription option is better at higher imbalance levels. Figure 4.1 can be explained loosely as follows. As the number of subscribers that are On increases, the number of subscribers attempting decreases. Consequently, at high capacity levels, the denial probability seen by

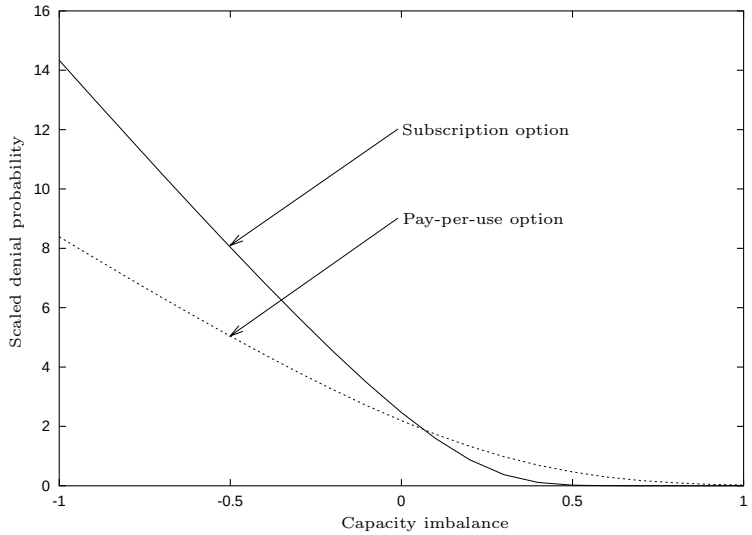


Figure 4.1: Comparison of quality-of-service in the two options

the subscribers is smaller than that seen by the pay-per-use customer stream whose attempt rate is independent of the state of the system. Arguing similarly, the denial probability can be explained to be higher in the subscription option at low capacity levels.

4.3.2 Firm's profits: subscription option always dominates

We now focus on establishing that the firm's profits are higher in the subscription option. We will normalize $r_s = r_p = 0$ and $m = 1$ for convenience. It is worth making a comment about the analysis for general reservation values. Note that r_s is a rate, while r_p is defined per usage. Hence, for these to be comparable we shall require that $r_s = mr_p$.

We shall first compare the nominal profits in the two options. Comparing the first order conditions that characterize the nominal solution (4.4) and (4.9), we observe that $\bar{p}_s = m\bar{p}_p$ and $\bar{k}_s = \bar{k}_p$. Further, the nominal profits are identical. Hence, we shall compare the asymptotic loss in profit in the two options. Noting that h'^{-1} is a function that does not have a closed form, proving such a claim in complete generality is quite difficult. Instead we prove this result for a specific demand function and will

provide numerical results for a few others. Denoting the optimal capacity imbalances in the subscription and pay-per-use options by θ_s^* and θ_p^* respectively, we state the primary result of this chapter.

Proposition 21. *For the exponential demand function with $F(x) = e^{-\alpha x}$ for $x > 0$, for any constant $\alpha > 0$, the firm's asymptotic loss in profit is higher in the pay-per-use option. Further, the capacity imbalance is lower in the subscription option, $\theta^{s*} < \theta^{p*}$.*

Noting that a linear combination of the corrections to the prices and capacity constitutes the capacity imbalance, for a fixed capacity level, low capacity imbalances imply lower prices. Hence, we observe that as firms can lock in subscribers, they have effectively reduced their hedge against variability and can offer the same capacity level at a lower price to obtain a higher profit.

Comparison of firm's profits: numerical experiments

We now provide numerical results that demonstrate that the profits are higher when the firm offers a subscription option for two demand functions: a linear demand function and a Pareto demand function with constant price elasticity.

Linear demand. To obtain a linear demand function, we choose a uniform distribution for the valuations, i.e., $F(x) = x$ for $0 \leq x \leq 1$. We numerically compute the loss in profit for a large number of parameter values. In particular, we vary both λ and the cost per server c, while choosing μ so that $m = 1$. In each instance we observed that the loss in profit as well as the capacity imbalance were lower in the subscription option. As a specific example, Table 4.1 compares the optimal imbalances and the loss in profits for the two options as the cost per server varies with $\lambda = 2$. We observe that the subscription option has a lower capacity imbalance and realizes a lower loss in profit.

Iso-elastic demand. We choose a Pareto distribution for the valuations, namely, $F(x) = 1 - x^{-2}$ for $x \geq 1$ which has a price elasticity of 2. Table 4.2 compares the optimal imbalances and the loss in profits for the two options as λ varies with the

	Subscription		Pay-per-use	
c	θ_s^*	Loss in Profit	θ_p^*	Loss in Profit
0.01	1.070	0.012	1.516	0.017
0.5	0.109	0.263	0.389	0.371
1.0	-0.704	0.298	-0.051	0.422
1.5	-2.766	0.202	-0.396	0.286

Table 4.1: Subscription versus pay-per-use: linear demand

	Subscription		Pay-per-use	
λ	θ_s^*	Loss in Profit	θ_p^*	Loss in Profit
1.1	-0.658	2.400	0.913	2.517
2	-0.208	0.759	0.389	1.073
5	-0.104	0.379	0.308	0.848
10	-0.069	0.253	0.290	0.800

Table 4.2: Subscription versus pay-per-use: Pareto demand

cost per server set at $c = 1$. We again choose μ so that $m = 1$. We observe that the subscription option has a lower capacity imbalance and realizes a lower loss in profit.

4.3.3 Comparison of consumer surplus and social welfare

We shall construct examples for the exponential demand function to show that the consumer surplus and social welfare can be higher or lower in the subscription option as compared to the pay-per-use option.

We shall begin by computing the consumer surplus and social welfare in the two options. We shall normalize $r_s = r_p = 0$ and $m = 1$ as before and denote $\bar{p} \equiv \bar{p}_s = \bar{p}_p$. For a system with n subscribers, the consumer surplus can be written out as

$$\begin{aligned} CS_s^n &= n \int_{\left(p^n\left(1+\frac{\gamma^n}{\lambda}\right)\right)}^{\infty} \bar{F}(p)dp \\ &= \frac{e^{-\alpha\bar{p}}}{\alpha} n - \left(\sqrt{n}\phi_s^{n*} + \sqrt{n}\frac{\gamma^n\bar{p}}{\lambda}\bar{p}\right)\frac{e^{-\alpha\bar{p}}}{\alpha}\sqrt{n}. \end{aligned}$$

Similarly, we can write out the consumer surplus for the pay-per-use customers as

$$\begin{aligned} CS_p^n &= n(1-\gamma^n)\int_{p^n}^{\infty} \bar{F}(p)dp \\ &= \frac{e^{-\alpha\bar{p}}}{\alpha} n - (\sqrt{n}\phi_p^{n*} + \sqrt{n}\gamma^n)\frac{e^{-\alpha\bar{p}}}{\alpha}\sqrt{n}. \end{aligned}$$

Hence,

$$\lim_{n\to\infty} \frac{CS_s^n}{n} = \lim_{n\to\infty} \frac{CS_p^n}{n} = \frac{e^{-\alpha\bar{p}}}{\alpha} \equiv \overline{CS}.$$

As the firm's profit can be completely characterized by the value of the capacity imbalance θ, using Propositions 16 and 20 the firm can choose to set any value of ϕ_i^{n*}, $i = s, p$, and adjust the value of κ^{n*} to obtain the same profit. Hence, denoting the loss in surplus (due to variability) from the nominal value as $\widehat{CS}_i = \lim_{n\to\infty} \sqrt{n}\left(\overline{CS} - \frac{CS_i^n}{n}\right)$, we see that the firm can decrease $\widehat{CS}_i$ in an unbounded fashion by increasing ϕ_i^* arbitrarily. Hence, the customers obtain a higher consumer surplus when the firm sets low prices. This result can be explained in a different manner. Note that if the firm sets any particular price, it can adjust the capacity level so that the capacity imbalance, and hence the quality-of-service remains unchanged. Thus, reducing prices while maintaining the same quality-of-service increases the consumer surplus. Nevertheless, we shall compare the consumer surplus at fixed price levels. In addition, we shall compare the social welfare, which in this case is the sum of the consumer surplus and the firm's profits. Table 4.3 displays the losses in consumer surplus and social welfare for different values of the parameter c at fixed $\lambda = 2$, prices set at the nominal values, and $\alpha = 1$. We observe that for low values of c, the loss in these measures are lower in the subscription option, while this is reversed for high values of c.

For completeness, we shall solve the firm's profit maximization problem when it offers both the subscription and pay-per-use options.

	Subscription		Pay-per-use	
c	Loss in CS	Loss in SW	Loss in CS	Loss in SW
0.1	0.012	0.097	0.016	0.136
1	0.172	0.611	0.163	0.783
5	0.647	1.123	0.261	0.935
10	0.530	0.735	0.125	0.415

Table 4.3: Comparison of consumer surplus and social welfare

4.4 Offering Both Subscription and Pay-per-use Options

We now consider the scenario when the firm offers both the options and wishes to maximize the profits. We shall assume the demand functions $N(p_s, \gamma)$ and $\Lambda(p_p, \gamma)$ associated with the subscribers and the pay-per-use customers respectively. We shall assume that the valuation of the pay-per-use customers is drawn from a distribution G, which may be different from F, with density g. The optimization problem of the system manager is

$$\begin{aligned} &\max_{(p_s,p_p,k)\in\mathbb{R}^3_+} \; p_s N(p_s,\gamma) + p_p(1-\gamma)\Lambda(p_p,\gamma) - kc \\ &\text{s.t. } \gamma = d(N(p_s,\gamma),\Lambda(p_p,\gamma),k), \end{aligned}$$

where $d(N(p_s, \gamma), \Lambda(p_p, \gamma), k)$ is the denial probability as a function of the number of subscribers, arrival rate of the pay-per-use customers, and the number of products. As in Section 4.1 we shall approximately solve this problem in the large market regime. We shall let n grow without bound and use the superscript n to make explicit the dependence of the various parameters on n, i.e., p_s^n, p_p^n k^n, γ^n, $N^n(p_s^n, \gamma^n)$ and $\Lambda(p_p^n, \gamma^n)$ denote the subscription fee, pay-per-use fee, the capacity level, the denial probability, the number of subscribers and the rate of arrival of the pay-per-use customers respectively. The optimization problem can then be written as

$$\begin{aligned} &\max_{(p_s^n,p_p^n,k^n)\in\mathbb{R}^3_+} \Pi^n(p_s^n,p_p^n,k^n) \equiv p_s^n N^n(p_s^n,\gamma^n) + p_p^n(1-\gamma^n)\Lambda^n(p_p^n,\gamma^n) - k^n c \\ &\text{s.t. } \gamma^n = d(N^n(p_s^n,\gamma^n),\Lambda^n(p_p^n,\gamma^n),k^n). \end{aligned}$$

With this setup, we try to find a sequence of prices, both subscription and pay-per-use, and number of products, $(p_s^{n*}, p_p^{n*}, k^{n*})$, that is optimal in the limiting regime as $n \to \infty$.

4.4.1 Nominal solution

We begin by solving the nominal problem for this system analogous to (4.3) given by

$$\begin{aligned} \max_{(p_s,p_p,k)\in\mathbb{R}^3_+} & \; p_s \bar{F}\left(\frac{r_s+p_s}{m}\right) + p_p m \bar{G}(r_p+p_p) - kc \\ \text{s.t. } & k = \frac{\lambda}{\lambda+\mu}\left(\bar{F}\left(\frac{r_s+p_s}{m}\right) + \bar{G}(r_p+p_p)\right). \end{aligned}$$

It is easy to verify using the first order conditions that the optimal prices $\bar{p}_s, \bar{p}_p$ satisfy the equations

$$\begin{aligned} \bar{F}\left(\frac{r_s+\bar{p}_s}{m}\right) &= \frac{1}{m}\left(\bar{p}_s - \frac{\lambda}{\lambda+\mu}c\right) f\left(\frac{r_s+\bar{p}_s}{m}\right), \\ \bar{G}(r_p+\bar{p}_p) &= \left(\bar{p}_p - \frac{c}{\mu}\right) g(r_p+\bar{p}_p), \end{aligned}$$

and the optimal number of products $\bar{k} = \frac{\lambda}{\lambda+\mu}\left(\bar{F}\left(\frac{r_s+\bar{p}_s}{m}\right) + \bar{G}(r_p+\bar{p}_p)\right)$. We then have the following result.

Proposition 22. *A sequence (p_s^n, p_p^n, k^n) is a nominal solution if*

$$\gamma^n \to 0 \quad \textit{and} \quad \left(p_s^n, p_p^n, \frac{k^n}{n}\right) \to (\bar{p}_s, \bar{p}_p, \bar{k}).$$

4.4.2 Refining the nominal solution

As before, we define the capacity imbalance in the system as the difference between the number of servers and the offered load in the absence of denials scaled by n. Thus, for any sequence (p_s^n, p_p^n, k^n), the capacity imbalance is given by

$$\theta^n = \frac{k^n}{n} - \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r_s+p_s^n}{m}\right) - \frac{\lambda}{\lambda+\mu}\bar{G}(r_p+p_p^n). \tag{4.13}$$

We now focus on constructing an asymptotically optimal sequence (p_s^n, p_p^n, k^n) with the associated denial probability γ^n. An analog to Lemma 17 holds in this setting as well which motivates us to restrict our search to θ^n such that $\lim_{n\to\infty} \theta^n\sqrt{n} = \theta$, where $\theta \in \mathbb{R}$. We choose $p_s^n = \bar{p}_s + \phi_s^n$, $p_p^n = \bar{p}_p + \phi_p^n$ and $k^n = (\bar{k} + \kappa^n)n$ such that $(p_s^n, p_p^n, k^n, \theta^n)$ satisfy (4.13) and $\phi_s^n, \phi_p^n, \kappa^n \to 0$. Further, we rewrite (4.13) using the Taylor series expansions of $\bar{F}\left(\frac{r_s+p_s^n}{m}\right)$ about $\left(\frac{r_s+\bar{p}_s}{m}\right)$ and $\bar{G}(r_p + p_p^n)$ about $r_p + \bar{p}_p$ to obtain the following.

$$\theta^n = \frac{\lambda}{\lambda+\mu}\left(\frac{\phi_s^n}{m} f\left(\frac{r_s+\bar{p}_s}{m}\right) + \phi_p^n g(r_p+\bar{p}_p)\right) + \kappa^n.$$

The asymptotic denial probability can now be characterized as follows.

Proposition 23. *For a sequence (p_s^n, p_p^n, k^n) such that $\theta \equiv \lim_{n\to\infty} \theta^n\sqrt{n}$ exists, we have*

$$\gamma^n\sqrt{n} \to \gamma = \frac{h\left(-\frac{\theta + f\left(\frac{r_s+\bar{p}_s}{m}\right)(r_s+\bar{p}_s)\frac{\gamma}{\lambda+\mu} + \frac{m}{\mu}g(r_p+\bar{p}_p)\gamma r_p}{\sqrt{\frac{m\bar{F}\left(\frac{r_s+\bar{p}_s}{m}\right)}{\lambda+\mu} + \frac{m\bar{G}(r_p+\bar{p}_p)}{\mu}}}\right)}{\sqrt{\frac{m\bar{F}\left(\frac{r_s+\bar{p}_s}{m}\right)}{\lambda+\mu} + \frac{m\bar{G}(r_p+\bar{p}_p)}{\mu}}}. \tag{4.14}$$

We now show that as in the case of the subscription and pay-per-use options, the profit is a function of the capacity imbalance alone, and hence pricing and capacity sizing are equivalent levers for the firm at the $O(\sqrt{n})$ scale.

Proposition 24. *For a sequence (p_s^n, p_p^n, k^n) such that $p_s^n = \bar{p}_s + \phi_s^n$, $p_p^n = \bar{p}_p + \phi_p^n$, $k^n = (\bar{k} + \kappa^n)n$ and $\theta \equiv \lim_{n\to\infty} \theta^n\sqrt{n}$ exists, the loss in profit function satisfies*

$$\lim_{n\to\infty} \tilde{\Pi}(p_s^n, p_p^n, k^n) = -c\theta - \left(\frac{(r_s+\bar{p}_s)}{\lambda}\bar{p}_s f\left(\frac{r_s+\bar{p}_s}{m}\right) + mg(r_p+\bar{p}_p)\bar{p}_p r_p + m\bar{p}_p\bar{G}(r_p+\bar{p}_p)\right)\gamma.$$

Hence, the asymptotically diffusion optimal sequence $(p_s^{n*}, p_p^{n*}, k^n)$ is characterized

by the optimal capacity imbalance θ^* that solves the following problem.

$$\min_{\theta\in\mathbb{R}} c\theta + \left(\frac{(r_s+\bar{p}_s)}{\lambda}\bar{p}_s f\left(\frac{r_s+\bar{p}_s}{m}\right) + mg(r_p+\bar{p}_p)\bar{p}_p r_p + m\bar{p}_p\bar{G}(r_p+\bar{p}_p)\right)\gamma \quad (4.15)$$
s.t (4.14).

We now characterize all asymptotically optimal sequences in the following result.

Proposition 25. *Any sequence* $(\bar{p}_s+\phi_s^{n*}, \bar{p}_p+\phi_p^{n*}, (\bar{k}+\kappa^{n*})n)$ *for which*

$$\frac{\lambda}{\lambda+\mu}\left(\frac{\phi_s^{n*}}{m}f\left(\frac{r_s+\bar{p}}{m}\right) + \phi_p^{n*}g(r_p+\bar{p}_p)\right) + \kappa^n = \frac{\theta^*}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

is asymptotically optimal, where θ^* *is the solution to* (4.15).

4.5 Discussion

This chapter provides a technique for computing prices and capacity levels that approximately maximize a large rental firm's profit. This technique is used to characterize the asymptotically optimal solution in a subscription and a pay-per-use environment and it is shown that the firm's profit is higher in the subscription setting.

We have only considered the special case of the subscriber's retrial rate, i.e., $\nu = \lambda$, in this chapter. The general case is quite difficult to analyze, even asymptotically (see Section 3.1). However, we can use numerical experiments to perform comparative statics on the retrial rate. In particular, as Table 4.4 demonstrates, for a linear demand function with $r_s = 0$ and cost per server $c = 1$, increasing the retrial rate leads to an increase in the firm's profits. Although increasing the retrial rate increases the denial probability, it also enables the subscribers to obtain service quicker, and hence leads to higher valuations. It is this latter effect that dominates and allows the firm to extract higher profits. An analytical analysis of this phenomenon is a topic for future work.

In this chapter, we only discussed static pricing and capacity sizing. One can envisage dynamic admission control as in Savin, Cohen, Gans and Katalan (2005)

		Retrial rate, ν			
		0.1	1	2	4
θ	-0.5	0.548 ± 0.014	0.329 ± 0.012	0.289 ± 0.008	0.287 ± 0.013
	0	0.594 ± 0.014	0.388 ± 0.010	0.340 ± 0.006	0.318± 0.004
	0.5	0.686 ± 0.011	0.566 ± 0.003	0.544 ± 0.002	0.529 ± 0.002

Table 4.4: Firm's loss in profits on the $O(\sqrt{n})$ scale with 95% confidence intervals for general retrial rates

and dynamic pricing in the spirit of Maglaras (2003) and Çelik and Maglaras (2005). These are topics for future work.

As a concluding note, we would like to comment on the distributional assumptions made in this chapter. Although, we have made Markovian assumptions, the steady-state distribution of the number-in-system for the subscription option depends on the distribution of On and Off times only through the means of these distributions, while for the pay-per-use option, the steady-state distribution of the number-in-system depends on the Poisson arrival assumption and the mean of the service time (see Section 3.1.1). In Section 5.2, we shall show this insensitivity of the steady-state distributions carries through to the denial probabilities.

Chapter 5

Some Numerical Results and Future Work

In this chapter our goal is to raise a bunch of questions that we have not answered theoretically and try to answer them using numerical experiments. Answering these questions theoretically is left for future work. All our questions are based on the basic subscriber model of Chapter 2.

The first questions is when can we use the limit theory developed in this thesis, or in other words, how large should the system in question be so that the asymptotic limits derived can be applied to give reasonable estimates for this system? We tackle this question in Section 5.1, where we show that the asymptotic estimates are valid even for fairly small systems.

The next question is about the distributional assumptions made in the thesis. Though the process level convergences depends heavily on the Markovian assumptions, we have seen that for the special case when the Off and Hold states are indistinguishable, the steady-state distribution depends only on the means of the distributions, and not on any other moments. This itself is not sufficient for the denial rates, and in turn the denial probabilities, to possess a similar insensitivity. In Section 5.2, we show that though the denial rate seems to depend only on the means of the distributions for the special case, this is not true for the general case when the Off and Hold states are distinguishable.

The third question that we raise is about the loss system model. As there might be systems where a queueing model is more appropriate, can our results be applied to these settings? We note that as the retrial rate increases without bound our loss system approaches a queueing system. In fact in Section 5.3 we show that this gap is bridged for fairly low retrial rates, leading us to believe that our analysis can be used to estimate parameters of interest for a queueing system as well.

Finally, we question the Poisson process approximation for arriving customers. One expects that when there are a large number of subscribers, each of whom have an extremely large mean time in the Off state, the resulting attempt process should look like a Poisson process. In Section 5.4, we use simulations to show that for such systems, the subscriber model asymptotic results yield better approximations than those using the Poisson assumption.

5.1 Rate of Convergence

In this thesis, we have seen convergences of processes and their invariant measures. However, nothing has been mentioned as to how quickly this convergence is observed, namely, how large (or how small) should n be for these asymptotic results to be valid? We again use simulation to demonstrate that these asymptotic results are valid for fairly small n values. To do so, we set $\nu = \lambda = 1$ and $\mu = 2$. We vary n from 10 to $1,000$ and compute the denial rate scaled by $\sqrt{n}$. We set the number of servers $k^n = \lfloor \frac{\lambda}{\lambda+\mu} n \rfloor = \lfloor n/3 \rfloor$ and the simulation run length to $10,000$ time units. The denial rates along with their 95% intervals are displayed in Table 5.1. For this choice of parameters, Proposition 6(a) yields that the scaled denial rate converges to $\sqrt{2}h(0) = 1.128$.

Observe that for $n = 10$ the asymptotic approximation is about 9% away from the true values, while for $n = 100$ this figure is about 3.5%, and it further reduces to 1% for $n = 1000$. This provides a basis for using the diffusion approximations even when n is fairly small.

Number of Subscribers, n			
10	100	1000	∞
1.030 ± 0.020	1.089 ± 0.014	1.116 ± 0.0155	1.128

Table 5.1: Scaled denial rates

		On Time			
		$c_v = 0$	$c_v = 0.1$	$c_v = 1$	$c_v = 10$
Off Time	$c_v = 0$	0 ± 0	0.739 ± 0.024	0.737 ± 0.007	0.742 ± 0.075
	$c_v = 0.1$	0.739 ± 0.011	0.728 ± 0.132	0.740 ± 0.006	0.729 ± 0.070
	$c_v = 1$	0.738 ± 0.007	0.738 ± 0.007	0.739 ± 0.008	0.726 ± 0.054
	$c_v = 10$	0.755 ± 0.083	0.739 ± 0.126	0.762 ± 0.118	0.731 ± 0.139

Table 5.2: Denial rates for subscribers with general holding times

5.2 Relaxing Markovian Assumptions

We know that for the special case when the Hold and Off states are indistinguishable for the subscribers, the steady-state distribution of the number-in-system process depends only on the means of the On and Off time distributions. This is proved in Cohen (1957) and can also be derived from Proposition 9 in Chapter 3. This result immediately implies that the denial rate and the denial probability also depend only on the means of these distributions when the On times are exponentially distributed. We conjecture that this result holds for general On time distributions that have densities as well, and we illustrate this via a simulation study.

We choose four different distributions with the coefficient of variant varying from 0 to 10. In each case we set the Hold and Off time distributions identical. We set the mean Off time at 1 and On time at 0.5. The four distributions that we use for the Off time are: (a) Deterministic, coefficient of variation (c_v)=0; (b) Erlang($n = 100, 2$), $c_v = 0.1$; (c) Exponential, $c_v = 1$; and (d) Hyper-exponential: distributed as Exponential(100) with probability $p = 0.9902$ and Exponential(0.009898) with probability $1-p$, $c_v = 10$. The four distributions that we use for the On time are: (a) Deterministic, $c_v = 0$; (b) Erlang($n = 100$, 2), $c_v = 0.1$; (c) Exponential, $c_v = 1$; and (d) Hyper-exponential: distributed as Exponential(100) with probability $p = 0.9904$ and Exponential(0.0196) with probability $1 - p$, $c_v = 10$.

		ν=10	ν =100	ν=1000	Queue
	10	8.033 ± 0.094	8.010 ± 0.0622	7.998 ± 0.06	8.000 ± 0.015
n	100	20.838 ± 0.551	20.065 ± 0.345	20.092 ± 0.338	20.0213 ± 0.129
	1000	67.068 ± 1.092	64.236 ± 1.078	63.922 ± 1.99	63.980 ± 0.455

Table 5.3: Using increasing retrial rates to model a queue

The length of the simulation is set at $100,000$ time units, which is sufficiently long to observe the steady-state behavior. We run simulations for systems with $n = 10$ subscribers and 5 servers, and the denial rates for each case are reported in Table 5.2 along with 95% confidence intervals based on 20 independent simulation runs. We can see that other than the completely deterministic case, where the On and Off times line up so that there are no denied attempts, the denial rates obtained are quite insensitive to the actual distribution.

5.3 Increasing Retrial Rates: A Queueing Limit

If the retrial rates increase in an unbounded fashion, we expect the system to begin to behave more like a queueing system, where upon being denied service people wait in a buffer instead of retrying. In fact, we can argue that in a system with retrial rate ν the denial rate, the rate at which customers are turned away or denied service, can be written as $\beta^n(\nu) = (n - k^n)\lambda\pi^n(Q_1^n(t) = k^n) + (\nu - \lambda)\mathbb{E}_{\pi^n}[Q_2^n(t)|Q_1^n(t) = k^n]$, where $Q_1^n(t)$ and $Q_2^n(t)$ are the number of subscribers in service and waiting to retry respectively, at time t, π^n is the stationary measure and $\mathbb{E}_{\pi^n}$ refers to the expectation with respect to the stationary measure. Hence, we obtain $\lim_{\nu\to\infty} \frac{\beta^n(\nu)}{\nu} = \lim_{\nu\to\infty} \mathbb{E}_{\pi^n}[Q_2^n|Q_1^n = k^n]$, which can be loosely argued to equal the expected queue length in the queueing model. We verify this convergence via simulation. We set $\lambda = \mu = 2$ and compute $\frac{\beta^n(\nu)}{\nu}$ for different values of ν and compare this with the expected queue length estimates for the queueing model. We perform these experiments for different values of n. The results displayed in Table 5.3 suggest the conjectured convergence. In fact, even for small values of the retrial rate, the ratio of the denial rate estimate to the retrial rate is quite close to the expected queue length estimate.

5.4 The Subscriber Model versus the Poisson Assumption

The justification for a Poisson arrival model stems from the Palm-Khintchine theory (see Heyman and Sobel (1982)) which implies that the superposition of many renewal processes, each with rate diminishing to zero, results in a Poisson process. This suggests that if for an infinite server system, the number of subscribers n grows without bound, while the rate at which each subscriber attempts for a server λ_n converges to 0 such that $n\lambda_n \to C$, where C is a constant, then the arrival process generated by the n subscribers converges to a Poisson process of rate C.

It can be shown for a finite server loss system that if $n\lambda_n = O(n^x)$, for some $0 < x < 1$ and the number of servers $k^n = k_1 n^x + k_2 n^{\frac{x}{2}}$, then denoting the number of servers in use at time t by $Q^n(t)$, we obtain the following asymptotic results.

Proposition 26. *If $n^{1-x}\lambda_n \to C$ for some $C > 0$, $0 < x < 1$, and $\frac{Q^n(0)}{n^x} \to \bar{q} = \min\left(k_1, \frac{C}{\mu}\right)$ a.s., then*

(a) $\frac{Q^n(\cdot)}{n^x} \to \bar{q}$.

(b) If $k_1 = \frac{C}{\mu}$ and $n^{-\frac{x}{2}}(Q^n(0) - k^n) \Rightarrow \hat{Q}(0)$, then $n^{-\frac{x}{2}}(Q^n(\cdot) - k^n) \Rightarrow \hat{Q}(\cdot)$, where $\hat{Q}(t)$ is given by

$$\hat{Q}(t) = \hat{Q}(0) - \mu \int_0^t (\hat{Q}(t) + k_2)dt + \sqrt{2C}B(t) - Y(t),$$

where B is a standard Brownian motion and $Y(\cdot)$ is the non-decreasing, non-negative process such that $\int_0^t \hat{Q}(u)dY(u) = 0$ for all $t \geq 0$ and $Y(0) = 0$.

(c) The denial rate converges as

$$\lim_{n\to\infty} n^{-\frac{x}{2}} D^n(k^n) = \sqrt{\mu C}\ h\left(-k_2\sqrt{\frac{\mu}{C}}\right).$$

We now construct systems with large number of subscribers n, each with small

n	λ	Servers	Simulation	Subscriber	Poisson
		10	13.81 ± 0.10	15.24	15.51
$1,000$	0.032	15	6.51 ± 0.13	7.38	7.71
		20	1.88 ± 0.06	2.07	2.30
		30	15.08 ± 0.20	16.27	16.60
5,000	0.014	35	8.37 ± 0.20	9.13	9.44
		40	3.60 ± 0.11	3.89	4.12
		40	24.63 ± 0.22	26.05	26.39
10,000	0.01	50	10.18 ± 0.22	10.97	11.28
		60	2.05 ± 0.09	2.11	2.25
		140	48.83 ± 0.51	50.41	50.80
100,000	0.0032	160	19.01 ± 0.38	19.86	20.19
		170	2.91 ± 0.16	2.95	3.07

Table 5.4: Comparison of denial rate approximations

attempt rates λ_n and the retrial rate $\nu_n = \lambda_n$, and compare the asymptotic approximations of the denial rates suggested by the subscriber model and the Poisson model to the true values, which we obtain via simulation. The results for $\mu = 2$ and $\lambda_n = \frac{1}{\sqrt{n}}$ are summarized in Table 5.4. The simulation results are presented along with 95% confidence intervals. The Poisson approximations are made using $x = 0.5$. These results demonstrate that the subscriber model provides a better approximation of the denial rates. Thus, by modeling customers as subscribers, we can improve the resolution of the performance measures without losing tractability.

Appendix A

Proofs of Results in Chapter 2

The main convergence results of this chapter, namely Propositions 1, 2, and 4 are special cases of Theorem 1 of Chapter 3 and are omitted for brevity. Proposition 4(a) and (b) are also proved, albeit in a different manner, in Theorem 4.1 in Srikant and Whitt (1996).

Proof of Lemma 2. We associate $N(\cdot)$ with a standard Brownian motion $B(\cdot)$ such that the strong approximation result in (1.1) of Lemma 1 holds, which suggests the following.

$$A^n(t) = \lambda \int_0^t (n - Q^n(s))ds + B\left(\lambda \int_0^t (n - Q^n(s))ds\right) + O\left(\log\left(2 \vee \lambda \int_0^t (n - Q^n(s))ds\right)\right).$$

Using Proposition 2(a) in addition to this relation implies

$$\lim_{n\to\infty} \frac{A^n(\cdot)}{n} \to m\cdot,$$

and further, using the weak convergence result in Proposition 2(b), we obtain

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{A^n(\cdot)}{n} - m\cdot\right) \Rightarrow -\lambda \int_0^{\cdot} (\hat{Q}(s) + k_2)ds + B\left(m\cdot\right).$$

■

Proof of Proposition 6. We shall only prove (a) as the proof of (b) follows in an identical fashion. Let us first obtain an expression for $D^n(k^n)$. To do so, we write $Q^n(t)$, $t \geq 0$, as follows

$$Q^n(t) = Q^n(0) + N^a\left(\lambda \int_0^t (n - Q^n(s))ds\right) - N^d\left(\mu \int_0^t Q^n(s)ds\right) - Y^n(t), \quad \text{(A.1)}$$

where

$$Y^n(\cdot) = N^a\left(\lambda \int_0^t (n - Q^n(s))ds\right) - N^a\left(\lambda \int_0^t 1_{\{Q^n(s)<k^n\}}(n - Q^n(s))ds\right)$$

counts the number of denied attempts. Denoting the invariant distribution of Q^n by π^n, we have

$$\begin{aligned}\lim_{t\to\infty} \frac{N^a\left(\lambda \int_0^t (n - Q^n(s))ds\right)}{t} &= \lambda\left(n - \frac{1}{t}\int_0^t Q^n(s)ds\right) \\ &= \lambda\left(n - \mathbb{E}_{\pi^n} Q^n(0)\right),\end{aligned} \quad \text{(A.2)}$$

where we use the fact that $Q^n(\cdot)$ is a positive recurrent Markov chain on a finite state-space, and hence ergodic. Similarly,

$$\lim_{t\to\infty} \frac{N^d\left(\int_0^t \mu Q^n(s)ds\right)}{t} = \mu \mathbb{E}_{\pi^n} Q^n(0). \quad \text{(A.3)}$$

Using (A.2) and (A.3) in (A.1), we obtain

$$D^n(k^n) = \lim_{t\to\infty} \frac{Y^n(t)}{t} = \lambda n - (\lambda + \mu)\mathbb{E}_{\pi^n} Q^n(0).$$

Hence, $D^n(k^n)/\sqrt{n} = -(\lambda + \mu)\mathbb{E}_{\pi^n}(\hat{Q}^n(0) + k_2)$. We now use Proposition 12 from Section 3.3 that proves that $\mathbb{E}_{\hat{\pi}^n}\hat{Q}^n(0) \to \mathbb{E}_{\hat{\pi}}\hat{Q}(0)$ to obtain

$$\frac{D^n(k_n)}{\sqrt{n}} \to -(\lambda + \mu)\mathbb{E}_{\hat{\pi}}(\hat{Q}(0) + k_2).$$

Using Proposition 3 we can compute

$$\mathbb{E}_{\hat{\pi}}(\hat{Q}(0)) = -k_2 - \sqrt{\frac{m}{\lambda+\mu}}\; h\left(-k_2\sqrt{\frac{\lambda+\mu}{m}}\right),$$

which implies that

$$\frac{D^n(k_n)}{\sqrt{n}} \to \sqrt{\lambda\mu}\; h\left(-k_2\sqrt{\frac{\lambda+\mu}{m}}\right).$$

■

Appendix B

Proofs of Results in Chapter 3

Proof of Lemma 3. For convenience we shall fix $n > 0$ and drop it from the notation. We shall argue on the same lines as in Theorem 9.2 in Mandelbaum, Massey and Reiman (1998). Let $\bar{Q}^r = \{\bar{Q}^r(t) : t \geq 0\}$ for $r = 0, 1, \ldots$, where $\bar{Q}^0(t) = Q(0)$ for $t \geq 0$ and for $r > 0$, $\bar{Q}^r$ is given by

$$\begin{aligned}\bar{Q}^r(t) \equiv & Q(0) + N^a\left(\int_0^{t\wedge T^r} 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\lambda(\bar{Q}^{r-1}(u))du\right) \\ & - N^d\left(\int_0^{t\wedge T^r}\mu(\bar{Q}^{r-1}(u))\,du\right) \\ & + N^r\left(\int_0^{t\wedge T^r} 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\nu\bar{Q}_3^{r-1}(u)\,du\right)(1,0,-1)' \\ & + \left(0,0,\Delta N_1^a(\bar{Q}_1^{r-1},\bar{Q}_2^{r-1},\bar{Q}_3^{r-1})(t\wedge T^r)\right)',\end{aligned} \tag{B.1}$$

where the time for r events to occur T^r is given by

$$\begin{aligned}T^r = \inf\Bigg\{ & t : N^a\left(\int_0^t 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\lambda(\bar{Q}^{r-1}(u))\,du\right)' e \\ & + N^d\left(\int_0^t\mu(\bar{Q}^{r-1}(u))\,du\right)' e + N^r\left(\int_0^t 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\nu\bar{Q}_3^{r-1}(u)\,du\right) \\ & + \Delta N_1^a(\bar{Q}_1^{r-1},\bar{Q}_2^{r-1},\bar{Q}_3^{r-1})(t) = r\Bigg\},\end{aligned}$$

where $e = (1, 1, 1)'$. Note that given $\bar{Q}^{r-1}$, $\bar{Q}^r$ is well defined for each r.

To complete the proof, we only need to show

(a) $\bar{Q}^r(t) = \bar{Q}^{r-1}(t)$ for all $0 \leq t < T^r$.

(b) $\lim_{r\to\infty} T^r = \infty$ a.s.

The solution to (3.1) Q can be constructed by setting

$$Q(t) = \bar{Q}^{r-1}(t) \text{ for all } 0 \leq t < T^r.$$

Uniqueness then follows by using induction on r and noting that (B.1) implies that the uniqueness of $\bar{Q}^r$ follows from uniqueness of $\bar{Q}^{r-1}$.

Note that the second claim follows trivially as $\bar{Q}^r$ is bounded. We prove the first claim using induction on r. The case $r = 1$ holds trivially as $\bar{Q}^1(t) = \bar{Q}^0(t) = Q(0)$ for $t < T^1$. Now assume that $\bar{Q}^r(t) = \bar{Q}^{r-1}(t)$ for all $0 \leq t < T^r$ for some $r \in \mathbb{N}$, then

$$\begin{aligned}
N^a\left(\int_0^{t\wedge T^r} 1_{\{\bar{Q}_1^r(u)+\bar{Q}_2^r(u)<k\}}\lambda(\bar{Q}^r(u))\,du\right) &= N^a\left(\int_0^{t\wedge T^r} 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\lambda(\bar{Q}^{r-1}(u))\,du\right)\\
N^d\left(\int_0^{t\wedge T^r} \mu(\bar{Q}^r(u))\,du\right) &= N^d\left(\int_0^{t\wedge T^r} \mu(\bar{Q}^{r-1}(u))\,du\right)\\
N^r\left(\int_0^{t\wedge T^r} 1_{\{\bar{Q}_1^r(u)+\bar{Q}_2^r(u)<k\}}\nu\bar{Q}_3^r(u)\,du\right) &= N^r\left(\int_0^{t\wedge T^r} 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\nu\bar{Q}_3^{r-1}(u)\,du\right),\\
\Delta N_1^a(\bar{Q}_1^r, \bar{Q}_2^r, \bar{Q}_3^r)(t\wedge T^r) &= \Delta N_1^a(\bar{Q}_1^{r-1}, \bar{Q}_2^{r-1}, \bar{Q}_3^{r-1})(t\wedge T^r)
\end{aligned}$$

for $0 \leq t \leq T^r$, which implies that $\bar{Q}^{r+1}(t) = \bar{Q}^r(t)$ for $0 \leq t \leq T^r$.

For $T^r \leq t < T^{r+1}$,

$$\begin{aligned}
N^a\left(\int_0^{t} 1_{\{\bar{Q}_1^r(u)+\bar{Q}_2^r(u)<k\}}\lambda(\bar{Q}_1^r(u))\,du\right) &= N^a\left(\int_0^{T^r} 1_{\{\bar{Q}_1^r(u)+\bar{Q}_2^r(u)<k\}}\lambda(\bar{Q}_1^r(u))\,du\right)\\
&= N^a\left(\int_0^{T^r} 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\lambda(\bar{Q}_1^{r-1}(u))\,du\right)\\
N^d\left(\int_0^{t} \mu(\bar{Q}^r(u))\,du\right) = N^d\left(\int_0^{T^r} \mu(\bar{Q}^r(u))\,du\right) &= N^d\left(\int_0^{T^r} \mu(\bar{Q}^{r-1}(u))\,du\right).
\end{aligned}$$

Similarly,

$$N^r\left(\int_0^t 1_{\{\bar{Q}_1^r(u)+\bar{Q}_2^r(u)<k\}}\nu\bar{Q}_3^r(u)\,du\right) = N^r\left(\int_0^{t\wedge T^r} 1_{\{\bar{Q}_1^{r-1}(u)+\bar{Q}_2^{r-1}(u)<k\}}\nu\bar{Q}_3^{r-1}(u)\,du\right),$$

$$\Delta N_1^a(\bar{Q}_1^r,\bar{Q}_2^r,\bar{Q}_3^r)(t) = \Delta N_1^a(\bar{Q}_1^{r-1},\bar{Q}_2^{r-1},\bar{Q}_3^{r-1})(t\wedge T^r)$$

Hence, $\bar{Q}^{r+1}(t) = \bar{Q}^r(t) = \bar{Q}^r(T^r)$ for $T^r \le t < T^{r+1}$. Therefore, $\bar{Q}^{r+1}(t) = \bar{Q}^r(t)$ for $0 \le t < T^{r+1}$ and the inductive hypothesis holds. ■

Proof of Lemma 4(a). We shall prove the result for the case $k_1 = \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$; the proof for the case $k_1 > \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$ follows in a similar fashion.

Using Lemma 7 and the fact that $\bar{q}(\cdot) = (\frac{\lambda}{\lambda+\mu}, \frac{\lambda_1}{\mu}, 0)'$ and $k_1 = \frac{\lambda}{\lambda+\mu} + \frac{\lambda_1}{\mu}$, we obtain

$$\begin{aligned}
&(q^n(0) - \bar{q}(0)) + \left(0, \frac{\lambda_2}{\sqrt{n}}t, 0\right)' - (1,1,0)' \sup_{0\le s\le t}(X_1^n(s) + X_2^n(s) - (\bar{q}_1(0) + \bar{q}_2(0)))^+ \\
&+ \alpha_n(t) + \delta_n(t) + \left(\int_0^t ((\lambda+\mu)(\bar{q}_1(u) - q_1^n(u)) + (\nu-\lambda)q_3^n(u))du, 0, 0\right)' \\
&\left(0, \mu\int_0^t(\bar{q}_2(u) - q_2^n(u))du, -\nu\int_0^t q_3^n(u)du\right)' \\
&\le (q^n(t) - \bar{q}(t)) \\
&\le (q^n(0) - \bar{q}(0)) + \left(0, \frac{\lambda_2}{\sqrt{n}}t, 0\right)' + (0,0,1)' \sup_{0\le s\le t}(X_1^n(s) + X_2^n(s) - (\bar{q}_1(0) + \bar{q}_2)(0))^+ \\
&+ \alpha_n(t) + \delta_n(t) + \left(\int_0^t ((\lambda+\mu)(\bar{q}_1(u) - q_1^n(u)) + (\nu-\lambda)q_3^n(u))du, 0, 0\right)' \\
&+ \left(0, \mu\int_0^t(\bar{q}_2(u) - q_2^n(u))du, -\nu\int_0^t q_3^n(u)du\right)'.
\end{aligned}$$

This suggests the following relation.

$$\begin{aligned}
\|\, q^n - \bar{q}\,\|_t \le |q^n(0) - \bar{q}(0)| + \max(\lambda+\mu, \nu, |\nu-\lambda|)\int_0^t \|\, q^n - \bar{q}\,\|_u du + \frac{\lambda_2}{\sqrt{n}}t \\
+ \|\,\alpha^n\,\|_t + \|\,\delta^n\,\|_t + 2\|\, X^n - \bar{q}\,\|_t
\end{aligned}$$

$$\leq 3\Big(|q^n(0)-\bar{q}(0)| + \max(\lambda+\mu,\nu,|\nu-\lambda|)\int_0^t \parallel q^n-\bar{q}\parallel_u du + \frac{\lambda_2}{\sqrt{n}}t$$
$$+ \parallel \alpha^n \parallel_t + \parallel \delta^n \parallel_t\Big).$$

Using Gronwall's Lemma we obtain

$$\parallel q^n-\bar{q}\parallel_T \leq \left(|q^n(0)-\bar{q}(0)| + \frac{\lambda_2}{\sqrt{n}}T + \parallel \alpha^n \parallel_T + \parallel \delta^n \parallel_T\right) e^{DT}, \tag{B.2}$$

where D is a constant. We will now show that $\parallel \alpha^n \parallel_T, \parallel \delta^n \parallel_T \to 0$, which will complete the result. We will first show that $\parallel M^{a,n} \parallel_T \to 0$ a.s. Now as $q_i^n \geq 0$ for $i = 1,2,3$, the following holds for $i = 1,2$:

$$\parallel \bar{N}_i^a\left(\int_0^{\cdot} 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\lambda^n(Q^n(u))\,du\right)\parallel_T \leq \parallel \bar{N}_i^a\,(Cn\cdot)\parallel_T$$

for some constant $C > 0$. Using the Functional Law of Large Numbers, we obtain $\parallel \frac{1}{n}\bar{N}_i^a(n\cdot)\parallel_T \to 0$ a.s, which implies that $\parallel \frac{1}{n}\bar{N}^a\,(Cn\cdot)\parallel_T \to 0$ a.s. Hence, $\parallel \delta^n \parallel_T \to 0$ a.s. Similarly as q_i^n are bounded above,

$$\parallel \bar{N}_i^d\left(\int_0^{\cdot} 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\mu(Q^n(u))\,du\right)\parallel_T \to 0,\ a.s.,\ and$$
$$\parallel \bar{N}^r\left(\int_0^{\cdot} 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}}\nu Q_3^n(u)\,du\right)\parallel_T \to 0,\ a.s.$$

Hence, $\parallel \alpha^n \parallel_T \to 0$ a.s. ■

Proof of Lemma 4(b). We use Lemma 7 to obtain the following relation for $q^n(\cdot) \equiv \frac{Q^n(\cdot)}{n}$:

$$\begin{aligned} q_1^n(t) = q_1^n(0) &+ \int_0^t [\lambda - (\lambda+\mu)q_1^n(s) + (\nu-\lambda)q_3^n(s)]ds + \alpha_1^n(t) \\ &- \int_0^t ((1-q_1^n(s))\lambda + (\nu-\lambda)q_3^n(s))1_{\{q_1^n(s)+q_2^n(s)=\kappa^n\}}ds \end{aligned} \tag{B.3}$$

$$
\begin{aligned}
q_2^n(t) &= q_2^n(0) + \int_0^t [\lambda_1 - \mu q_2^n(s)]ds + \frac{\lambda_2}{\sqrt{n}}t \\
&\quad + \alpha_2^n(t) - \int_0^t \left(\lambda_1 + \frac{\lambda_2}{\sqrt{n}}\right) 1_{\{q_1^n(s)+q_2^n(s)=\kappa^n\}} ds, \qquad \text{(B.4)} \\
q_3^n(t) &= q_3^n(0) - \int_0^t \nu q_3^n(s)ds + \delta_3^n(t) \\
&\quad + \int_0^t ((1 - q_1^n(s))\lambda + (\nu - \lambda)q_3^n(s)) 1_{\{q_1^n(s)+q_2^n(s)=\kappa^n\}} ds. \qquad \text{(B.5)}
\end{aligned}
$$

where α^n and δ^n are defined in (3.18) and (3.19) respectively.

Define $y^n = \{\int_0^t 1_{\{q_1^n(s)+q_2^n(s)=\kappa^n\}} ds : t \geq 0\} \in D_{\mathbb{R}}[0,\infty)$. Observe that y^n is uniformly Lipschitz, and hence absolutely continuous. Further, $\{y^n(\cdot)\}$ is a family of equicontinuous and point-wise bounded functions defined on a separable space. Hence, using the Ascoli-Arzelá theorem (see Corollary 7.41 in Royden (1988)), we obtain the existence of a uniformly convergent subsequence. Let n_k index the subsequence such that $y^{n_k} \to y$. We shall now show that the sequence $\{q^{n_k}(\cdot)\}$ is Cauchy in the uniform norm. Using (B.3-B.5) we obtain

$$
\begin{aligned}
&\| q^{n_k} - q^{n_l} \|_t \\
&\leq \| q^{n_k}(0) - q^{n_l}(0) \| + K \int_0^t \| q^{n_k} - q^{n_l} \|_s ds + \lambda_2 |1/\sqrt{n_k} - 1/\sqrt{n_l}| t + \| \alpha^{n_k} - \alpha^{n_l} \|_t \\
&+ \| \delta^{n_k} - \delta^{n_l} \|_t + \lambda \| \int_0^{\cdot} (1 - q_1^{n_k}(s)) dy^{n_k}(s) - \int_0^{\cdot} (1 - q_1^{n_l}(s)) dy^{n_l}(s) \|_t \\
&+ (\nu - \lambda) \| \int_0^{\cdot} q_3^{n_k}(s) dy^{n_k}(s) - \int_0^{\cdot} q_3^{n_l}(s) dy^{n_l}(s) \|_t + \lambda_1 \| y^{n_k} - y^{n_l} \|_t \\
&+ \lambda_2 \| y^{n_k}/\sqrt{n_k} - y^{n_l}/\sqrt{n_l} \|_t \\
&\leq \| q^{n_k}(0) - q^{n_l}(0) \| + K \int_0^t \| q^{n_k} - q^{n_l} \|_s ds + \lambda_2 |1/\sqrt{n_k} - 1/\sqrt{n_l}| t + \| \alpha^{n_k} - \alpha^{n_l} \|_t \\
&+ \| \delta^{n_k} - \delta^{n_l} \|_t + \lambda \| \int_0^{\cdot} (1 - q_1^{n_k}(s)) dy^{n_k}(s) - \int_0^{\cdot} (1 - q_1^{n_l}(s)) dy^{n_k}(s) \|_t \\
&+ \lambda \| \int_0^{\cdot} (1 - q_1^{n_l}(s)) dy^{n_k}(s) - \int_0^{\cdot} (1 - q_1^{n_l}(s)) dy^{n_l}(s) \|_t \\
&+ (\nu - \lambda) \| \int_0^{\cdot} q_3^{n_k}(s) dy^{n_k}(s) - \int_0^{\cdot} q_3^{n_l}(s) dy^{n_k}(s) \|_t
\end{aligned}
$$

$$
\begin{aligned}
&+ (\nu-\lambda)\| \int_0^{\cdot} q_3^{n_l}(s)dy^{n_k}(s) - \int_0^{\cdot} q_3^{n_l}(s)dy^{n_l}(s) \|_t + \lambda_1\| y^{n_k} - y^{n_l} \|_t \\
&+ \lambda_2\| y^{n_k}/\sqrt{n_k} - y^{n_l}/\sqrt{n_l} \|_t \\
&\leq \| q^{n_k}(0) - q^{n_l}(0) \| + K\int_0^t \| q^{n_k} - q^{n_l} \|_s ds + \lambda_2|1/\sqrt{n_k} - 1/\sqrt{n_l}|t + \| \alpha^{n_k} - \alpha^{n_l} \|_t \\
&+ \| \delta^{n_k} - \delta^{n_l} \|_t + K_1\int_0^t \| q^{n_k} - q^{n_l} \|_s ds + K_2\| y^{n_k} - y^{n_l} \|_t \\
&+ \lambda_2\| y^{n_k}/\sqrt{n_k} - y^{n_l}/\sqrt{n_l} \|_t \\
&\leq (\| q^{n_k}(0) - q^{n_l}(0) \| + \lambda_2|1/\sqrt{n_k} - 1/\sqrt{n_l}|t + \| \alpha^{n_k} - \alpha^{n_l} \|_t + \| \delta^{n_k} - \delta^{n_l} \|_t \\
&+ K_2\| y^{n_k} - y^{n_l} \|_t + \lambda_2\| y^{n_k}/\sqrt{n_k} - y^{n_l}/\sqrt{n_l} \|_t)e^{K_3 t},
\end{aligned}
$$

where K, K_1, K_2 and K_3 are constants. As $q^n(0)$ is assumed to converge and the fact that $\alpha^n, \delta^n \to 0$ (as in Lemma 4(a)), we obtain that $\{q^{n_k}(\cdot)\}$ is Cauchy, and as $D_{\mathbb{R}^3}[0,\infty)$ is complete, $q^{n_k} \to \bar{q} \in D_{\mathbb{R}^3}[0,\infty)$. Further, note that $\{y^n\}$ also describes a family of measures defined on the Borel σ-field on $[0,\infty)$. Using the bounded convergence theorem, we can now argue the existence of a uniformly convergent subsequence for

$$
\tilde{y}^n \equiv \left\{ \int_0^t (\lambda(1 - q_1^n(s)) + (\nu-\lambda)q_3^n(s))dy^n(s) : t \geq 0 \right\}.
$$

Specifically,

$$
\tilde{y}^{n_k} \to \tilde{y} \equiv \left\{ \int_0^t ((1-\bar{q}_1(s))\lambda + (\nu-\lambda)\bar{q}_3(s)dy(s) : t \geq 0 \right\}. \tag{B.6}
$$

Then $\bar{q}(\cdot)$ satisfies

$$
\begin{aligned}
\bar{q}_1(t) &= \bar{q}_1(0) + \int_0^t [\lambda - (\lambda+\mu)\bar{q}_1(s) + (\nu-\lambda)\bar{q}_3(s)]ds - \tilde{y}(t) \\
\bar{q}_2(t) &= \bar{q}_2(0) + \int_0^t [\lambda_1 - \mu\bar{q}_2(s)]ds - \lambda_1 y(t) \\
\bar{q}_3(t) &= \bar{q}_3(0) - \nu\int_0^t \bar{q}_3(s)ds + \tilde{y}(t).
\end{aligned}
$$

Differentiating the above with respect to t and using (B.6), we obtain

$$\begin{aligned}
\dot{\bar{q}}_1(t) &= \lambda - (\lambda+\mu)\bar{q}_1(t) + (\nu-\lambda)\bar{q}_3(t) - ((1-\bar{q}_1(t))\lambda + (\nu-\lambda)\bar{q}_3(t))\dot{y}(t) \\
\dot{\bar{q}}_2(t) &= \lambda_1 - \mu\bar{q}_2(t) - \lambda_1\dot{y}(t) \\
\dot{\bar{q}}_3(t) &= -\nu\bar{q}_3(t) + ((1-\bar{q}_1)\lambda + (\nu-\lambda)\bar{q}_3(t))\dot{y}(t).
\end{aligned}$$

If $\bar{q}_1(t) + \bar{q}_2(t) = k_1$ for any $t \geq 0$, then $\dot{\bar{q}}_1(t) + \dot{\bar{q}}_2(t) \leq 0$, which implies that

$$\dot{y}(t) \geq 1 - \frac{\mu k_1}{(1-\bar{q}_1(t))\lambda + (\nu-\lambda)\bar{q}_3(t) + \lambda_1} > 0.$$

Hence, by the definition of y we obtain $\dot{y}(t) = 1 - \frac{\mu k_1}{(1-\bar{q}_1(t))\lambda+(\nu-\lambda)\bar{q}_3(t)+\lambda_1}$ for t such that $\bar{q}_1(t) + \bar{q}_2(t) = k_1$. Hence, for $\bar{q}(0) = \bar{\bar{q}}(0)$, we obtain $\dot{\bar{q}}(0) = 0$, which implies that $\bar{q}(t) = \bar{q}(0)$ for all $t \geq 0$. As the limit is independent of the convergent subsequence, we obtain $q^n \to \bar{\bar{q}}$. ■.

Proof of Lemma 5. Using the Lipschitz continuity of the reflection map $\tilde{\Phi}_0$ and the fact that the drift terms are linear, we obtain the existence of a unique strong solution to (3.5-3.7) by applying Theorem 2.1 in Atar, Budhiraja and Dupuis (2001). ■

Proof of Lemma 6. We shall first show that $\hat{\pi}$ must be the density of an invariant distribution of $X(\cdot)$, and then prove that $X(\cdot)$ must have a unique invariant distribution to complete the proof. Proceeding as in the proof of Theorem 1 in Heyman, Lakshman and Neidhardt (1997), we observe that for $\hat{\pi}$ to be density corresponding to an invariant distribution, the following balance relations must hold for any feasible state (j^s, x^s, j^p, x^p). If $|j^s| + j^p < k$,

$$\begin{aligned}
&\sum_{u\in U} \partial_{x_u^s}\hat{\pi}(j^s,x^s,j^p,x^p) + \sum_{i=1}^{j^p} \partial_{x_i^p}\hat{\pi}(j^s,x^s,j^p,x^p) \\
&+ \sum_{u\in j^s} \hat{\pi}(j^s\setminus\{u\}, x^s - x_u^s e^{n,u}, j^p, x^p)G'(x_u) + \sum_{u\notin j^s}\hat{\pi}(j^s\cup\{u\}, x^s - x_u^s e^{n,u}, j^p, x^p)F'(x_u) \\
&+ \sum_{i=1}^{j^p} \hat{\pi}(j^s, x^s, j^p-1, x^p - x_i^p e^{j^p,i})\lambda_p^n G'(x_i^p) + \hat{\pi}(j^s, x^s, j^p+1, x^p) \\
&- \lambda_p^n\hat{\pi}(j^s,x^s,j^p,x^p) = 0,
\end{aligned} \tag{B.7}$$

where $e^{\ell,m} \in \mathbb{R}^\ell$ with $e^{\ell,m}_k = 0$ for $k \neq m$ and $e^{\ell,m}_m = 1$, and if $|j^s| + j^p = k$,

$$\begin{aligned}
&\sum_{u\in U} \partial_{x^s_u} \hat{\pi}(j^s, x^s, j^p, x^p) + \sum_{i=1}^{j^p} \partial_{x^p_i} \hat{\pi}(j^s, x^s, j^p, x^p) \\
&+ \sum_{u\in j^s} \hat{\pi}(j^s \setminus \{u\}, x^s - x^s_u e^{n,u}, j^p, x^p) G'(x_u) + \sum_{u\notin j^s} \hat{\pi}(j^s, x^s - x^s_u e^{n,u}, j^p, x^p) F'(x_u) \\
&+ \sum_{i=1}^{j^p} \hat{\pi}(j^s, x^s, j^p - 1, x^p - x^p_i e^{j^p,i}) \lambda^n_p G'(x^p_i) = 0.
\end{aligned} \tag{B.8}$$

Using the value of $\hat{\pi}$ from (3.13) and using the local balance equations that π^s and π^p satisfy, we see that (B.7) and (B.8) hold.

Now, all that is left is to prove the uniqueness. We shall switch to a discrete time version of X to prove this result. To this effect, fix a sequence of time instants t_n such that $t_n \uparrow \infty$. Then $\{X_n = X(t_n) : n \geq 1\}$ is an irreducible, discrete time Markov chain with the same state space as X. As any invariant distribution for X is invariant for X_n, the chain X_n must have at least one invariant distribution.This allows us to conclude that X_n must have a unique invariant distribution (c.f. Proposition 10.1.1 and Theorem 10.0.1 in Meyn and Tweedie (1993)). This directly implies that X must have a unique invariant distribution. ■

Proof of Lemma 8. Using (3.26), we have for any process $Z \in D_{\mathbb{R}^3}[0,\infty)$, $\tilde{\Phi}_a(Z)(t) = Z(t) + \tilde{R}Y(t)$, where Y is a non-negative, non-decreasing process such that $\hat{n}'\tilde{R}Y(t) = -\sup_{0\leq s\leq t} (Z_1(s) + Z_2(s) - a)^+$. Using this representation of $\tilde{\Phi}$, we obtain

$$\begin{aligned}
&\| \tilde{\Phi}_a(Z^1) - \tilde{\Phi}_a(Z^2) \|_T \\
&\leq \| Z^1 - Z^2 \|_T \\
&+ \frac{\max(m,\lambda_1)}{m+\lambda_1} \| \sup_{0\leq s\leq t} \left(Z^2_1 + Z^2_2 - a\right)^+(s) - \sup_{0\leq s\leq t} \left(Z^1_1 + Z^1_2 - a\right)^+(s) \|_T \\
&\leq K \| Z^1 - Z^2 \|_T,
\end{aligned}$$

where the second inequality follows by noting that for any two real functions f, g we have $|\sup_{0\leq s\leq t} f(s) - \sup_{0\leq s\leq t} g(s)| \leq \sup_{0\leq s\leq t} |f(s) - g(s)|$ and the fact that

$\| Z_1^2 + Z_2^2 - (Z_1^1 + Z_2^1) \|_T \leq 2\| Z^1 - Z^2 \|_T$. Hence, the mapping is Lipschitz continuous with the constant $K = 1 + 2\frac{\max(m,\lambda_1)}{m+\lambda_1}$. ■

Proof of Lemma 9. We associate the independent unit rate Poisson processes $N_i^a(\cdot)$, $i = 1, 2$, $N_j^d(\cdot)$, $j = 1, 2$, and $N^r(\cdot)$ with a family of independent standard Brownian motions $B_i^a(\cdot)$, $i = 1, 2$, $B_j^d(\cdot)$, $j = 1, 2$, and $B^r(\cdot)$ such that the strong approximation result in (1.1) of Lemma 1 holds. In particular, there exist random variables X_i^z for $z \in \{a, d\}$ and $i = 1, 2$ and X^r, such that

$$X_i^z \equiv \sup_{t\geq 0} \frac{|N_i^z(t) - t - B_i^z(t)|}{\log(2 \vee t)} < \infty,$$

and

$$X^r \equiv \sup_{t\geq 0} \frac{|N^r(t) - t - B^r(t)|}{\log(2 \vee t)} < \infty,$$

Moreover, X_i^z and X^r are i.i.d and have finite means. We can write

$$\begin{aligned}
\tilde{M}_1^n(t) &= \frac{1}{\sqrt{n}}(\tilde{M}_1^{a,n}(t) - \tilde{M}_1^{d,n}(t)) \\
&= \frac{1}{\sqrt{n}} B_1^a \left(\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}} (1 - \tilde{Q}_1^n(u) - \tilde{Q}_3^n(u))\lambda du \right) \\
&+ \frac{1}{\sqrt{n}} B^r \left(\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}} \nu \tilde{Q}_3^n(u)^+ du \right) - \frac{1}{\sqrt{n}} B_1^d \left(\mu \tilde{Q}_1^n(u)^+ du \right) \\
&+ \frac{1}{\sqrt{n}} \log(2 \vee Knt) \\
&\stackrel{d}{=} B^{*a} \left(\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}} (1 - \tilde{q}_1^n(u) - \tilde{q}_3^n(u))\lambda du \right) \\
&+ B^{*r} \left(\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}} \nu \tilde{q}_3^n(u)^+ du \right) - B^{*d} \left(\mu \tilde{q}_1^n(u)^+ du \right) \\
&+ \frac{1}{\sqrt{n}} \log(2 \vee Knt),
\end{aligned}$$

where K is a constant and B^{*z} for $z \in \{a, d, r\}$ is a family of independent standard Brownian motions. The equality in distribution holds uniformly on compact sets and is due to the self-similar nature of standard Brownian motion. Note the argument of

the Brownian motion B^{*a} can be rewritten as follows.

$$\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}}(1-\tilde{q}_1^n(u)-\tilde{q}_3^n(u))\lambda du = \int_0^t (1-\tilde{q}_1^n(u)-\tilde{q}_3^n(u))\lambda du - \int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)=\kappa^n\}}(1-\tilde{q}_1^n(u)-\tilde{q}_3^n(u))\lambda du.$$

Lemma 4(a) implies that the first term converges to mt, while the second vanishes. Thus,

$$B^{*a}\left(\int_0^{\cdot} 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}}(1-\tilde{q}_1^n(u)-\tilde{q}_3^n(u))\lambda du\right) \to B^{*a}(m\cdot).$$

Similarly, $B^{*r}\left(\int_0^{\cdot} 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}}\nu\tilde{q}_3^n(u)^+du\right) \to 0$ and $B^{*r}\left(\mu\tilde{q}_1^n(u)^+du\right) \to B^{*d}(m\cdot)$. Hence, $\tilde{M}_1^n \Rightarrow W_1$. The convergence of $\tilde{M}_i^n$ for $i = 2, 3$ follow similarly, which completes the proof of (a). The proof of (b) follows in a similar manner. ■

Proof of Lemma 10. Using the Lipschitz continuity of $\tilde{\Phi}_{\kappa^n}$, we have

$$\| Q^{*n} \|_t \leq K \| \tilde{Z}^n \|_t, \quad 0 \leq t \leq T,$$

where K is the Lipschitz constant derived in Lemma 8. Further, using the fact that θ is Lipschitz continuous we obtain

$$\| \tilde{Z}^n \|_t \leq \| Q^{*n}(0) \| + K_1 \int_0^t \| Q^{*n} \|_u du + \lambda_2 t + \| \tilde{M}^n \|_t + \| \sqrt{n}\tilde{\delta}^n \|_t,$$

where K_1 is a constant. Combining the above and using Gronwall's Lemma we get

$$\| Q^{*n} \|_T \leq K(\| Q^{*n}(0) + (k_2, 0, 0)' \| + \| \tilde{M}^n \|_T + \| \sqrt{n}\tilde{\delta}^n \|_T + \lambda_2 T)e^{K_2 T},$$

where K_2 is another constant. Now as $\{Q^{*n}(0)\}$, $\{\tilde{M}^n\}$, and $\{\sqrt{n}\tilde{\delta}^n\}$ converge weakly the result follows. ■

Proof of Lemma 11. Fix any $T > 0$. We shall first show that $\{\tilde{D}_\theta^n\}$ given by (3.38) is tight in $D_{\mathbb{R}^3}[0, T]$. As θ is Lipschitz continuous, (3.38) implies

$$|\tilde{D}_\theta^n(t) - \tilde{D}_\theta^n(s)| \leq (C\| Q^{*n} \|_T + \lambda_2)(t - s), \quad 0 \leq s \leq t \leq T,$$

for some $C > 0$. Hence, using Lemma 10 we obtain the tightness of $\{\tilde{D}^n_\theta\}$. Then, as Lemma 9 holds, we obtain the tightness of $\tilde{Z}^n$. Finally, using (3.40) and the Lipschitz continuity of $\tilde{\Phi}_0$, we establish the tightness of Q^{*n} to complete the proof. ■

Proof of Lemma 12. As $\tilde{Z}^n \Rightarrow Z^*$, using the continuous mapping theorem we obtain $Q^{*n} \Rightarrow Q^*$. Thus,

$$\tilde{D}^n_\theta(\cdot) \Rightarrow \int_0^{\cdot} (-(\lambda+\mu)(Q^*_1(u)+k_2) + (\nu-\lambda)Q^*_3(u), \lambda_2 - \mu Q^*_2(u), -\nu Q^*_3(u))du.$$

Then, using Lemma 9, we obtain the desired result. ■

Proof of Lemma 13. We shall assume $\tilde{Q}^n(0) = Q^n(0)$ for all n for convenience. We have

$$\hat{Q}^n(t) - Q^{*n}(t) = \sqrt{n}(X^n(t) - \tilde{X}^n(t)) + \sqrt{n}\left(\int_0^t R^n(u)\, dY^n(u) - \tilde{R}^n(t)\tilde{Y}^n(t)\right). \quad \text{(B.9)}$$

Consider the second term. As $q^n(\cdot) \to \bar{q}$, for every $0 < \gamma \in \mathbb{R}^3$ small enough, there exists $n^* \in \mathbb{N}$ such that for each $n \geq n^*$,

$$-(m+\gamma_1, \lambda_1+\gamma_2, -m+\gamma_3)' \leq R^n \leq -(m-\gamma_1, \lambda_1-\gamma_2, -m-\gamma_3)'.$$

This implies the following.

$$\begin{aligned}
&-((m+\gamma_1)Y^n(t) - m\tilde{Y}^n(t), (\lambda_1+\gamma_2)Y^n(t) - \lambda_1\tilde{Y}^n(t), (-m+\gamma_3)Y^n(t) + m\tilde{Y}^n(t)) \\
&\leq \left(\int_0^t R^n(u)\, dY^n(u) - \tilde{R}^n(t)\tilde{Y}^n(t)\right) \\
&\leq -((m-\gamma_1)Y^n(t) - m\tilde{Y}^n(t), (\lambda_1-\gamma_2)Y^n(t) - \lambda_1\tilde{Y}^n(t), (-m-\gamma_3)Y^n(t) \\
&+ m\tilde{Y}^n(t)).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\| \int_0^t R^n(u)\, dY^n(u) - \tilde{R}^n(t)\tilde{Y}^n(t) \|_T \\
&\leq (m+\lambda_1+\gamma_1+\gamma_2+\gamma_3)\| Y^n - \tilde{Y}^n \|_T + \max_i \gamma_i \| \tilde{Y}^n \|_T
\end{aligned}$$

$$\begin{aligned}
&\leq C\| - \sup_{0\leq s\leq t}(X_1^n(s)+X_2^n(s)-\kappa^n)^+ + \sup_{0\leq s\leq t}\left(\tilde{X}_1^n(s)+\tilde{X}_2^n(s)-\kappa^n\right)^+ \|_T \\
&+ \max_i \gamma_i \| \tilde{Y}^n \|_T \\
&\leq 2C\| X^n - \tilde{X}^n \|_T + \max_i \gamma_i \| \tilde{Y}^n \|_T,
\end{aligned}$$

where $C = m+\lambda_1+\gamma_1+\gamma_2+\gamma_3$ and the last inequality follows by observing that for any two real functions $f,\ g$ we have $|\sup_{0\leq s\leq t} f(s) - \sup_{0\leq s\leq t} g(s)| \leq \sup_{s\leq t}|f(s)-g(s)|$.

Using the above equation in (B.9) we obtain

$$\| \hat{Q}^n - Q^{*n} \|_T \leq (2C+1)\| \sqrt{n}(X^n - \tilde{X}^n) \|_T + \max_i \gamma_i \| \sqrt{n}\tilde{Y}^n \|_T. \tag{B.10}$$

Now consider the term $\sqrt{n}(X^n - \tilde{X}^n)$. Using the definition of X^n and $\tilde{X}^n$ is (3.22) and (3.27) respectively, we obtain

$$\begin{aligned}
\sqrt{n}(X^n - \tilde{X}^n) = \frac{1}{\sqrt{n}} \int_0^t \theta^n(nq^n(u)) - \theta^n(n\tilde{q}^n(u))\, du + \sqrt{n}(\alpha^n(t) - \tilde{\alpha}^n(t)) \\
+ \sqrt{n}(\delta^n(t) - \tilde{\delta}^n(t)),
\end{aligned}$$

where A is a constant. Hence, using the definition of θ^n we obtain

$$\| \sqrt{n}(X^n - \tilde{X}^n) \|_T \leq \int_0^t A \| \sqrt{n}(q^n - \tilde{q}^n) \|_u \ du + \| \sqrt{n}(\alpha^n - \tilde{\alpha}^n) \|_T + \| \sqrt{n}(\delta^n - \tilde{\delta}^n \|_T.$$

Now, using this equation with (B.10), we obtain

$$\begin{aligned}
\|\sqrt{n}(\hat{Q}^n - Q^{*n}) \|_T \leq &(2C+1)A \int_0^T \|\sqrt{n}(\hat{Q}^n - Q^{*n}) \|_u \, du + (2C+1)\|\sqrt{n}(\alpha^n - \tilde{\alpha}^n) \|_T \\
&+(2C+1)\| \sqrt{n}(\delta^n - \tilde{\delta}^n) \|_T + \max_i \gamma_i \| \sqrt{n}\tilde{Y}^n \|_T.
\end{aligned}$$

Now using Gronwall's Lemma we obtain

$$\begin{aligned}
\|\sqrt{n}(\hat{Q}^n - Q^{*n}) \|_T \leq C_1\Big(&\|\sqrt{n}(\alpha^n - \tilde{\alpha}^n) \|_T + \|\sqrt{n}(\delta^n - \tilde{\delta}^n) \|_T \\
&+ \max_i \gamma_i \|\sqrt{n}\tilde{Y}^n \|_T\Big) e^{C_2 T},
\end{aligned} \tag{B.11}$$

where C_1, C_2 are constants.

The following holds.

$$\| \sqrt{n}(\alpha^n - \tilde{\alpha}^n) \|_T \leq \| \frac{1}{\sqrt{n}}(M^{a,n} - \tilde{M}^{a,n}) \|_T + \| \frac{1}{\sqrt{n}}(M^{d,n} - \tilde{M}^{d,n}) \|_T.$$

As in Lemma 9, we associate the independent unit rate Poisson processes $N_i^a(\cdot)$, $i = 1, 2$, $N_j^d(\cdot)$, $j = 1, 2$, and $N^r(\cdot)$ with a family of independent standard Brownian motions $B_i^a(\cdot)$, $i = 1, 2$, $B_j^d(\cdot)$, $j = 1, 2$, and $B^r(\cdot)$ such that the strong approximation result in (1.1) of Lemma 1 holds. Then, using (1.1) we obtain

$$\begin{aligned}
\frac{1}{\sqrt{n}}(M_1^{a,n}(t) - \tilde{M}_1^{a,n}(t)) &= \frac{1}{\sqrt{n}} B_1^a \left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}} \lambda_1^n(Q^n(u))\, du \right) \\
&\quad - \frac{1}{\sqrt{n}} B_1^a \left(\int_0^t 1_{\{\tilde{Q}_1^n(u)+\tilde{Q}_2^n(u)<k^n\}} \lambda_1^n(\tilde{Q}^n(u))\, du \right) \\
&\quad + \frac{1}{\sqrt{n}} B^r \left(\int_0^t 1_{\{Q_1^n(u)+Q_2^n(u)<k^n\}} \nu Q_3^n(u)^+ du \right) \\
&\quad - \frac{1}{\sqrt{n}} B^r \left(\int_0^t 1_{\{\tilde{Q}_1^n(u)+\tilde{Q}_2^n(u)<k^n\}} \nu \tilde{Q}_3^n(u)^+ du \right) \\
&\quad + \frac{1}{\sqrt{n}} O(\log(2 \vee Knt)),
\end{aligned}$$

where $K > 0$ is some constant. This implies that the following holds.

$$\begin{aligned}
&\frac{1}{\sqrt{n}}(M_1^{a,n}(t) - \tilde{M}_1^{a,n}(t)) \\
&\stackrel{d}{=} B^{*a} \left(\int_0^t 1_{\{q_1^n(u)+q_2^n(u)<\kappa^n\}} \frac{\lambda_1^n(nq^n(u))}{n}\, du \right) \\
&- B^{*a} \left(\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}} \frac{\lambda_1^n(n\tilde{q}^n(u))}{n}\, du \right) \\
&+ B^{*r} \left(\int_0^t 1_{\{q_1^n(u)+q_2^n(u)<\kappa^n\}} \nu q_3^n(u)^+ du \right) - B^{*r} \left(\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}} \nu \tilde{q}_3^n(u)^+ du \right) \\
&+ \frac{1}{\sqrt{n}} O(\log(2 \vee Bnt))),
\end{aligned} \tag{B.12}$$

where B^{*a} and B^{*r} are independent standard Brownian motions, and the equality in distribution holds uniformly on compact sets and is due to the self-similar nature of standard Brownian motion. Now,

$$\int_0^t 1_{\{q_1^n(u)+q_2^n(u)<\kappa^n\}}\frac{\lambda_1^n(nq^n(u))}{n}du = \int_0^t \frac{\lambda_1^n(nq^n(u))}{n}du - \int_0^t 1_{\{q_1^n(u)+q_2^n(u)=\kappa^n\}}\frac{\lambda_1^n(nq^n(u))}{n}du.$$

Thus, using Lemma 4(a), we obtain

$$\int_0^t 1_{\{q_1^n(u)+q_2^n(u)<\kappa^n\}}\frac{\lambda_1^n(nq^n(u))}{n}du \to \lambda_1(\bar{q}(0))t.$$

Similarly,

$$\int_0^t 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}}\frac{\lambda_1^n(n\tilde{q}^n(u))}{n}du \to \lambda_1(\bar{q}(0))t.$$

Then using the uniform continuity of B^{*a} on the compact set $[0,(\lambda+\nu)T]$ in (B.12), we obtain

$$\|B^{*a}\left(\int_0^{\cdot} 1_{\{q_1^n(u)+q_2^n(u)<\kappa^n\}}\frac{\lambda_1^n(nq^n(u))}{n}\,du\right) - B^{*a}\left(\int_0^{\cdot} 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}}\frac{\lambda_1^n(n\tilde{q}^n(u))}{n}\,du\right)\|_T \to 0.$$

Similarly,

$$\|\,B^{*r}\left(\int_0^{\cdot} 1_{\{q_1^n(u)+q_2^n(u)<\kappa^n\}}\nu q_3^n(u)^+du\right) - B^{*r}\left(\int_0^{\cdot} 1_{\{\tilde{q}_1^n(u)+\tilde{q}_2^n(u)<\kappa^n\}}\nu\tilde{q}_3^n(u)^+du\right)\|_T \to 0,$$

and thus

$$\|\,\frac{1}{\sqrt{n}}(M_1^{a,n}-\tilde{M}_1^{a,n})\,\|_T \Rightarrow 0.$$

Continuing in a similar manner we can show that

$$\|\,\frac{1}{\sqrt{n}}(M_i^{a,n}-\tilde{M}_i^{a,n})\,\|_T \Rightarrow 0,\ i=2,3.$$

Similarly, we obtain $\| \frac{1}{\sqrt{n}}(M_i^{d,n} - \tilde{M}_i^{d,n}) \|_T \Rightarrow 0$, i=1,2,3. Hence,

$$\| \frac{1}{\sqrt{n}}(\alpha^n - \tilde{\alpha}^n) \|_T \Rightarrow 0.$$

A similar argument shows that $\| \sqrt{n}(\delta^n - \tilde{\delta}^n) \|_T \Rightarrow 0$. Proposition 10 establishes the weak convergence of Q^{*n}, which ensures that $\| \sqrt{n}\tilde{Y}^n \|_T$ has a weak limit. Then, using these results in (B.11) and noting that γ is arbitrary, we obtain

$$\| \hat{Q}^n - Q^{*n} \|_T \Rightarrow 0.$$

■

Proof of Lemma 14. The proof follows directly using the following result along with Definition 1.

Lemma 18. *There exist constants $t_0, c_0 > 0$ that are independent of n such that for all sufficiently large n*

$$\sup_{\{z \in \mathbb{R}^2 \times \mathbb{R}_+ : e'|z| \geq c_0\sqrt{n}\}} \{\mathbb{E}\left[e'|\hat{q}^n(t_0)||\hat{q}^n(0) = z\right] - e'|z|\} \leq -\sqrt{n}.$$

In addition, there exists $\beta_0 > 0$ such that

$$\limsup_{n\to\infty} \sup_{\{z \in \mathbb{R}^2 \times \mathbb{R}_+\}} \mathbb{E}\left[\exp(n^{-\frac{1}{2}}\beta_0(e'|\hat{q}^n(t_0)| - e'|z|)^+)|\hat{q}^n(0) = z\right] < \infty, \tag{B.13}$$

and

$$\limsup_{n\to\infty} \sup_{\{z \in S\}} n^{-1}\mathbb{E}\left[(e'|\hat{q}^n(t_0)| - e'|z|)^2 \exp(n^{-\frac{1}{2}}\beta_0(e'|\hat{q}^n(t_0)| - e'|z|)^+)|\hat{q}^n(0) = z\right] < \infty, \tag{B.14}$$

where $S = \mathbb{R}^2 \times \mathbb{R}_+$.

■

Proof of Lemma 16. Corollary 1 proves that $\{\hat{\pi}^n\}$ is tight, and thus has a converging subsequence. Note that the limit point of this convergent subsequence must be invariant for the limiting diffusion process $\hat{Q}(\cdot)$. Hence, we have obtained the existence of an invariant distribution for this diffusion process. Uniqueness then follows using the argument in the proof of Lemma 6 that proves the uniqueness of the invariant distribution for the process $X(\cdot)$. ■

Proof of Lemma 18. We begin by defining a fluid trajectory $\bar{q}^n(\cdot)$ that starts at $q^n(0)$ and then obtaining bounds on the deviations of the process q^n from this fluid trajectory. Formally, $\bar{q}^n$ is defined as follows.

$$\begin{aligned}\bar{q}^n &= \Phi_{\kappa^n}(\bar{x}^n)\\ \bar{x}^n &= \bar{q}^n(0) + \int_0^t \tilde{\theta}(\bar{q}^n(u))\,du,\end{aligned}$$

where Φ is given by (3.21) and $\tilde{\theta}(x,y,z) = (-(\lambda+\mu)x - (\lambda-\nu)z, -\mu y, -\nu z)'$. $\bar{q}^n$ satisfies the following ordinary differential equation:

$$\begin{aligned}\frac{d\bar{q}_1^n}{dt} &= \lambda - (\lambda+\mu)\bar{q}_1^n - (\lambda-\nu)\bar{q}_3^n,\\ \frac{d\bar{q}_2^n}{dt} &= \lambda_1 - \mu\bar{q}_2^n,\\ \frac{d\bar{q}_3^n}{dt} &= -\nu\bar{q}_3^n,\\ \bar{q}^n(0) &= \bar{q}(0) - \left(\frac{k_2}{\sqrt{n}}, 0\right)' - \frac{z}{n},\end{aligned}$$

which is solved by

$$\begin{aligned}\bar{q}_1^n(t) &= \bar{q}_1(t) - \left(\frac{z_1 + k_2\sqrt{n}}{n} + \frac{z_3}{n(\lambda+\mu-\nu)}\right)\exp\left(-(\lambda+\mu)t\right)\\ &\qquad - \frac{z_3}{n(\lambda+\mu-\nu)}\exp\left(-\nu t\right),\\ \bar{q}_2^n(t) &= \bar{q}_2(t) - \frac{z_2}{n}\exp\left(-\mu t\right),\\ \bar{q}_3^n(t) &= \frac{z_3}{n}\exp(-\nu t),\end{aligned} \tag{B.15}$$

if $\nu \neq \lambda + \mu$. For the case $\nu = \lambda + \mu$, the solution is given by

$$\bar{q}^n(t) = \bar{q}(0) - \left(\left(\frac{z_1 + k_2\sqrt{n}}{n} + t\frac{z_3}{n} \right) \exp\left(-(\lambda+\mu)t\right), \frac{z_2}{n}\exp\left(-\mu t\right), \frac{z_3}{n}\exp(-\nu t) \right)'.$$

For the remainder of the proof, we shall assume $\nu \neq \lambda + \mu$. The case $\nu = \lambda + \mu$ can be handled in an analogous manner.The following result will be useful in completing the proof.

Lemma 19. *The following results hold.*

$$\limsup_{n\to\infty} \sup_{z\in\mathbb{R}^2\times\mathbb{R}_+} n^{\frac{1}{2}}\mathbb{E}\left[\| \bar{q}^n - q^n \|_{t_0} | \hat{q}^n(0) = z\right] < \infty, \quad \text{(B.16)}$$

$$\limsup_{n\to\infty} \sup_{z\in\mathbb{R}^2\times\mathbb{R}_+} \mathbb{E}\left[\exp(n^{\frac{1}{2}}\beta_0 \| \bar{q}^n - q^n \|_{t_0}) | \hat{q}^n(0) = z\right] < \infty, \quad \text{(B.17)}$$

$$\limsup_{n\to\infty} \sup_{z\in\mathbb{R}^2\times\mathbb{R}_+} n\mathbb{E}\left[\| \bar{q}^n - q^n \|_{t_0}^{\,2} \exp(n^{\frac{1}{2}}\beta_0 \| \bar{q}^n - q^n \|_{t_0}) | \hat{q}^n(0) = z\right] < \infty. \quad \text{(B.18)}$$

Using (B.16), we obtain the existence of a constant $C_2 > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left[|e'|\hat{q}^n(t_0)| - e'|n(\bar{q}(0) - \bar{q}^n(t_0)) - (k_2\sqrt{n}, 0)'| \mid |\hat{q}^n(0) = z\right] \\
&\leq 2n\mathbb{E}\left[\| \bar{q}^n - q^n \|_{t_0} | \hat{q}^n(0) = z\right] \\
&\leq C_2\sqrt{n}.
\end{aligned}$$

Combining this with (B.15), we obtain that

$$\begin{aligned}
&\mathbb{E}\left[e'|\hat{q}^n(t_0)| \; |\hat{q}^n(0) = z\right] \leq C_2\sqrt{n} + |z_1|\exp(-(\lambda+\mu)t_0) + |z_2|\exp(-\mu t_0) \\
&+ z_3\left(\exp(-\nu t_0) + \frac{1}{|\lambda+\mu-\nu|}\left(\exp(-(\lambda+\mu)t_0) + \exp(-\nu t_0)\right)\right) \\
&+ |k_2|\sqrt{n}(1 - \exp(-(\lambda+\mu)t_0)) \\
&= e'|z| + \tilde{C}_2\sqrt{n} - |z_1|(1 - \exp(-(\lambda+\mu)t_0)) - |z_2|(1 - \exp(-\mu t_0)) \\
&- |z_3|\left(1 - \left(\exp(-\nu t_0) + \frac{1}{|\lambda+\mu-\nu|}\left(\exp(-(\lambda+\mu)t_0) + \exp(-\nu t_0)\right)\right)\right) \\
&\leq e'|z| + \tilde{C}_2\sqrt{n} - (|z_1| + |z_2|)(1 - \exp(-\mu t_0)) \\
&- |z_3|\left(1 - \left(\exp(-\nu t_0) + \frac{1}{|\lambda+\mu-\nu|}\left(\exp(-(\lambda+\mu)t_0) + \exp(-\nu t_0)\right)\right)\right),
\end{aligned}$$

where $\tilde{C}_2 = C_2 + |k_2|(1-\exp(-(\lambda+\mu)t_0))$. Hence, choosing $c_0 = \frac{\tilde{C}_2+3}{2}$ and t_0 such that $c_0 \exp(-\mu t_0) \leq 1$ and $c_0 \left(\exp(-\nu t_0) + \frac{1}{|\lambda+\mu-\nu|}\left(\exp(-(\lambda+\mu)t_0) + \exp(-\nu t_0)\right)\right) \leq 1$ and noting that $e'|z| > c_0\sqrt{n}$, we obtain

$$\mathbb{E}\left[e'|\hat{q}^n(t_0)| \; |\hat{q}^n(0) = z\right] - e'|z| \leq -\sqrt{n}.$$

To prove (B.13) and (B.14), note that for any $\beta > 0$ following relation holds.

$$\begin{aligned} n^{-\frac{1}{2}}\beta(e'|\hat{q}^n(t_0)| - e'|z|)^+ &\leq n^{-\frac{1}{2}}\beta(e'|\hat{q}^n(t_0)| - e'|n(\bar{q}(0) - \bar{q}^n(t_0)) - (k_2\sqrt{n}, 0, 0)'|)^+ \\ &\quad + n^{-\frac{1}{2}}\beta(e'|n(\bar{q}(t_0) - \bar{q}^n(t_0)) - (k_2\sqrt{n}, 0, 0)'| - e'|z|)^+ \\ &\leq 2n^{\frac{1}{2}}\beta\| \; \bar{q}^n - q^n \; \|_{t_0} + \beta|k_2|, \end{aligned} \tag{B.19}$$

as (B.15) implies $(e'|n(\bar{q}(0) - \bar{q}^n(t_0)) - (k_2\sqrt{n}, 0, 0)'| - e'|z|)^+ \leq |k_2|\sqrt{n}$.

Use of Lemma 19 along with (B.19) completes the proof. ■

Proof of Lemma 19. We will begin by demonstrating that (B.17) implies both (B.16) and (B.18). To see this, for a fixed $\beta_1 > 0$, choose C large enough so that $\exp(\beta_1 x) > x^2 > x$ for $x > C$. Replacing $n^{\frac{1}{2}}\| \; \bar{q}^n - q^n \; \|_{t_0}$ by $\max(C, n^{\frac{1}{2}}\| \; \bar{q}^n - q^n \; \|_{t_0})$ to obtain that (B.18) and (B.16) hold provided the following holds.

$$\limsup_{n\to\infty} \sup_{z\in\mathbb{R}^2\times\mathbb{R}_+} \mathbb{E}\left[\exp\left(2\beta_0 \max(C, n^{\frac{1}{2}}\| \; \bar{q}^n - q^n \; \|_{t_0})\right) |\hat{q}^n(0) = z\right] < \infty,$$

which holds if and only if (B.17) holds. Hence, we will focus on proving (B.17).

For $i = 1, 2, 3$, we have the following

$$\begin{aligned} &\mathbb{P}\left(\sup_{0\leq t\leq t_0} \exp\left(\beta n^{\frac{1}{2}}|\bar{q}_i^n(t) - q_i^n(t)|\right) \geq u|\hat{q}^n(0) = z\right) \\ &= \mathbb{P}\left(\sup_{0\leq t\leq t_0} |\bar{q}_i^n(t) - q_i^n(t)| \geq \beta^{-1}n^{-\frac{1}{2}}\log u \,\middle|\, \hat{q}^n(0) = z\right) \\ &\leq \mathbb{P}\left(a(t_0)\frac{\lambda_2}{\sqrt{n}} + b(t_0)\| \; \alpha^n \; \|_{t_0} + c(t_0)\| \; \delta^n \; \|_{t_0} \geq \beta^{-1}n^{-\frac{1}{2}}\log u\right) \end{aligned}$$

where the inequality follows by repeating the argument used to obtain (B.2) noting

that $e'\bar{q}^n \leq e'\bar{q}$, with $a(t_0)$, $b(t_0)$, and $c(t_0)$ being the analogs of the coefficients of $\frac{\lambda_2}{\sqrt{n}}$, $\| \alpha^n \|_{t_0}$, and $\| \delta^n \|_{t_0}$ in the analog of (B.2). We will now use the strong approximation result to bound the terms $\| \alpha^n \|_{t_0}$ and $\| \delta^n \|_{t_0}$. First, define the independent and identically distributed random variables $X_j^i = \sup_{t>0} \frac{|N_j^i(t)-t-B_j^i(t)|}{\log(2\vee t)}$, for $i \in \{a, d\}$ and $j = 1, 2$, and $X^r = \sup_{t>0} \frac{|N^r(t)-t-B^r(t)|}{\log(2\vee t)}$. Lemma 1 ensures that $\mathbb{E}e^{\theta X} < \infty$ for θ small enough, where $X = X_j^i, X^r$.

Observe that the definition of α^n in (3.18), the fact $q^n \leq \max(1, e'\bar{q})$ and the strong approximation result imply for a constant $\gamma > 0$, $\| n^{\frac{1}{2}}\alpha^n \|_{t_0} \leq_{st} \sum_{i=1}^{2}(\| B_i^a(\gamma\cdot) \|_{t_0} + \| B_i^d(\gamma\cdot) \|_{t_0}) + \| B^r(\gamma\cdot) \|_{t_0} + (\sum_{i,j} X_j^i + X^r)\log(2 \vee \gamma n t_0)/\sqrt{n}$ and $\| n^{\frac{1}{2}}\delta^n \|_{t_0} \leq_{st} \| B_1^a(\gamma\cdot) \|_{t_0} + 2X_1^+ \log(2 \vee \gamma n t_0)/\sqrt{n}$, where $Y \leq_{st} Z$ if $\mathbb{P}(Y > a) \leq \mathbb{P}(Z > a)$ for $a \in \mathbb{R}$. This combined with the above relation suggests the following

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{0\leq t\leq t_0} \exp\left(\beta n^{\frac{1}{2}}|\bar{q}_i^n(t) - q_i^n(t)|\right) \geq u | \hat{q}^n(0) = z\right) \\
&\leq 5\mathbb{P}\left((b(t_0) + c(t_0))\| B_1^a(\gamma\cdot) \|_{t_0} \geq \frac{\beta^{-1}\log u - \lambda_2 a(t_0)}{10}\right) \\
&+ 5\mathbb{P}\left((b(t_0) + 2c(t_0))X_1^+ \geq \frac{\beta^{-1}\log u - \lambda_2 a(t_0)}{10}\frac{\sqrt{n}}{\log(2 \vee \gamma n t_0)}\right) \\
&= C_1 \exp\left(-c_1(\beta^{-1}\log u - \zeta)^2\right) + K_1 \exp\left(-\theta(\beta^{-1}\log u - \zeta)\frac{\sqrt{n}}{\log(2 \vee \gamma n t_0)}\right),
\end{aligned}
$$

where $\zeta = \lambda_2 a(t_0)$ and K_1, C_1, c_1 are constants. The last relation holds using tail bounds on the maximum of the Brownian motion and using $\mathbb{E}e^{\theta X_1^+} < \infty$ for some $\theta > 0$. Now,

$$
\int_2^\infty C_1 \exp\left(-c_1(\beta^{-1}\log u - \zeta)^2\right) du = \int_2^\infty C_2 u^{-c_1\beta^{-1}\left(\beta^{-1}\log u - 2\zeta\right)du)} < \infty,
$$

where C_2 is a constant and the finiteness follows as for $u > u_0 \equiv \exp\left(\frac{2\beta^2}{c_1} + 2\zeta\beta\right)$, the exponent of u in the above relation is less than -2. We also observe that

$$
\int_{e^{\zeta\beta}}^\infty K_1 \exp\left(-\theta(\beta^{-1}\log u - \zeta)\frac{\sqrt{n}}{\log(2 \vee \gamma n t_0)}\right) du < \infty
$$

. This implies the following.

$$\begin{aligned}
&\limsup_{n\to\infty}\sup_{z\in\mathbb{R}^2\times\mathbb{R}_+}\mathbb{E}\left[\sup_{0\le t\le t_0}\exp(n^{\frac{1}{2}}\beta_0(\bar{q}_i^n(t)-q_i^n(t))^+)|\hat{q}^n(0)=z\right]\\
&\le e^{\zeta\beta}+\int_{e^{\zeta\beta}}^{\infty}K_1\exp\left(-\theta(\beta^{-1}\log u-\zeta)\frac{\sqrt{n}}{\log(2\vee\gamma n t_0)}\right)du+2\\
&+\limsup_{n\to\infty}\sup_{z\in\mathbb{R}^2\times\mathbb{R}_+}\int_2^{\infty}\mathbb{P}\left(\sup_{0\le t\le t_0}(\bar{q}_i^n(t)-q_i^n(t))^+\ge\beta^{-1}n^{\frac{1}{2}}\log u|\hat{q}^n(0)=z\right)du\\
&<\infty,
\end{aligned}$$

and hence

$$\limsup_{n\to\infty}\sup_{\{z:e'z\ge 0\}}\mathbb{E}\left[\exp(n^{\frac{1}{2}}\beta_0(\|\ \bar{q}_i^n-q_i^n\ \|_{t_0})^+)|\hat{q}^n(0)=z\right]<\infty,$$

which concludes the proof of (B.17). ■

Appendix C

Proofs of Results in Chapter 4

Before embarking on the proofs of the results in this chapter, we shall state and prove some asymptotic results that will be quite useful.

C.1 Asymptotic Results

Consider a sequence of systems with the n^{th} system consisting of $\lceil an + b^n\sqrt{n} \rceil > 0$ subscribers, where $a > 0$ and $b^n \in \mathbb{R}$ such that $b^n \to b$. Let $Q^n(t)$ represent the number of servers in use at time t in the system when there are n subscribers, i.e., $Q^n(t) = \sum_{i=1}^{\lceil an+b^n\sqrt{n}\rceil} 1_{\{\text{Subscriber } i \text{ is on at time } t\}}$. We consider capacity levels of the form $k^n = \bar{k}n + \kappa^n\sqrt{n}$ for some $\bar{k} \in \mathbb{R}_+$ and $\kappa^n \in \mathbb{R}$ such that $\kappa^n \to \kappa$. Define $q^n(\cdot) = \frac{Q^n(\cdot)}{n}$, the centered and scaled process

$$\hat{Q}^n(t) = \frac{Q^n(t) - k^n}{\sqrt{n}} \leq 0$$

and $\bar{q}(\cdot) = \min\left(\bar{k}, \frac{\lambda}{\lambda+\mu}a\right)$. We then have the following asymptotic results for this system.

Lemma 20. *If $q^n(0) \to \bar{q}(0)$ a.s., then*

(a) $q^n \to \bar{q}$.

(b) If $\bar{k} = \frac{\lambda}{\lambda+\mu}a$ *and* $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$*, then* $\hat{Q}^n \Rightarrow \hat{Q}$*, where* $\hat{Q}(\cdot)$ *is a reflected affine-drift diffusion process with an upper reflecting barrier at* 0*. That is,*

$$\hat{Q}(t) = \hat{Q}(0) - (\lambda+\mu)\int_0^t \left(\hat{Q}(s) + \kappa - \frac{\lambda}{\lambda+\mu}b\right) ds + \sqrt{2am}\; B(t) - Y(t),$$

where B *is a standard Brownian motion and* Y *is the non-negative, non decreasing process such that* $\int_0^t \hat{Q}(u)dY(u) = 0$*,* $\forall t \geq 0$ *and* $Y(0) = 0$*.*

(c) The invariant distribution of $\hat{Q}^n(\cdot)$*,* $\hat{\pi}^n \to \hat{\pi}$*, where* $\hat{\pi}$ *is the unique invariant distribution of the diffusion process* $\hat{Q}$*. Further,* $\mathbb{E}_{\hat{\pi}^n}\hat{Q}^n(0) \to \mathbb{E}_{\hat{\pi}}\hat{Q}(0)$*.*

(d) The density corresponding to $\hat{\pi}$ *is given by*

$$\hat{p}(x) = \frac{\exp\left(-\frac{1}{2am}(\lambda+\mu)\left(x+\kappa-\frac{\lambda}{\lambda+\mu}b\right)^2\right)}{\int_{-\infty}^0 \exp\left(-\frac{1}{2am}(\lambda+\mu)\left(z+\kappa-\frac{\lambda}{\lambda+\mu}b\right)^2\right)}, \quad x \leq 0.$$

Proof. Perform a change in parameter with $\ell = an + b\sqrt{n}$. We now have a system with ℓ subscribers and $k^\ell = \frac{\bar{k}}{a}\ell + \frac{\left(\kappa - \bar{k}\frac{b}{a}\right)}{\sqrt{a}}\sqrt{\ell}$ servers. (Note that k^ℓ is only approximately equal to k^n, but the difference is asymptotically negligible at the diffusion scale.) We can now derive the asymptotic limits using Proposition 2 in Section 2.1. These limits, when scaled appropriately, give us the limits in (a) and (b). The convergence results in (c) follow from Theorem 1(b) in Section 3.1 and Proposition 12 in Section 3.3 respectively. Proposition 1 in Ward and Glynn (2003) gives us (d). ■

Suppose instead of the subscribers, there is an exogenous stream of customers that arrive according to a Poisson process with rate $\lambda_p = \lambda_1 n + \lambda_2^n\sqrt{n}$, where $\lambda_1 > 0$ and $\lambda_2^n \in \mathbb{R}$ such that $\lambda_2^n \to \lambda_2$. Let $Q(\cdot) \in D_{\mathbb{R}}[0,\infty)$ be the process that denotes the number of servers in use by the exogenous customers. Define $q^n(\cdot)$ and $\hat{Q}^n$ as before and $\bar{q}(\cdot) \equiv \min\left(\bar{k}, \frac{\lambda_1}{\mu}\right)$. We then have the following asymptotic results for this system.

Lemma 21. *If* $q^n(0) \to \bar{q}(0)$ *a.s., then*

(a) $q^n \to \bar{q}$.

(b) If $\bar{k} = \frac{\lambda_1}{\mu}$ *and* $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$, *then* $\hat{Q}^n \Rightarrow \hat{Q}$, *where* $\hat{Q}(\cdot)$ *is a reflected affine-drift diffusion process with an upper reflecting barrier at* 0. *That is,*

$$\hat{Q}(t) = \hat{Q}(0) + \lambda_2 t - \mu \int_0^t (\hat{Q}(s) + \kappa)ds + \sqrt{2\lambda_1}B(t) - Y(t),$$

where B *is a standard Brownian motion and* Y *is the non-negative, non decreasing process such that* $\int_0^t \hat{Q}(u)dY(u) = 0$, $\forall t \geq 0$ *and* $Y(0) = 0$.

(c) The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \to \hat{\pi}$, *where* $\hat{\pi}$ *is the unique invariant distribution of the diffusion process* $\hat{Q}$. *Further,* $\mathbb{E}_{\hat{\pi}^n}\hat{Q}^n(0) \to \mathbb{E}_{\hat{\pi}}\hat{Q}(0)$.

(d) The density corresponding to $\hat{\pi}$ *is given by*

$$\hat{p}(x) = \frac{\exp\left(-\frac{1}{2}\frac{\mu}{\lambda_1}\left(x + \kappa - \frac{\lambda_2}{\mu}\right)^2\right)}{\int_{-\infty}^0 \exp\left(-\frac{1}{2}\frac{\mu}{\lambda_1}\left(z + \kappa - \frac{\lambda_2}{\mu}\right)^2\right)dz}, \quad x \leq 0.$$

This result can be derived from Proposition 4 in Section 2.2 just as Lemma 20 is derived from Proposition 2.

Suppose now we have a system with both subscribers and the exogenous stream of customers. Let $Q(\cdot) \in D_{\mathbb{R}^2}[0, \infty)$ be the process that denotes the number of servers in use by the subscribers and the exogenous customers. Define $q^n = \frac{Q^n}{n}$, $\hat{Q}^n = \frac{Q^n - k^n}{\sqrt{n}}$, and $\bar{q}(\cdot) = \left(\frac{\lambda}{\lambda+\mu}a, \frac{\lambda_1}{\mu}\right)$. We then have the following asymptotic results for this system that can be derived from Propositions 7 and 8 in Section 3.1.1.

Lemma 22. *If* $\bar{k} = \frac{\lambda}{\lambda+\mu}a + \frac{\lambda_1}{\mu}$ *and* $q^n(0) \to \bar{q}(0)$ *a.s., then*

(a) $q^n \to \bar{q}$.

(b) If $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$*, then* $\hat{Q}^n \Rightarrow \hat{Q}$*, where* $\hat{Q}(\cdot)$ *is given by*

$$\hat{Q}_1(t) = \hat{Q}_1(0) - (\lambda+\mu)\int_0^t \left(\hat{Q}_1(s) + \kappa - \frac{\lambda}{\lambda+\mu}b\right) ds + \sqrt{2am}B_1(t) - amY(t),$$
$$\hat{Q}_2(t) = \hat{Q}_2(0) + \lambda_2 t - \mu\int_0^t \hat{Q}_2(s)ds + \sqrt{2\lambda_1}B_2(t) - \lambda_1 Y(t),$$

where B_1 *and* B_2 *are two independent standard Brownian motions and* Y *is the non-negative, non-decreasing process such that* $\int_0^t (\hat{Q}_1(u) + \hat{Q}_2(u))dY(u) = 0$*,* $\forall t \geq 0$ *and* $Y(0) = 0$*.*

(c) The invariant distribution of $\hat{Q}^n(\cdot)$*,* $\hat{\pi}^n \to \hat{\pi}$*, where* $\hat{\pi}$ *is the unique invariant distribution of the diffusion process* $\hat{Q}$*. Further,* $\mathbb{E}_{\hat{\pi}^n}\hat{Q}_i^n(0) \to \mathbb{E}_{\hat{\pi}}\hat{Q}_i(0)$ *for* $i = 1, 2$*.*

(d) The density corresponding to $\hat{\pi}$ *is given by*

$$\hat{p}(x,y) = \begin{cases} \dfrac{\exp\left(-\frac{1}{2}\left(\frac{(x+\kappa-b\lambda/(\lambda+\mu))^2(\lambda+\mu)}{am} + \frac{\left(y-\frac{\lambda_2}{\mu}\right)^2\mu}{\lambda_1}\right)\right)}{\int_{-\infty}^0 \exp\left(-\frac{(z+\kappa-b\lambda/(\lambda+\mu)-\lambda_2/\mu)^2}{2(am/(\lambda+\mu)+\lambda_1/\mu)}\right)dz}, & \text{if } x+y \leq 0, \\ 0, & \text{else.} \end{cases}$$

Armed with Lemmas 20, 21, and 22, we are now ready to prove the results of Chapter 4.

C.2 Proofs of Results in Section 4.1

Let π^n denote the invariant distribution of $Q^n(\cdot)$. We shall now state and prove the following result that will be useful in proving the results of this section.

Lemma 23. *For a sequence* (p^n, k^n) *with associated denial probability* γ^n*, we have*

$$\gamma^n = \frac{\lambda N^n(p^n,\gamma^n) - (\lambda+\mu)\mathbb{E}_{\pi^n}Q^n(0)}{\lambda\left(N^n(p^n,\gamma^n) - \mathbb{E}_{\pi^n}Q^n(0)\right)}.$$

Proof. We can write $Q^n(t)$, for any $t > 0$, as follows

$$Q^n(t) = Q^n(0) + A\left(\lambda \int_0^t \left(N^n(p^n, \gamma^n) - Q^n(s)\right) ds\right) - D\left(\mu \int_0^t Q^n(s) ds\right) - Y^n(t), \text{ a.s,} \tag{C.1}$$

where $A(\cdot)$ and $D(\cdot)$ are two independent Poisson processes with unit rate, and $Y^n(\cdot) = A\left(\lambda \int_0^{\cdot} (N^n(p^n, \gamma^n) - Q^n(s))\, ds\right) - A\left(\lambda \int_0^{\cdot} 1_{\{Q^n(s) < k^n\}} (N^n(p^n, \gamma^n) - Q^n(s))\, ds\right)$ counts the number of denied attempts. Note that

$$\begin{aligned} \lim_{t\to\infty} \frac{A\left(\lambda \int_0^t \left(N^n(p^n, \gamma^n) - Q^n(s)\right) ds\right)}{t} &= \lambda N^n(p^n, \gamma^n) - \lambda \lim_{t\to\infty} \frac{1}{t} \int_0^t Q^n(s) ds \\ &= \lambda \left(N^n(p^n, \gamma^n) - \mathbb{E}_{\pi^n} Q^n(0)\right). \end{aligned} \tag{C.2}$$

Similarly, we get

$$\lim_{t\to\infty} \frac{D\left(\mu \int_0^t Q^n(s) ds\right)}{t} = \mu \mathbb{E}_{\pi^n} Q^n(0). \tag{C.3}$$

Using (C.2) and (C.3) in (C.1), we obtain

$$\lim_{t\to\infty} \frac{Y^n(t)}{t} = \lambda N^n(p^n, \gamma^n) - (\lambda + \mu) \mathbb{E}_{\pi^n} Q^n(0).$$

Thus, using (4.1) we obtain

$$\begin{aligned} \gamma^n &= \lim_{t\to\infty} \frac{Y^n(t)}{A\left(\lambda \int_0^t \left(N^n(p^n, \gamma^n) - Q^n(s)\right) ds\right)} \\ &= \frac{\lambda N^n(p^n, \gamma^n) - (\lambda + \mu) \mathbb{E}_{\pi^n} Q^n(0)}{\lambda \left(N^n(p^n, \gamma^n) - \mathbb{E}_{\pi^n} Q^n(0)\right)}, \end{aligned}$$

which proves the claim. ■

The following result provides a justification of the demand function in (4.2).

Lemma 24. *For a system with n subscribers and $\nu = \lambda$, the mean time between successful attempts to obtain a product in steady-state for a subscriber is $\frac{1}{\lambda(1-\gamma)} + \frac{1}{\mu}$.*

Proof. For convenience we shall drop the superscript n. The mean time between successful attempts to obtain a product in steady-state for a subscriber can be written

as

$$\begin{aligned}
&\lim_{t\to\infty} \frac{t}{\text{No. of successful attempts per subscriber by } t} \\
&= \lim_{t\to\infty} \frac{\text{No. of attempts by } t}{\text{No. of successful attempts by } t} \frac{t}{\text{No. of attempts per subscriber by } t} \\
&= \frac{1}{(1-\gamma)} \frac{1}{\lambda\left(1-\frac{\mathbb{E}_\pi Q(0)}{n}\right)},
\end{aligned} \tag{C.4}$$

where we use the definition of γ given in (4.1) and the arguments in Lemma 23 to arrive at this relation. Lemma 23 also implies that

$$\gamma = \frac{\lambda n - (\lambda+\mu)\mathbb{E}_\pi Q(0)}{\lambda(n - \mathbb{E}_\pi Q(0))}. \tag{C.5}$$

Writing $\mathbb{E}_\pi Q(0)$ in terms of γ in (C.5) and substituting this in (C.4) gives us the desired result. ■

Proof of Proposition 13. Consider any sequence (p^n, k^n). Choose a subsequence n' such that $\gamma^{n'} \to \tilde{\gamma}$, $p^{n'} \to \tilde{p}$, and $k^{n'}/n' \to \tilde{k}$ as $n' \to \infty$. For this scenario,

$$\lim_{n'\to\infty} \frac{\Pi^{n'}(p^{n'}, k^{n'})}{n'} = \tilde{p}\bar{F}\left(\left(\frac{1}{\lambda(1-\tilde{\gamma})} + \frac{1}{\mu}\right)(r+\tilde{p})\right) - \tilde{k}c,$$

where using Lemma 23, $\tilde{\gamma} = \frac{\lambda\bar{F}\left(\left(\frac{1}{\lambda(1-\tilde{\gamma})}+\frac{1}{\mu}\right)(r+\tilde{p})\right)-(\lambda+\mu)\tilde{k}}{\lambda\left(\bar{F}\left(\left(\frac{1}{\lambda(1-\tilde{\gamma})}+\frac{1}{\mu}\right)(r+\tilde{p})\right)-\tilde{k}\right)}$.

Consider the following optimization problem

$$\begin{aligned}
\max_{(p,k)\in\mathbb{R}^2_+} \quad & p\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)} + \frac{1}{\mu}\right)(r+p)\right) - kc \\
\text{s.t.} \quad & \\
& \gamma = \frac{\lambda\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)} + \frac{1}{\mu}\right)(r+p)\right) - (\lambda+\mu)k}{\lambda\left(\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)} + \frac{1}{\mu}\right)(r+p)\right) - k\right)}.
\end{aligned} \tag{C.6}$$

We will show that the above problem is maximized at $(\bar{p}, \bar{k})$ and has $\gamma = 0$, which proves the claim.

Fix the number of servers at k, and denote $\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right)(r+p)\right)$ by x. Using the definition of γ, we obtain

$$x = k\mu\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right).$$

The optimization problem can thus be rewritten as

$$\max_{\gamma\geq 0} k\mu\bar{F}^{-1}\left(k\mu\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right)\right) - rk\mu\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right) - kc.$$

It is easy to see that the above is maximized at $\gamma = 0$, and hence (C.6) has the same solution as (4.3), which is $(\bar{p}, \bar{k})$. ■

We adapt the proof technique in Plambeck and Ward (2005) to prove the following result.

Proof of Lemma 17. We shall prove that for any (p^n, k^n) and the associated γ^n if $\limsup_{n\to\infty}|\sqrt{n}\theta^n| = \infty$, then $\limsup_{n\to\infty}\tilde{\Pi}^n(p^n,k^n) = \infty$, and hence (p^n,k^n) cannot be asymptotically optimal.

Given (p^n, k^n) we can write

$$\begin{aligned}
\tilde{\Pi}^n(p^n,k^n) &= \sqrt{n}\left(\bar{\Pi} - \left(p^n\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)(r+p^n)\right) - (k^n/n)c\right)\right)\\
&= \sqrt{n}\Bigg(\bar{\Pi} - \Bigg(p^n\bar{F}\left(\frac{r+p^n}{m}\right)\\
&\quad + p^n\left(\bar{F}\left(\frac{r+p^n}{m}\right) - \bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)(r+p^n)\right)\right)\Bigg)\Bigg) + (k^n/\sqrt{n})c\\
&\overset{(a)}{=} \sqrt{n}\left(\bar{\Pi} - \left(p^n\bar{F}\left(\frac{r+p^n}{m}\right) - (k^n/n)c\right)\right)\\
&\quad + \sqrt{n}p^n\left(\bar{F}\left(\frac{r+p^n}{m}\right) - \bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)(r+p^n)\right)\right)\\
&\geq \sqrt{n}\left(\theta^n c + p^n\left(\bar{F}\left(\frac{r+p^n}{m}\right) - \bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)(r+p^n)\right)\right)\right),
\end{aligned} \tag{C.7}$$

where the inequality follows from the fact that the second term in (a) is upper bounded

by the solution to the problem

$$\max_{(p,k)\in\mathbb{R}^2_+} p\bar{F}\left(\frac{r+p}{m}\right) - kc$$

s.t.

$$k = \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+p}{m}\right) + \theta^n,$$

which is $\bar{\Pi} - \theta^n c$ at $p = \bar{p}$. We now divide the proof based on two cases.

Case 1: $\limsup_{n\to\infty} \sqrt{n}\theta^n = \infty$

Since $\gamma^n \geq 0$, this implies that the second term in the lower bound for $\tilde{\Pi}^n(p^n, k^n)$ in (C.7) is non-negative. Hence,

$$\tilde{\Pi}^n(p^n, k^n) \geq \sqrt{n}\theta^n c,$$

and hence $\limsup_{n\to\infty} \tilde{\Pi}^n(p^n, k^n) = \infty$.

Case 2: $\liminf_{n\to\infty} \sqrt{n}\theta^n = -\infty$

Consider a subsequence n' such that $\lim_{n'\to\infty} \sqrt{n'}\theta^{n'} = -\infty$. Using (4.5), we split each capacity imbalance $\theta^{n'}$ into corrections in prices and number of servers, i.e., set $p^{n'} = \bar{p} + \phi^{n'}$, and $k^{n'} = (\bar{k} + \kappa^{n'})n'$, such that

$$\theta^{n'} = \frac{\phi^{n'}}{\mu} f\left(\frac{r+\bar{p}}{m}\right) + \kappa^{n'} + o(\phi^{n'}).$$

Using Lemma 23, Taylor's expansion, and the fact that $Q^{n'}(\cdot) \leq (\bar{k} + \kappa^{n'})n'$, we have the following relation for n' large enough

$$\begin{aligned}
\gamma^{n'} &\geq \frac{\lambda n' \left(\bar{F}\left(\frac{r+\bar{p}}{m}\right) - f\left(\frac{r+\bar{p}}{m}\right)\frac{\phi^{n'}}{m} - f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\gamma^{n'}/\lambda + o(\phi^{n'}) + o(\gamma^{n'})\right)}{\lambda(N^{n'}(p^{n'},\gamma^{n'}) - \mathbb{E}_{\pi^{n'}}Q^{n'}(0))} \\
&\quad - \frac{(\lambda+\mu)n'\left(\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+\bar{p}}{m}\right) + \kappa^{n'}\right)}{\lambda(N^{n'}(p^{n'},\gamma^{n'}) - \mathbb{E}_{\pi^{n'}}Q^{n'}(0))} \\
&\geq \frac{-\lambda f\left(\frac{r+\bar{p}}{m}\right)\frac{\phi^{n'}}{m} - f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\gamma^{n'} + o(\phi^{n'}) + o(\gamma^{n'}) - (\lambda+\mu)\kappa^{n'}}{\left(\lambda\bar{F}\left(\frac{r+\bar{p}}{m}\right) - \lambda\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+\bar{p}}{m}\right)\right) + O(1/n)} \\
&= \frac{-(\lambda+\mu)\theta^{n'} - f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\gamma^{n'} + o(\gamma^{n'})}{m\bar{F}\left(\frac{r+\bar{p}}{m}\right) + O(1/n)},
\end{aligned} \tag{C.8}$$

where the second inequality follows as $\lim_{n'\to\infty} p^{n'} = \bar{p}$, $\lim_{n'\to\infty} k^{n'}/n' = \bar{k}$, and

$$\lim_{n'\to\infty} \mathbb{E}_{\pi^{n'}} Q^{n'}(0)/n' = \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{r+\bar{p}}{m}\right),$$

which follows as $\frac{Q^{n'}(0)}{n} \leq \bar{k}+1$ for n' large enough.

We can simplify (C.8) to obtain

$$\frac{\theta^{n'}}{\gamma^{n'}} \geq -\frac{1}{\lambda+\mu}\left(m\bar{F}\left(\frac{r+\bar{p}}{m}\right) + f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p}) + o(\gamma^{n'})/\gamma^{n'} + O(1/n)\right). \quad \text{(C.9)}$$

Since $\lim_{n'\to\infty}\sqrt{n'}\theta^{n'} = -\infty$, the above implies $\lim_{n'\to\infty}\sqrt{n'}\gamma^{n'} = \infty$.

Using Taylor's expansion in (C.7), and then using (C.9) we get

$$\begin{aligned}
\lim_{n'\to\infty} \tilde{\Pi}^{n'}(p^{n'},k^{n'}) &\geq \lim_{n'\to\infty} \sqrt{n'}\left(\theta^n c + (\bar{p}+\phi^{n'})f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\gamma^{n'}/\lambda + o(\gamma^{n'})\right) \\
&\geq \lim_{n'\to\infty} \sqrt{n'}\gamma^{n'}\Bigg(-\frac{c}{\lambda+\mu}\left(m\bar{F}\left(\frac{r+\bar{p}}{m}\right) + f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\right) \\
&\quad + \bar{p}f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})/\lambda + O(\phi^{n'}) + o(\gamma^{n'})/\gamma^{n'} + O(1/n)\Bigg) \\
&\overset{(b)}{=} \left(\lim_{n'\to\infty}\sqrt{n'}\gamma^{n'}\frac{f\left(\frac{r+\bar{p}}{m}\right)}{\lambda}\right)\left(\bar{p}-\frac{\lambda c}{\lambda+\mu}\right)\left(r+\bar{p}-\frac{\lambda c}{\lambda+\mu}\right),
\end{aligned}$$

where (b) follows by using (4.4). Again referring to (4.4), we must have $\left(\bar{p}-\frac{\lambda c}{\lambda+\mu}\right) \geq 0$. In addition noting that $r \geq 0$, we obtain $\lim_{n'\to\infty}\tilde{\Pi}^{n'}(p^{n'},k^{n'}) = \infty$, which implies

$$\limsup_{n\to\infty} \tilde{\Pi}^n(p^n,k^n) = \infty.$$

■

Proof of Proposition 14. We shall first compute the asymptotic denial probability when the number of subscribers joining the system is independent of the denial probability. Once, we have this limit we shall incorporate the dependence to complete the proof.

Consider a sequence of systems with $N^n = \lceil an + b^n\sqrt{n} \rceil \geq 0$ subscribers in the n^{th} system, where $a > 0$ and $b^n \in \mathbb{R}$ such that $b^n \to b$, and number of servers given by $k^n = \frac{\lambda}{\lambda+\mu}an + \kappa\sqrt{n}$ for some $\kappa \in \mathbb{R}$. We can then characterize the asymptotic scaled denial probability for this system as follows.

Lemma 25. *The scaled denial probability converges as follows.*

$$\lim_{n\to\infty} \gamma^n\sqrt{n} = \sqrt{\frac{\lambda+\mu}{am}}\; h\left(-\left(\kappa - \frac{\lambda}{\lambda+\mu}b\right)\sqrt{\frac{\lambda+\mu}{am}}\right).$$

We now let the number of subscribers in the n^{th} system be $N^n(p^n, \gamma^n)$. To complete the proof we need to establish that $\gamma^n\sqrt{n}$ converges. To see this convergence, note that $N^n(p^n, \gamma^n) \leq N^n(p^n, 0)$, and hence performing a birth-death chain analysis, we obtain $\gamma^n = d(N^n(p^n, \gamma^n), k^n) \leq d(N^n(p^n, 0), k^n)$. A direct application of Lemma 25 allows us to conclude that $d(N^n(p^n, 0), k^n)\sqrt{n}$ converges as both $\phi^n\sqrt{n}$ and $\kappa^n\sqrt{n}$ converge. Hence, $\limsup_{n\to\infty} \gamma^n\sqrt{n} < \infty$. Further, as the relation (4.7) must hold for any convergent subsequence of $\gamma^n\sqrt{n}$, the result follows. ■

Proof of Lemma 25. Arguing as in Lemma 23, we can show that

$$\gamma^n = \frac{\lambda N^n - (\lambda+\mu)\mathbb{E}_{\pi^n}Q^n(0)}{\lambda\left(N^n - \mathbb{E}_{\pi^n}Q^n(0)\right)}.$$

Hence, using Lemma 20(c), we have

$$\gamma^n\sqrt{n} \to -\frac{(\lambda+\mu)^2}{a\lambda\mu}\mathbb{E}_{\hat{\pi}}\left(\hat{Q}(0) + \kappa - \frac{\lambda}{\lambda+\mu}b\right).$$

Using Lemma 20(c) we can compute

$$\mathbb{E}_{\hat{\pi}}\left(\hat{Q}(0) + \kappa - \frac{\lambda}{\lambda+\mu}b\right) = \sqrt{\frac{am}{\lambda+\mu}}\; h\left(-\left(\kappa - \frac{\lambda}{\lambda+\mu}b\right)\sqrt{\frac{\lambda+\mu}{am}}\right).$$

Hence, we have

$$\lim_{n\to\infty} \gamma^n\sqrt{n} = \sqrt{\frac{\lambda+\mu}{am}}\; h\left(-\left(\kappa - \frac{\lambda}{\lambda+\mu}b\right)\sqrt{\frac{\lambda+\mu}{am}}\right).$$

■

Proof of Proposition 15. We can express $\tilde{\Pi}^n(p^n, k^n)$ as

$$\begin{aligned}\tilde{\Pi}^n(p^n, k^n) &= \sqrt{n}\left(\bar{\Pi} - \left(p^n \bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)} + \frac{1}{\mu}\right)(r+p^n)\right) - (k^n/n)c\right)\right) \\ &= \sqrt{n}\left(\theta^n c + \bar{p} f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})(\gamma^n/\lambda) + o(1/\sqrt{n})\right),\end{aligned}$$

where the second equality follows by application of Taylor's expansion and using (4.6). Using $\theta = \lim_{n\to\infty} \theta^n \sqrt{n}$ and Proposition 14, we obtain

$$\lim_{n\to\infty} \tilde{\Pi}^n(p^n, k^n) = \theta c + \bar{p} f\left(\frac{r+\bar{p}}{m}\right)(r+\bar{p})\gamma/\lambda,$$

where γ satisfies (4.7). ■

C.3 Proofs of Results in Section 4.2

The key difference between the subscription option and the pay-per-use options is in the computation of the denial probability. For the pay-per-use option, the following analog of Lemma 23 holds.

Lemma 26. *For a sequence (p_p^n, k^n) with associated denial probability γ^n, we have*

$$\gamma^n = \frac{\Lambda^n(p_p^n, \gamma^n) - \mu \mathbb{E}_{\pi^n} Q^n(0)}{\Lambda^n(p_p^n, \gamma^n)}.$$

Proof. We can write the number of customers using the product at time t, $Q^n(t)$, for any $t > 0$, as follows

$$Q^n(t) = Q^n(0) + A\left(\Lambda^n(p_p^n, \gamma^n)t\right) - D\left(\mu \int_0^t Q^n(s)ds\right) - Y^n(t) \quad \text{a.s.},$$

where $A(\cdot)$ and $D(\cdot)$ are two independent Poisson processes with unit rate, and $Y^n(\cdot) = A\left(\Lambda^n(p_p^n, \gamma^n)\cdot\right) - A\left(\Lambda^n(p_p^n, \gamma^n)\int_0^{\cdot} 1_{\{Q^n(s)<k^n\}}ds\right)$ counts the number of denied attempts. The result now follows by continuing as in the proof of Lemma 23. ■

Using this result and Lemma 21, all other results in this section can be proved in a manner similar to those in Section 4.1 and are omitted for brevity.

C.4 Proofs of Results in Section 4.4

For this system, the following analog of Lemma 23 holds.

Lemma 27. *For a sequence (p_p^n, k^n) with associated denial probability γ^n, we have*

$$\gamma^n = \frac{\lambda N^n(p_s^n, \gamma^n) + \Lambda^n(p_p^n, \gamma^n) - (\lambda+\mu)\mathbb{E}_{\pi^n} Q_1^n(0) - \mu \mathbb{E}_{\pi^n} Q_2^n(0)}{\lambda(N^n(p_p^n, \gamma^n) - \mathbb{E}_{\pi^n} Q_1^n(0)) + \Lambda^n(p_p^n, \gamma^n)}.$$

Using this result and Lemma 22, all other results in this section can be proved in a manner similar to those in Section 4.1 and are omitted for brevity.

C.5 Proof of Proposition 21

The nominal solution is given by $\bar{p}_s = \bar{p}_p = \frac{1}{\alpha} + \frac{c}{\mu} \equiv \bar{p}$. Using this we characterize the optimal capacity imbalances θ_i^* and denial probabilities γ_i^*, $i = s, p$ for the two options. (4.8) implies that

$$\gamma_s^* = \sqrt{(\lambda+\mu)e^{\alpha\bar{p}}}\; h(z),$$

where $z = -\sqrt{(\lambda+\mu)e^{\alpha\bar{p}}}\left(\theta_s^* + \alpha e^{-\alpha\bar{p}}\bar{p}\frac{\gamma_s^*}{\lambda+\mu}\right)$. Then using the fact that for a standard normal distribution $h'(x) = h^2(x) - xh(x)$, we can rewrite the optimality condition on z, $h'(z) = \frac{\alpha c}{\mu+\alpha c}$ in conjunction with (4.8) to obtain the following relation

$$(\gamma_s^*)^2 \left(\frac{1+\alpha\bar{p}}{\lambda+\mu}\right) e^{-\alpha\bar{p}} + \theta_s^*\gamma_s^* = \frac{\alpha c}{\mu+\alpha c},$$

which implies

$$\gamma_s^* = (\lambda+\mu)e^{\alpha\bar{p}} \frac{-\theta_s^* + \sqrt{(\theta_s^*)^2 + \frac{4\alpha c(1+\alpha\bar{p})}{(\mu+\alpha c)(\lambda+\mu)} e^{-\alpha\bar{p}}}}{2(1+\alpha\bar{p})}. \tag{C.10}$$

Performing a similar analysis for the pay-per-use option, we first observe that the optimality conditions are identical, i.e., $h'(z) = \frac{\alpha c}{\mu+\alpha c}$. Hence, equating the corresponding z values for the two options, we obtain

$$-\sqrt{(\lambda+\mu)e^{\alpha\bar{p}}}\left(\theta_s^* + \alpha e^{-\alpha\bar{p}}\bar{p}\frac{\gamma_s^*}{\lambda+\mu}\right) = -\sqrt{\mu e^{\alpha\bar{p}}}\ \theta_p^*.$$

Using the value of γ_s^*, this can be simplified to obtain

$$\theta_p^* = \sqrt{\lambda}\left(\frac{1}{2}\left(\frac{2+\alpha\bar{p}}{1+\alpha\bar{p}}\right)\theta_s^* + \frac{\alpha\bar{p}}{2(1+\alpha\bar{p})}\sqrt{(\theta_s^*)^2 + \frac{4\alpha c(1+\alpha\bar{p})}{(\mu+\alpha c)(\lambda+\mu)}e^{-\alpha\bar{p}}}\right).$$

Note that this implies $\theta_p^* > \theta_s^*$ as $\lambda > 1$. Further, we can derive an analog of (C.10) to write out γ_p^* in terms of θ_p^* to obtain

$$\gamma_p^* = \frac{\mu e^{\alpha\bar{p}}}{2}\left(-\theta_p^* + \sqrt{(\theta_p^*)^2 + \frac{4\alpha c}{(\mu+\alpha c)\mu}e^{-\alpha\bar{p}}}\right),$$

which can be recast as a function of θ_s^* alone. Hence, the denial probabilities and the optimal capacity imbalance in the pay-per-use system can be written as a function of θ_s^*. This enables to write out the difference in the loss of profits between the two options as

$$P(\theta_s^*) = c\theta_p^*(\theta_s^*) + e^{-\alpha\bar{p}}\bar{p}\gamma_p^*(\theta_s^*) - \left(\theta_s^* c + \alpha e^{-\alpha\bar{p}}\frac{\bar{p}^2}{\lambda}\gamma_s^*(\theta_s^*)\right).$$

The claim is equivalent to proving $P(\theta_s^*) > 0$. Noting that computing the value of θ_s^* requires us to deal with the hazard rate function, instead we will prove that $P(\theta) > 0$ for any $\theta \in \mathbb{R}$. Via algebraic manipulations we can show that $P(0) > 0$ and $P(\theta) = 0$ has no real roots. This along with the continuity of P completes the proof.

Bibliography

Armony, M. and Maglaras, C. (2004), 'On customer contact centers with a call-back option: customer decisions, routing rules and system design', *Operations Research* **52**, 271–292.

Atar, R., Budhiraja, A. and Dupuis, P. (2001), 'On positive recurrence of constrained diffusion processes', *The Annals of Probability* **29**(2), 979–1000.

Borst, S., Mandelbaum, A. and Reiman, M. I. (2004), 'Dimensioning large call centers', *Operations Research* **52**, 17–34.

Çelik, S. and Maglaras, C. (2005), 'Dynamic pricing and lead-time quotation for a multi-class make-to-order queue', *Submitted* .

Cohen, J. W. (1957), 'The generalized engset formulae', *Philips Telecommunication Review* **18**(4), 158–170.

de Véricourt, F. and Jennings, O. B. (2005), 'Large-scale membership services', *Submitted to Operations Research* .

Erlang, A. K. (1917), 'Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges', *Electroteknikeren* **13**, 5–13.

Gallego, G. and van Ryzin, G. (1994), 'Optimal dynamic pricing of inventories with stochastic demand over finite horizons', *Management Science* **40**, 999–1020.

Gamarnik, D. and Zeevi, A. (2006), 'Validity of heavy traffic steady-state approximations in generalized Jackson networks', *Annals of Applied Probability* **16**, 56–90.

Gross, D. and Harris, C. (1998), *Fundamentals of Queueing Theory*, John Wiley and Sons, New York.

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

Harrison, J. M. and Williams, R. J. (1996), 'A multiclass closed queueing network with unconventional heavy traffic behavior', *Annals of Applied Probability* **6**(1), 1–47.

Heyman, D. P., Lakshman, T. V. and Neidhardt, L. (1997), 'A new method for analyzing feedback-based protocols with applications to engineering web traffic over the internet', *ACM SIGMETRICS* pp. 24–38.

Heyman, D. P. and Sobel, M. J. (1982), *Stochastic Models in Operations Research, Vol. 1: Stochastic Processes and Operating Characteristics*, McGraw-Hill, New York.

Hunt, P. J. and Kelly, F. P. (1989), 'On critically loaded loss networks', *Advances in Applied Probability* **21**, 831–841.

Kleinrock, L. (1975), *Queueing Systems, Volume I: Theory*, John Wiley & Sons, New York.

Kogan, Y. A., Lipster, R. S. and Smorodinskii, A. V. (1986), 'Gaussian diffusion approximations of closed Markov models of computer networks', *Problems in Information Transmission* **22**, 38–51.

Komlós, J., Major, P. and Tusnady, G. (1975), 'An approximation of partial sums of independent random variables and the sample distribution I', *Z. Wahr. und Verw. Gebiete* **32**, 111–131.

Krichagina, E. V. and Puhalskii, A. A. (1997), 'A heavy-traffic analysis of a closed queueing system with a GI/∞ service center', *Queueing Systems: Theory and Applications* **25**, 235–280.

Kumar, S. (2000), 'Two-server closed networks in heavy traffic: diffusion limits and asymptotic optimality', *Annals of Applied Probability* **10**(3), 930–961.

Kurtz, T. (1978), 'Strong approximation theorems for density dependent Markov chains', *Stochastic Processes and Their Applications* **6**, 223–240.

Maglaras, C. (2003), 'Revenue management for a multi-class make-to-order queue', *Working paper, Columbia University* .

Maglaras, C. and Zeevi, A. (2003), 'Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations', *Management Science* **49**, 1018–1038.

Mandelbaum, A., Massey, W. A. and Reiman, M. I. (1998), 'Strong approximations for Markovian service networks', *Queueing Systems* **30**, 149–201.

Mandelbaum, A. and Pats, G. (1998), 'State-dependent stochastic networks. Part I: Approximations and applications with continuous diffusion limits', *The Annals of Applied Probability* **8**(2), 569–646.

Mendelson, H. and Whang, S. (1990), 'Optimal incentive-compatible priority pricing for the M/M/1 queue', *Operations Research* **38**, 870–883.

Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

Naor, P. (1969), 'The regulation of queue sizes by levying tolls', *Econometrica* **37**, 15–24.

Paschalidis, I. and Tsitsiklis, J. (2000), 'Congestion-dependent pricing of network services', *IEEE/ACM Trans. on Networking* **8**, 171–184.

Plambeck, E. L. and Ward, A. R. (2005), Optimal control of high volume assemble-to-order systems, Technical Report 1890, Graduate School of Business, Stanford University.

Puhalskii, A. A. and Reiman, M. I. (1998), 'A critically loaded multirate link with trunk reservation', *Queueing Systems: Theory and Applications* **28**, 157–190.

Reiman, M. I. (1991), 'A critically loaded multiclass erlang loss system', *Queueing Systems* **9**, 65–82.

Royden, H. L. (1988), *Real Analysis*, Prentice Hall, Englewood Cliffs, NJ.

Savin, S. V., Cohen, M. A., Gans, N. and Katalan, Z. (2005), 'Capacity management in rental businesses with two customer bases', *Operations Research* **53**(4), 617–631.

Srikant, R. and Whitt, W. (1996), 'Simulation run lengths to estimate blocking probabilities', *ACM Transactions on Modelling and Computer Simulation* **6**(1), 7–52.

Ward, A. and Glynn, P. (2003), 'Properties of the reflected Ornstein-Uhlenbeck process', *Queueing Systems* **44**, 109–123.

Whitt, W. (2003), 'How multiserver queues scale with growing congestion-dependent demand', *Operations Research* **51**(4), 531–542.

Zeltyn, S. and Mandelbaum, A. (2005), 'Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue'. Internet Supplement, available at `http://iew3.technion.ac.il/serveng/References/regimes_supplement.pdf`.