NORTHWESTERN UNIVERSITY

Queueing Models for Service Systems with Dependencies

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering & Management Science

By

Chenguang (Allen) Wu

EVANSTON, ILLINOIS

2018

ABSTRACT

Queueing Models for Service Systems with Dependencies

Chenguang (Allen) Wu

One of the main drivers of complexity in a service system is the dependence between different random variables describing the system. For example, the queue lengths at different time points and the waiting times of different items (jobs, customers) in queue are strongly dependent. To reduce dependence-related complexities, it is customary to assume that the system primitives (such as the arrival processes, service times, patience of different customers, etc.) are independent from one another as well as from the systems dynamics and state. However, in many settings, dependencies across different primitive processes, or between a primitive process and the state of the system, should clearly hold. For example, it stands to reason, and has been shown empirically in call centers, hospitals and restaurants, that the service requirement of a customer may depend on that customer's patience or on the time that customer spends waiting in queue. A natural question to ask is then: To what extent do these types of dependencies impact the performance, control and optimal design of a service system? In this thesis, I aim to answer this question for fundamental queueing models.

I start by considering two relevant dependence structures in large service systems: In the first, customers' patience depends on their individual service requirement; in the second, the service requirement of each customer depends on that customer's delay in queue. Since either dependence structure renders exact analysis intractable, I employ a fluid approach to approximate the mean-field behavior of the stochastic queueing systems, which illustrates a first-order impact of both dependencies on the system's performance. Using a stationary analysis, I demonstrate a fundamental difference between the two dependencies in their stationary behavior for the corresponding fluid models. Despite this difference, surprisingly, a unified model can be developed to describe the two dependencies simultaneously, which is further used to characterize the relation between the two dependencies using the concept of *equivalence class*. Such relation yields important insights to the empirical identification of the dependence, which is otherwise very difficult given that any dataset of service systems is *censored* due to customer abandonment.

To manifest the role of pricing in efficiently exploiting the underlying system structure, I consider another type of dependence in service systems in which a customer's value for service depends on that customer's service requirement. In a queueing-game framework, I analyze the impact of such dependence on the service provider's revenue performance. I show that a positive dependence between the service value and the service requirement may hurt the provider's revenue if customers are charged the same price. In response to the positive dependence, I propose a novel service-based pricing scheme which prices a customer's service based on that customer's realized service time. I demonstrate that the positive dependence could be exploited under service-based pricing to generate more revenue compared to the case of no dependence.

Acknowledgements

Special thanks go to my advisors, Dr. Ohad Perry and Dr. Achal Bassamboo, for their advising on this research; their dedication to my work in the past years are highly appreciated. I will never be able to survive academia without their guidance and input throughout. I am grateful to my committee members, Dr. Barry L. Nelson and Dr. Itai Gurvich, for their strong support in my research agenda and job search. Thanks are also given to the professors at Northwestern whose insightful courses showed me the way to the field of Operations. I acknowledge the help from my peer PhD students at Northwestern as well as my friends and colleagues in the academic community. Lastly, I would like to thank my family and my girlfriend whose support helped me through the difficult time during the job search.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The last century has experienced a significant growth in the service industry worldwide: Currently, the service sector accounts for approximately 80.2% of the GDP (Factbook (2017)) and 80.3% of workforce* in the United States. In fast-developing economies, such as China and India, the service industry also plays an increasingly important role in creating jobs and generating social value. It is therefore vital to have a comprehensive understanding of service systems, with the ultimate goal to improve the efficiency in the operations of services.

A central question in the management of service systems is to determine the optimal trade-off between satisfactory service levels and a reasonable profile of operational costs. Among the various important problems that are of key interest to service centers, one of them is the staffing problem, which is concerned with the solution to the optimal number of service agents hired to operate the service center. The answer to the staffing problem is crucial because hiring labor is often very costly in developed countries and regions. For example, Bocklund and Hinton (2008) estimate that 77% of the operational costs are devoted to human sources in contact centers in the US. Hence, contact-center managers seeks to find the minimal number of working representatives that ensures a prespecified service level for their customers. In healthcare settings where the service quality becomes more crucial, the task to allocate ambulances and their crews, available doctors and nurses,

*Data from The World Bank.

as well as inpatient beds, have to be solved simultaneously, rendering the staffing problem more complicated.

1.1. Queueing Models for Service Systems with Impatient Customers

Queueing models are frequently employed to model the dynamics of service systems and are further used to provide guidance to the operational decisions. A fundamental phenomenon in service systems is that customers are often impatient while waiting for their service to commence. A customer who has to wait for the service in a queue may choose to balk, i.e., leave the system immediately after seeing a queue, or abandon the queue while waiting. For example, in call centers, incoming callers may hang up the phone immediately after they realize that they need to wait until a service representative becomes available. In emergency departments, patients may choose to leave without being seen by a doctor if their waiting time in the waiting room exceeds their patience. From a service-quality perspective, customer impatience leads to lost sales which otherwise could be transformed into real sales if the waiting time can be significantly reduced. Also, customer dissatisfaction that arises from excessive waiting and abandonment is harmful to customer loyalty and the firm's reputation. From an analytical perspective, customer impatience makes a nontrivial impact on system performance, and thus should be accounted for in the modeling and analysis of service systems.

The queueing literature has studied customer impatience extensively. A typical approach in the modeling of customers' impatience is to assume that each arriving customer is endowed with a finite patience time and the customer abandons the queue if his waiting time exceeds that patience time. A second approach is to assume that each customer

arrives with a personal valuation for service and decides to balk if his valuation is smaller than the full price of the service, which is computed as the sum of the price charged for service and the disutility incurred by the delay in queue. Therefore, the higher the valuation of a customer, the less likely he is to balk and the more patient he is. In either approach, a basic assumption in most existing queueing models is that the primitives in the system (arrival, service, patience time, valuation, etc.) are independent from one another as well as from the system's dynamics and state.

1.2. Research Question

In a wide variety of services, however, one may expect dependencies across different primitives, or between a primitive process and the state of the system. Following this idea, I will explore two directions in this thesis. In one direction, I consider customers' service and patience times to be dependent, a fact empirically observed by Reich (2012). Reich (2012) demonstrates a statistically significant correlation between the service and patience times of incoming callers in a call center that serves customers of a large commercial bank. In a many-server queueing framework, I explicitly model the dependence between customers' service and patience times and study its impact on various key performance measures in the service system and the resulting implications on the staffing problem. My results provide important insights for the optimal design of large-scale service systems with dependencies between underlying primitives, such as large contact centers with hundreds of service representatives and large hospitals with dozens of doctors and inpatient beds.

Motivated by the empirical work in Reich (2012), I also propose another dependence structure in service systems which is relevant and practically reasonable. I consider customers' service times to depend on their actual delays in queue, which is also empirically observed in other contexts. Chan et al. (2016) show that in some Intensive Care Units (ICU), patients' lengths of stay increase with the delay they experience before their admission to the ICU. It is significant to note that although the two forementioned dependencies are very different regarding how customers' service times are determined, they are in fact indistinguishable in data. This is because all datasets, wherever collected, are necessarily *censored* due to customer abandonment. In other words, both dependencies can be used to explain the empirical finding in Reich (2012). Despite this empirical difficulty, one of the key contributions of this thesis is to establish a unified model which describes the two dependencies simultaneously. The analysis of the unified model identifies a fundamental difference as well as an interesting but deep relation between the two dependencies.

In the other direction, I focus on the customers' balking behavior and consider a customer's service value to depend on his service requirement. This dependence has been studied by Anand et al. (2011) in a similar setting. Anand et al. (2011) assume that the customer population has the same service value, which increases with the service quality, which in turn increases with the average service time, which is a decision variable of the service provider. I find this assumption to be restrictive. From a customer's perspective, the average service time is only a statistical indicator of the service quality and may not be very useful to that particular customer since he is more concerned with his own service. Therefore, it is more straightforward to model a direct dependence between a

customer's service value and his individual service time, a key feature captured in my game-theoretical queueing model.

I remark that despite the existing empirical evidence that demonstrates the existence of various dependencies in service systems, the implications of these dependencies have not been well explored with only a few exceptions. Chan et al. (2016) provide an upper bound on the expected workload in a system where a customer's service time depends on the her waiting time. But the bound becomes very loose even when the system size is less than 20. Bassamboo and Randhawa (2015) discuss the optimal control when customers' service and patience may be correlated. This thesis complements the existing empirical studies and contributes to the literature by explicitly modeling and analyzing these dependencies in service systems. The insights developed in this thesis provide important guidelines for the implementation on key operational problems, such as capacity decisions, pricing strategies and routing policies.

1.3. Organization

I briefly discuss the organization of this thesis. In Chapter 2, I consider a many-server queueing system with homogeneous service agents. The key feature is that each customer arrives with a service time and patience time that are dependent. An exact analysis for such a system is intractable due to the dependence. For analytical tractability, I utilize a many-server (deterministic) fluid model to approximate the mean-field behavior of the stochastic queueing systems. Numerical experiments are given to validate the accuracy of the fluid model. I use the results derived in performance analysis to solve the service manager's capacity sizing problem.

In Chapter 3, I propose another dependence structure in which customers' service times depend on their actual delays in queue. I develop a unified transient fluid model which is able to capture the two dependencies forementioned simultaneously. The fluid analysis exposes a fundamental difference between the two dependencies, which suggests caution to operational decisions given that the two dependencies are indistinguishable in censored data.

In Chapter 4, I comment on two directions for future research. In one direction, I establish the relation between the two dependencies, which has important implications on the empirical identification of the exact form of dependence on a censored dataset. In the second, I consider a single-server service system in which customers' service values depend on their service requirement. I give structural results on the impact of such dependence on service provider's optimal pricing decision and revenue performance. I also propose a novel service-based pricing scheme, which manifests the role of pricing in efficiently exploiting dependencies in service systems.

CHAPTER 2

Service Systems with Dependent Service and Patience Times

2.1. Introduction

Customers arriving to a service system are often impatient and may choose to abandon the queue while waiting for their service to commence. A typical approach in the queueing literature to model this phenomenon is to assume that each customer is endowed with a finite patience and will abandon if his delay in queue exceeds that patience time. It is further assumed that the patience time of each customer is random, and is *independent* of all other random variables comprising the system, and in particular, of that customer's service requirement. However, in many settings one expects to have customers' patience be dependent on their individual service requirements, as is indeed observed empirically in Reich (2012) and Vries et al. (2017). In this chapter, I study the impact of such dependence on system performance and optimal staffing.

To motivate my analytical study, I start by considering the following question: To what extent does the dependence between patience and service requirement impact various system measures? To demonstrate the significant effects that such a dependency has on fundamental performance measures, I compare three systems, differing from one another only by the joint distribution of the service time and the (im)patience of the customers. The three systems I consider all have $s = 100$ agents, a Poisson arrival process with rate $\lambda = 110$, and marginal service and patience times that are exponentially distributed with

rates $\mu = 1$ and $\theta = 1/2$, respectively. The *nominal traffic intensity*, defined as the usual traffic intensity when there is no dependence, is $\rho := \lambda/s\mu = 1.1$. I remark that, under mild assumptions on the abandonment distribution, the system is always stable (reaches a steady state), regardless of the value of the nominal traffic intensity.

While there are many metrics to measure dependence, a commonly used one is Pearson's correlation coefficient, and it is adopted in the current example. In particular, recall that for random variables S and T having finite second moments with covariance $\mathrm{Cov}(S, T)$ and variances $\mathrm{Var}(S)$ and $\mathrm{Var}(T)$, Pearson's coefficient of correlation is defined via

$$r := \frac{\mathrm{Cov}(S, T)}{\sqrt{\mathrm{Var}(S)\mathrm{Var}(T)}}.$$

In my simulation study I compare the standard model, in which patience and service times are independent (with $r = 0$), to a system with positive correlation ($r = 0.4$) and a system with a negative correlation ($r = -0.4$).

Table 2.1 reports estimations for the following steady-state performance metrics: expected queue length, throughput rate (defined as the average number of service completions per unit time), expected waiting time of served customers; and the probability that an arriving customer is delayed in queue before entering service. The results are based on ten independent simulation runs, each of $3,000$ time units, with the first $1,000$ time units serving as a warm up period*. The 95% confidence intervals, calculated using the t distribution with nine degrees of freedom, are also given.

Observe that, under positive correlation, the expected queue length and expected offered wait (defined as the average time that an infinitely patient customer would wait

*In fact, the convergence of the stochastic systems to the steady state is fairly fast; see Appendix A.1.2 for more details.

Table 2.1. Simulation estimations of stationary performance measures ($\lambda = 110$, $s = 100$)

Correlation	Queue Length	Throughput	Waiting Time	Prob. of Waiting
Negative ($r = -0.4$)	10.4 ± 0.11	104.8 ± 0.04	0.09 ± 0.001	$78.7\% \pm 0.11\%$
Independent ($r = 0$)	21.0 ± 0.20	99.5 ± 0.10	0.20 ± 0.002	$93.4\% \pm 0.19\%$
Positive ($r = 0.4$)	39.9 ± 0.45	90.1 ± 0.19	0.39 ± 0.005	$99.0\% \pm 0.48\%$

before entering service) are approximately twice as large as those in the independent case, and four times as large as those in the negatively correlated case. Observe also the substantial differences in the probability that customers find all agents busy upon arrival. Since the nominal traffic intensity is greater than 1, one expects almost all arrivals to be delayed in queue. However, when $r = -0.4$, roughly 21% of the customers enter service immediately upon arrival, a statistic that is typically associated with critically loaded many-server systems (or even slightly underloaded systems), but not with overloaded ones; see, e.g., Garnett et al. (2002).

The reason for the substantial differences between the performance metrics under different correlations above can be attributed to the dramatic decrease in the throughput rate as the correlation increases. In particular, the throughput under negative correlation is approximately 5% higher than it is in the independent case, and 15% higher than the case with a positive correlation. Furthermore, under negative correlation, the throughput is larger than 100 per unit time, which is the maximum achievable throughput in systems with independent service and patience times. (In the standard independent model, the throughput is bounded by the minimum of the arrival rate and total service capacity of the pool, namely, by $\min\{\lambda, s\mu\}$. Since $\lambda > s\mu$ in this example, the throughput in the independent model equals $s\mu = 100$.) In addition, even though there is rarely any idleness in the system with a positive correlation, so that all agents are working almost all the time,

the throughput in this case is about 10% smaller than 100. These simulation results are easy to explain: Patient customers are those who get served; they require longer-than-average service times under positive correlation but shorter-than-average service times under negative correlation. As a result, the throughput is lower under positive correlation and higher under negative correlation than in the independent model. I conclude that *even moderate correlation can substantially affect the system performance*, and therefore staffing decisions, due to its impact on the total service rate, and thus the throughput.

Of course, Pearson's coefficient of correlation is only one of various metrics that measure dependence between random variables. Therefore, a second natural question to address is whether, given the arrival process, number of agents and marginal service and patience distributions, knowledge of the correlation coefficient between those latter two distributions is sufficient to determine the performance. To answer this question, I perform another simulation study in which I consider systems having the same correlation between the service time and patience, but differing in the corresponding joint distributions. Specifically, I simulate nine groups of systems, where each group consists of four systems, all four having arrival rate $\lambda = 110$, number of agents $s = 100$ and marginal service and patience times that are exponentially distributed with means 1 and 2, respectively, but different dependencies between service and patience times. The correlation coefficient is identical among the four systems within each group, but varies across the nine groups. I use a Gaussian copula and three different t-copulas, all with the same correlation coefficient, to generate the four different joint distributions for each of the nine groups; see Appendix A.1.1 for more details. The simulated steady-state expected queue length for each of the 36 systems is shown in Figure 2.1. I make two important observa-

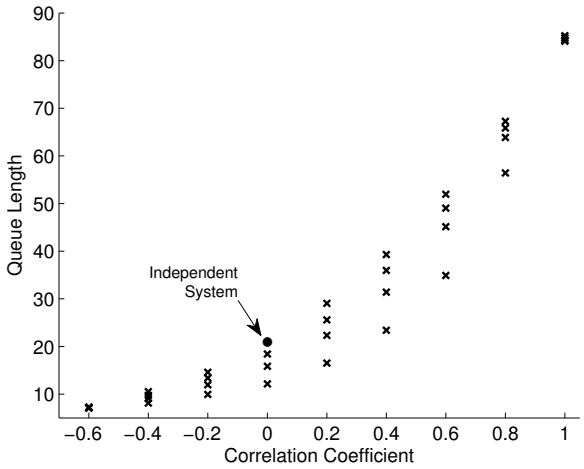


Figure 2.1. Simulated Expected Queue Length for Different Joint Distributions

tions: First, even though the correlation (and marginal distributions) of the service time and patience are the same within each of the nine groups, the queue lengths of the four systems within each group may differ significantly. The differences between the queue lengths are particularly large when the correlation is moderately positive (r is between 0.4 and 0.6). In particular, for the group with $r = 0.4$, the min-to-max ratio of queue length is almost 60%. Moreover, the case $r = 0$ demonstrates that the dependency indeed matters, even when the corresponding random variables are uncorrelated. I conclude that *the correlation coefficient is not a sufficient statistic to determine system performance.*

Second, I find that I cannot compare systems *across the groups*, namely, the correlation coefficient is not a sufficient statistic to compare systems, even if their correlations are different. Specifically, even though one intuitively expects to have the queue length increase as the correlation increases, this is not true in general. For example, the expected steady-state queue length in the independent model could be larger than the expected

queue length in a system with $r = 0.2$, and could be roughly equal to the expected queue length in a system with $r = 0.4$.

To summarize, the two simulation examples above suggest that (i) dependency between patience time and service requirement can have substantial impacts on system performance, and that (ii) to isolate its effects, one must consider more refined measures of dependencies than simple correlation.

The Setting. To gain insights, I consider a many-server queueing system with a single pool of statistically homogeneous agents. I assume that each arriving customer is endowed with a bivariate random variable whose marginals represent that customer's service requirement and patience time, and that those bivariate random variables are independently and identically distributed (IID) across the customers. In the presence of the dependence, it is essential to distinguish between the *nominal service rate*, denoted by μ, which I define to be the reciprocal of the (unconditional) expected service time of all arrivals, and the *effective service rate*, denoted by μ_{eff}, which is the reciprocal of the actual mean service time in steady state, averaged over the customers that end up receiving service. The key to analyzing the system is to characterize this effective service rate, or alternatively, the throughput rate, defined to be the long-run average number of service completions per unit time. Equivalently, the throughput rate can be defined as the average number of service completions per unit time when the system is stationary. (I will use the terms "stationary" and "steady state" interchangeably.)

Of course, the dependence between the service requirement and patience of each customer only matters if sufficiently many customers need to wait in queue for a sufficiently long time. (Otherwise, the effective service rate μ_{eff} will be approximately equal to the

nominal service rate μ.) Therefore, *I focus on overloaded systems*, where an overload is defined to hold when the arrival rate λ satisfies $\lambda > s\mu$ with s being the number of agents. It is significant that $\lambda > s\mu$ implies that the system is overloaded, even if μ_{eff} can be substantially larger than μ. Indeed, if this was not the case, i.e., if the system was to stabilize at a non-overloaded equilibrium, then waiting times would necessarily be negligible *in a sufficiently large system*, in which case it would hold that $\mu_{\text{eff}} \approx \mu$. In turn, this implies that $\lambda > s\mu \approx s\mu_{\text{eff}}$, so that the system is overloaded, and waiting times are nonnegligible. This heuristic contradictory argument is formalized in Proposition 2.2 below.

I note that the dependence may also have substantial impacts on critically loaded systems in some cases, because a nonnegligible proportion of customers are delayed in queue. Since my analysis is motivated by asymptotic considerations, and in particular, by a weak law of large numbers (which I do not formally prove here), the stochastic fluctuations of the queue in a critically loaded system are negligible for sufficiently large systems; see also Remark 2.1 below. For applications in which those stochastic fluctuations are nevertheless significant (because the systems is not sufficiently large), I propose a heuristic refinement in §2.6.3 below.

2.1.1. Main Goals and Contribution

Goals. In this chapter I aim to quantify the impacts of a dependency between the service requirement and the patience of customers on key performance measures and on optimal staffing decisions, when capacity and abandonment costs are incurred. (Henceforth,

dependence or correlation refer to that between the service time and patience time distributions.) As the simulation study depicted in Figure 2.1 shows, quantifying the impact of the dependence on the queueing system requires more refined measures of dependency than simple correlation.

To this end, I must first develop an effective approximation for the analytically intractable queueing system. Indeed, even if the arrival process is Poisson and the marginal distributions of the service and patience times are both exponential (distributional assumptions that I do not make), the number-in-system-process is not Markovian, since the service-time distribution of a customer in service is related to his delay in queue. Hence, the service-time distribution of each customer in service at any given time is in general different than that of any other customer in service, rendering exact analysis intractable. Thus, building on the fluid model for non-Markovian many-server systems proposed in Whitt (2006a) and Bassamboo and Randhawa (2015), I employ a stationary fluid model to approximate the steady-state distribution of the stochastic queueing system. It is important to note that the fluid model is characterized via the full joint distribution of the service time and patience (see §2.4.1 below), so that the dependence structure and its impact on the fluid model can be studied.

Contribution. With respect to the goals above, my contribution here is fourfold:

(I) I explicitly characterize the effective service rate of the fluid model in stationarity, from which the value of the throughput rate follows immediately. Given the throughput rate, all other key performance measures in the fluid model, e.g., the stationary queue length and the waiting time of served customers, can be easily

computed. I demonstrate via simulation experiments that my fluid model is an effective and accurate approximation. See §2.4 for my fluid model.

(II) I provide a novel framework to measure the impact of the dependence on the fluid model, and in turn, on the stochastic system it approximates. First, for a given system, I study how the structure of the conditional expected service time, conditioned on the waiting time in queue, impacts the throughput rate (which determines other performance measures). Second, I compare systems differing from each other only by the dependence structure. To this end, I rank the "strength" of the dependence by utilizing the Positive Quadrant Dependence (PQD) stochastic order; see §2.3 for background of PQD order and §2.5 for performance analysis.

(III) I apply the fluid model and the framework described in (I) and (II) to study the economic implications of the service-patience dependency by analyzing an optimal-staffing problem when costs for staffing and abandonment are incurred. In particular, I compute the fluid-optimal staffing, as well as provide structural results regarding how the dependence affects that optimal staffing. In addition, based on my fluid analysis, I provide a heuristic safety-staffing rule for settings in which the fluid-optimal solution is to process all the input (implying that it is optimal to have the stochastic system be critically loaded), in which case second-order stochastic fluctuations have a dominant impact on the optimal solution. See §2.6 for the capacity sizing problem and the heuristic refinement.

(IV) Estimating the exact joint distribution can be hard in practice. Thus, a parametric approach is warranted. I therefore demonstrate that my main structural

results hold for important classes of bivariate random variables generated by copulas, facilitating simulation experiments that can be used to estimate possible scenarios for different joint distributions. In particular, I focus on the class of Gaussian copulas (see §2.3.2 for details), whose relative tractability makes them attractive, and thus prevalent, in modeling.

2.2. Related Literature

Related Queueing Models. As was mentioned above, the fluid model I employ builds on the fluid model proposed in Whitt (2006a) to approximate the non-Markovian $G/GI/s+GI$, which has a general stationary arrival process (the G), IID service times with general distribution (the first GI), s statistically homogeneous agents, and IID times for waiting customers to abandon the queue while waiting for service (the $+GI$). Whitt's fluid model is shown to hold as a bona-fide fluid limit in the many-server heavy-traffic limiting regime in Kang and Ramanan (2010) and Zhang (2013). The fluid model in Whitt (2006a) is employed to optimize staffing decisions when the arrival rate and number of agents in a call center are uncertain in Whitt (2006b), and to study the impact of delay announcements in Armony et al. (2009). The stationary point of a fluid model in which the service time and patience can be dependent is characterized in Bassamboo and Randhawa (2015), which considers scheduling policies for customers based on their waiting times. Liu and Whitt (2011a,b) adapt the approach in Whitt (2006a) to study systems in which the arrivals and staffing may vary with time. The two papers Bassamboo and Randhawa (2010) and Bassamboo et al. (2010) use a fluid approach to study capacity-sizing problems, and show that the fluid model yields accurate approximations for large overloaded systems.

Although most of the literature assumes that the random variables comprising the primitive processes of queueing models (arrivals, service times, and patience times when abandonment is considered) are independent, there are a few exceptions. Both Whitt (1990) and Boxma and Vlasiou (2007) consider a $G/G/1$ system in which the service rate depends linearly on the delay process. More recently, heavy-traffic limits for infinite-server models in which successive service times are dependent were developed in Pang and Whitt (2012, 2013). Li and Whitt (2014) build on the latter references to approximate blocking probabilities in loss models when successive service times and successive interarrival times are allowed to be dependent. Whitt and You (2016) employ a robust optimization approach to consider the impact of serial dependence between interarrival and service times in a single-server queue.

Motivated by empirical evidence that long waiting times for admissions often lead to increased hospitalization times in intensive-care units, Chan et al. (2016) analyze an $M/M(f)/n$ queueing model (with no abandonment) in which service times are exponentially distributed with a mean which increases with congestion according to a given "inflation" function f (the notation $M(f)$ for the service time). Upper bounds for the waiting times in queue are developed, and are shown to be fairly accurate for small systems (with a small number of servers) or systems with low utilization. I, on the other hand, consider large and overloaded systems.

Bivariate Stochastic Order and Copulas. Recall that one of the goals in this chapter is to compare and rank systems having identical marginal distributions of service and patience times but different dependence structures. To this end, I employ the PQD order mentioned above and copulas. I refer to Scarsini and Shaked (1996) and Shaked

and Shanthikumar (2007) for a surveys of positive dependence orders in general, and PQD in particular, and to Joe (1997) and Nelsen (2013) for overviews of the theory and applications of copulas. Stochastic orders for multivariate random variables generated by a common copula can be found in Müller and Scarsini (2001).

The multivariate Gaussian copula is applied in Clemen and Reilly (1999) for decision and risk analysis. Both Corbett and Rajaram (2006) and Mak and Shen (2014) study the benefits of inventory pooling by adopting the supermodular order to compare the dependence of demand at multiple locations. In the queueing literature, Müller (2000) uses the PQD order to rank the dependence between the service time of a customer and the subsequent interarrival time. It is shown that stronger dependence between interarrival and service times leads to decreasing waiting times in the increasing convex ordering sense.

2.3. Measures of Dependence

In this section, I describe the measures of dependence that I will use in this chapter. I provide more details in Appendix A.1.1. Let S and T be two random variables with finite second moment. Let $f := f(S,T)$ denote the joint density of S and T having marginal densities f_S and f_T, respectively.

I consider the set of all bivariate distributions with the same marginal densities f_S and f_T, which I denote by $\mathcal{F}(f_S, f_T)$. (Note that, if S and T are independent, then their joint distribution function is in $\mathcal{F}(f_S, f_T)$, so that this set is not empty; it can be shown that there are many other joint distributions in this set; see §2.3.2.) I first employ a stochastic order, introduced in §2.3.1 below, to rank the strength of the dependence of the elements in $\mathcal{F}(f_S, f_T)$. I then discuss how to use copulas to represent joint distributions in §2.3.2.

2.3.1. Measuring Dependence via Bivariate Dependence Orders

A natural dependence concept is achieved by comparing the joint distribution of two dependent random variables X_1 and X_2 to the distribution of two independent random variables with the same marginals. In particular, X_1 and X_2 are said to be Positive Quadrant Dependent (PQD) if

$$\mathbb{P}(X_1 > x_1, X_2 > x_2) \geq \mathbb{P}(X_1 > x_1)\mathbb{P}(X_2 > x_2) \quad \text{for all } x_1, x_2.$$

Similarly, X_1 and X_2 are said to be Negative Quadrant Dependent (NQD) if $\mathbb{P}(X_1 > x_1, X_2 > x_2) \leq \mathbb{P}(X_1 > x_1)\mathbb{P}(X_2 > x_2)$ for all x_1, x_2.

Loosely speaking, PQD means that large values of X_1 tend to go together with large values of X_2, namely, both random variables are more likely to be large together than if they were independent.

The notion of PQD leads to the following bivariate stochastic dependence order; see, e.g., Shaked and Shanthikumar (2007, Chapter 9).

Definition 2.1 (PQD order)**.** *For random vectors (X_1, X_2) with joint cdf G and (Y_1, Y_2) with joint cdf H, suppose that G and H have the same marginal cdf's F_1 and F_2. I say that (X_1, X_2) is smaller than (Y_1, Y_2) in the PQD order, denoted by $(X_1, X_2) \leq_{PQD} (Y_1, Y_2)$, if*

$$G(x_1, x_2) \leq H(x_1, x_2), \text{ or equivalently, } \bar{G}(x_1, x_2) \leq \bar{H}(x_1, x_2) \quad \text{for all } x_1, x_2$$

where $\bar{G}(x_1, x_2) := \mathbb{P}(X_1 > x_1, X_2 > x_2)$ and $\bar{H}(x_1, x_2) := \mathbb{P}(Y_1 > x_1, Y_2 > x_2)$.

One can analogously define NQD order by switching the inequalities between G and H, namely, $(X_1, X_2) \leq_{NQD} (Y_1, Y_2)$ if $\bar{G}(x_1, x_2) \geq \bar{H}(x_1, x_2)$ for all x_1, x_2.

It is worth noting that, even though PQD (NQD) order is a partial order on $\mathcal{F}(f_S, f_T)$ (not all the bivariate distributions in $\mathcal{F}(f_S, f_T)$ can be ranked by PQD order), it is widely considered to be the *most fundamental stochastic dependence order*; see Colangelo et al. (2006). Indeed, Joe (1997) postulates that PQD order possesses all the desirable properties that a multivariate positive dependence order should satisfy, and that any other stochastic positive dependence order should imply PQD order.

One can relate PQD order to Pearson's correlation in the following lemma, whose proof can be found in Shaked and Shanthikumar (2007, p. 389).

Lemma 2.1. *If $(S_1, T_1) \leq_{PQD} (S_2, T_2)$, then $r_1 \leq r_2$, where r_i is the Pearson's correlation coefficient of (S_i, T_i), $i = 1, 2$.*

2.3.2. Measuring Dependence via Copulas

A d-dimensional copula C, associated with a random vector $(X_1, \ldots, X_d)$ having joint cdf F and marginal cdf's $F_1, \ldots, F_d$, is a joint cdf on the unit cube $[0, 1]^d$ with uniformly distributed marginals, such that

$$C(F_1(x_1), \ldots, F_d(x_d)) = F(x_1, \ldots, x_d) \quad \text{for all } x_1, \ldots, x_d, \quad d \geq 2. \tag{2.1}$$

By Sklar's theorem (e.g., Clemen and Reilly (1999, §2)), a copula exists uniquely for any given joint cdf F if the marginals are continuous (as I assume). Moreover, for any marginal distribution F_i, $i = 1, \ldots, d$, and a copula C, there exists a joint distribution function F, such that (2.1) holds. Thus, the use of copulas provides great modeling flexibility for

practical purposes as it places no restriction on the marginal distributions. (In principle, I could choose any marginal distributions for S and T, and construct a joint distribution having those marginals.) Furthermore, copulas offer increased tractability, since they allow us to "decouple" a joint distribution of a multivariate random variable into its univariate marginal distributions and the copula, which fully captures the dependence structure between the marginals. In my setting, copulas are useful not only in generating joint distributions, but also because many classes of copulas can be associated with PQD order. In particular, let $\mathcal{P} := \mathcal{P}(f_S, f_T)$ denote a subset of $\mathcal{F}(f_S, f_T)$ that can be ranked by PQD order; the existence of a nonempty set $\mathcal{P}$ can be deduced from (9.A.6) in Shaked and Shanthikumar (2007). It is significant that a set $\mathcal{P}$ can be chosen to be the set of bivariate distributions generated by one of many commonly used copulas, e.g., the Guassian copula, t-copula and various Archimedean copulas (such as Frank, Joe, AMH and Gumbel copulas.) Due to its tractability, the Guassian copula plays a fundamental role in modeling dependent distributions. I will therefore focus on this class of copulas, and demonstrate how my results translate to the corresponding joint distributions.

I denote the set of joint distributions generated by the Gaussian copula with fixed marginals f_S and f_T by $\mathcal{G} := \mathcal{G}(f_S, f_T)$. For a given $r_G \in [-1, 1]$, a Gaussian copula can be written as

$$C(x_1, x_2) = \Phi_{r_G}(\Phi^{-1}(x_1), \Phi^{-1}(x_2)), \quad x_1, x_2 \in [0, 1],$$

where Φ is cdf of the standard normal random variable and Φ^{-1} is its inverse, and Φ_{r_G} is the joint cdf of a bivariate normal with mean vector zero and correlation coefficient r_G. (Note that r_G is not the correlation coefficient of the resulting joint distribution, which I denote by r.) It follows that a Gaussian copula can be used to construct a

bivariate distribution for any predetermined marginals and any *attainable* correlation coefficient r[†]. Moreover, Lemma A.2 in Appendix A.1.2 proves that the elements in $\mathcal{G}(f_S, f_T)$ can be ranked by r, namely, by a single parameter. This latter property makes the Gaussian copula an attractive object of study, because it implies that the complicated high-dimensional dependence structure of the random variables generated by the copula can be quantified by a scalar.

2.4. Model

I consider a multi-server queueing system with s statistically identical agents. Customers arrive to the system according to a general stationary arrival process; upon arrival, a customer enters service immediately if an agent is available, and joins the queue if all agents are busy. I assume that each customer has a finite patience for waiting to be served, and will abandon the queue if his waiting time exceeds that patience. A key feature of my model is that the patience time of a customer depends on that customer's service requirement, although the bivariate random variables of service and patience times are independent across customers.

More specifically, letting S_i and T_i denote the service requirement and patience time of customer i, respectively, I assume that $\{(S_i, T_i) : i \geq 1\}$ are IID bivariate random variables, all having the same continuous joint density f and marginal densities f_S and f_T for service time and patience time, respectively. The support of both marginal densities is assumed to be the entire positive half of the real line. I use S and T to denote generic

[†]In general, for given marginals there can be values in $[-1, 1]$ that r cannot achieve. For example, if both the marginals are exponential distributions, then r cannot attain values smaller than -0.64. Moreover, marginal distributions together with a correlation coefficient do not uniquely determine a joint distribution. Extreme examples in Sharakhmetov and Ibragimov (2002) and Embrechts et al. (2002) give a continuum of bivariate distributions with the same marginals and correlation coefficient.

random variables having joint density f, and marginals f_S and f_T. I further assume that $\mathbb{E}[S^2] < \infty$ and $\mathbb{E}[T^2] < \infty$, so that both random variables have finite expectations, and the correlation coefficient between S and T, denoted by r, is well defined. I refer to $\mu := 1/\mathbb{E}[S] > 0$ as the *nominal service rate*, because μ would be the service rate if there was no waiting, namely, if the system had sufficient capacity to operate as an infinite-server queue.

Let λ denote the arrival rate, and let $\rho := \lambda/s\mu$ denote the *nominal traffic intensity*. I consider overloaded systems in which the arrival rate is larger than the total service capacity and thus a non-negligible fraction of customers abandon the system. It will be shown in Proposition 2.2 below that if $\lambda > s\mu$, or equivalently $\rho > 1$, then the system is overloaded for any joint distribution f.

2.4.1. The Fluid Model

As was mentioned above, if S and T are dependent, the number-in-system process is necessarily non-Markovian, rendering stochastic analysis prohibitively hard. I therefore employ a deterministic fluid model, as in Whitt (2006a) and Bassamboo and Randhawa (2015), to approximate the stationary queueing system, and demonstrate the effectiveness of that fluid model via simulations. To construct the fluid model, I replace the stochastic arrival, service and abandonment processes by corresponding deterministic flows. In particular, I start by taking the number of agents s to be a positive real number (not necessarily an integer), and imagine that fluid flows into the system at rate λ. Since each of the s agents processes work at rate μ, fluid flows out of service at rate $s\mu$, so that, by the assumption $\rho > 1$, the rate at which fluid arrives is greater than the processing rate of

all agents combined, implying that a nonnegligible proportion of fluid leaves the system via abandonment.

In my setting, the workload in the system depends on the waiting time; to characterize it, I define the *work evolution function*

$$\phi(w) := \int_w^\infty \int_0^\infty x f(x,y) dx dy, \tag{2.2}$$

which represents the work of a unit of fluid that remains in the system after waiting for w time units in the queue. To see this, observe that $\phi(w) = F_T^c(w)\mathbb{E}[S|T > w]$, where $F_T^c := 1 - F_T$ is the proportion of fluid that remains in queue after waiting w time units, and $\mathbb{E}[S|T > w]$ is the average work of that remaining fluid. In steady state, the work flow *into service* must be equal to the work flow *out of service*, giving rise to the steady state fluid equation

$$\lambda\phi(w) = s. \tag{2.3}$$

Observe that $\phi(w)$ is strictly decreasing in w due to my assumption that f_S and f_T are strictly positive over $[0, \infty)$, implying the following result.

Proposition 2.1. *If $\rho > 1$, then there exists a unique $\bar{w} > 0$ that solves equation (2.3).*

I refer to the unique solution $\bar{w}$ to (2.3) as the *offered wait*. It represents the time in queue that a virtual customer endowed with infinite patience would wait before entering service when the fluid model is stationary. In other words, in the fluid model, customers with patience greater than or equal to $\bar{w}$ enter service after waiting exactly $\bar{w}$ time units

in queue, whereas the remaining customers, whose patience is smaller than $\bar{w}$, abandon the queue.

Given the steady state offered wait, I can characterize other key performance measures for the fluid model. Let $a(w)$ denote the conditional expected service time, conditioned on the patience being larger than w, i.e.,

$$a(w) := \mathbb{E}[S|T > w]. \tag{2.4}$$

Then $a_{\text{eff}} := a(\bar{w})$ is the *average effective service time* in steady state, so that $\mu_{\text{eff}} := 1/a_{\text{eff}}$ is the *effective service rate* in steady state. Given the effective service rate μ_{eff} I can characterize the *effective traffic intensity* to the system

$$\rho_{\text{eff}} := \frac{\lambda}{s\mu_{\text{eff}}}. \tag{2.5}$$

Next, dividing both sides of the equality in (2.3) by $s\mu$ gives

$$\rho\phi(\bar{w}) = 1/\mu. \tag{2.6}$$

Noting that $\phi(w) = F_T^c(w)a(w) = F_T^c(w)/\mu_{\text{eff}}$, where F_T^c is the complement of the cumulative distribution function (cdf) F_T of patience time, I see that (2.6) can be represented via

$$\rho F_T^c(\bar{w}) = \frac{\mu_{\text{eff}}}{\mu}, \quad \text{or equivalently,} \quad \frac{\lambda F_T^c(\bar{w})}{s} = \mu_{\text{eff}}. \tag{2.7}$$

The first equality in (2.7) is a generalization of Equation (3.9) in Whitt (2006a), which states $\rho F_T^c(w) = 1$ in the independent model. The second equality in (2.7) can be interpreted as follows: Since $F_T^c(\bar{w})$ is the proportion of fluid that remains in the queue after $\bar{w}$ time units, and thus gets served, $\lambda F_T^c(\bar{w})/s$ represents the rate per agent at which fluid flows into service, and this rate must equal the effective service rate of an agent μ_{eff}.

I note that when S and T are positively dependent, $a(w) = \mathbb{E}(S|T > w)$ might increase to infinity as $w \to \infty$. However, the assumption that $\mathbb{E}[S] < \infty$ ensures that $F_T^c(w)a(w)$ is strictly decreasing and converges to 0 as $w \to \infty$.

Next, I compute the throughput and stationary fluid queue which I denote by R and Q, respectively. Clearly, I have $R = s\mu_{\text{eff}}$, so that

$$R = s\mu_{\text{eff}} = s\mu\rho F_T^c(\bar{w}) = \lambda F_T^c(\bar{w}), \tag{2.8}$$

where the second equality follows from (2.7). The expression for the steady-state fluid queue length Q is derived as follows: The amount of fluid that enters the queue over an interval $[t, t + dx)$ is λdx, and the proportion of that fluid remaining in the queue t time units later after arrival is $F_T^c(t)$. Since all arriving fluid that is served waits exactly $\bar{w}$, it holds that

$$Q = \lambda \int_0^{\bar{w}} F_T^c(x)dx. \tag{2.9}$$

Observe that the fluid model is completely determined by the three elements in the *primitive data set* $\mathcal{D} := (\lambda, s, f)$. (Note that the marginal distributions of S and T and the nominal service rate μ are easily recovered from f.) Indeed, given the model data in $\mathcal{D}$, I

can compute the offered wait $\bar{w}$ via (2.3), from which a_{eff} and μ_{eff} can be easily recovered via (2.7). Given these latter two variables, I can compute the stationary throughput R in (2.8) and fluid queue Q in (2.9).

Note that in an overloaded system, i.e., with $\bar{w} > 0$ (Proposition 2.1), my fluid model captures "predictable" queueing effects, which are due to insufficient service capacity. This is different than non-overloaded systems, in which queueing is due to stochasticity associated with the arrival and service process. Specifically, the fluid model does not capture queueing effects that are due to random fluctuations. The following remark elaborates on this point from an asymptotic perspective.

Remark 2.1. *Even though I do not prove limit theorems here, it is helpful to think of the stationary fluid model as a weak law of large numbers for a sequence of stationary stochastic systems. More formally, consider a sequence of stochastic systems as described above indexed by the number of agents s. Assume that the arrival rate to system s is $\lambda_s := s\lambda + o(s)$ (where $o(s)$ denotes a function that increases slower than s, i.e., $o(s)/s \to 0$ as $s \to \infty$), but that the joint distribution f is fixed along the sequence. Letting $Q_s(\infty)$ denote a random variable which is distributed as the stationary queue in the s system, I conjecture that $Q_s(\infty)/s$ converges in distribution to Q in* (2.9)*, and that a similar result holds for the stationary distribution of the service process. In particular, I expect my fluid model to become more accurate as the size of the system increases, although my simulation experiments (depicted in Figures 2.2 and 2.3 below) demonstrate that the system need not be too large. It is readily seen from the spatial scaling by s of the prelimit that the fluid model does not capture fluctuations of order $o(s)$. Hence, the fluid queue*

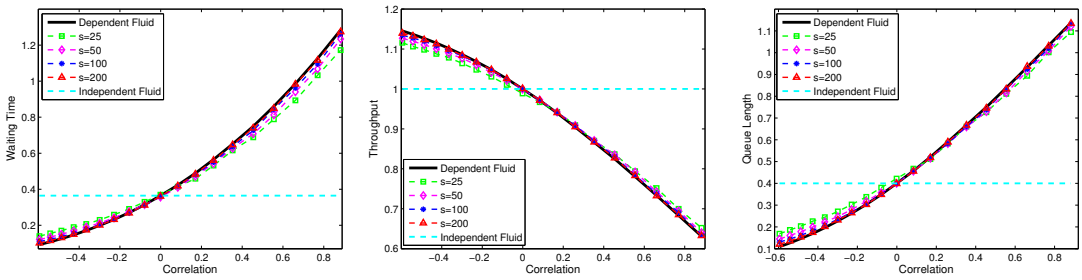
and the offered wait are both zero when the system is not overloaded, i.e., when $\rho \leq 1$. *See also Proposition 2.2 below.*

2.4.2. Numerical Examples

I now examine the accuracy of the fluid approximation for overloaded systems via simulation. To conduct the numerical experiments, I vary the size of the system (number of agents) from 25 to 200 and the arrival rate such that $\rho = 1.2$ for all the systems I consider. In the first numerical study, depicted in Figure 2.2, the arrival process is Poisson with rate λ and the service time S and the patience time T are exponentially distributed with means 1 and 2, respectively. (Recall that the number-in-system process is not Markovian, so that steady-state quantities cannot be computed for the stochastic systems.) To move away from the exponential assumption, I perform another numerical study, depicted in Figure 2.3, in which I consider a renewal arrival process with Erlang$(2, 2\lambda)$ interarrival-time distribution (namely, Erlang with a shape parameter 2 and a rate parameter 2λ, so that the arrival rate is λ); service time S is lognormal with $LN(1, 2)$; and the patience time T is lognormal with $LN(2, 2)$, where I use $LN(a, b)$ to denote the lognormal distribution with mean a and variance b. (Note that the mean service time is 1 and mean time to abandon is 2 for the given lognormal distributions.) In both numerical studies I plot the simulated average waiting time of *served customers*, the average throughput and average queue length in steady state, and compare those simulation results (curves indicated by the number of agents s) to the corresponding fluid estimates (the 'Dependent Fluid' curves). The fluid estimates are obtained by numerically computing ϕ in (2.6) and solving $\bar{w}$ with a bisection search. The throughput and queue length are both plotted

scaled by the number of agents s. To compare the result to the independent model, I also plot the fluid estimates of the independent model (the 'Independent Fluid' curves).

It is clear from the simulations that the fluid model is accurately predicting the steady-state metrics of overloaded systems, even for relatively small systems (with 25 agents), and that the accuracy does not depend on exponential-distributions and Poisson-process assumptions. Further, as was already demonstrated in §2.1, the independent model does not give useful approximations even for systems with moderate dependence.

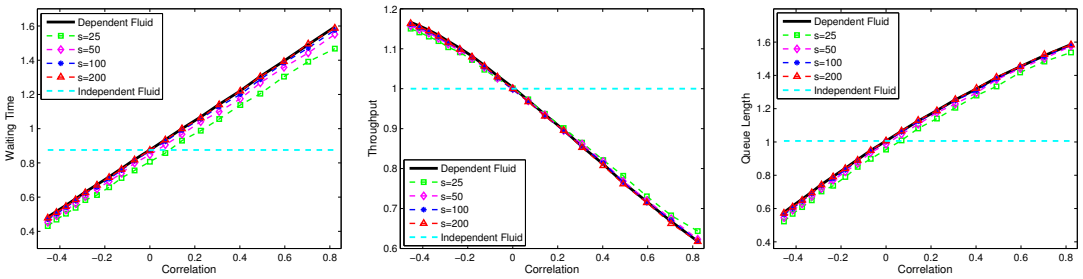


Waiting Time Throughput Queue Length

Figure 2.2. Simulation and fluid model under different system sizes and dependencies, $\rho = \lambda/s\mu = 1.2$, $s \in \{25, 50, 100, 200\}$. Poisson arrival with rate λ, service time distribution exp(1), and patience time distribution exp(1/2). (The joint distribution of service time and patience time is generated via Guassian copula.)

I next demonstrate how the effective traffic intensity ρ_{eff} in (2.5) changes with the nominal traffic intensity ρ and the joint distribution f; the results are shown in Table 2.2. As before, S and T are taken to be exponentially distributed with means 1 and 2, respectively, and two different joint distributions are generated via Gaussian copulas, one with $r = -0.4$ and the second with $r = 0.4$.

It is seen that even moderate dependence (as captured by the correlation) may have a large impact on the effective system load. For example, when $\rho = 1.2$ and $r = -0.4$,



Waiting Time Throughput Queue Length

Figure 2.3. Simulation and fluid model under different system sizes and dependencies, $\rho = \lambda/s\mu = 1.2$, $s \in \{25, 50, 100, 200\}$. Interarrival time distribution $Erlang(2, 2\lambda)$, service time distribution $LN(1, 2)$, patience time distribution $LN(2, 2)$. (The joint distribution of service time and patience time is generated via Guassian copula.)

Table 2.2. A comparison of ρ_{eff} for different ρ, $\rho \in \{1, 1.05, 1.1, 1.2.1.3, 1.5\}$

	1	1.05	1.1	1.2	1.3	1.5
ρ_{eff} (r=-0.4)	1.0	1.02	1.04	1.08	1.13	1.22
ρ_{eff} (r=0.4)	1.0	1.12	1.23	1.42	1.60	1.97

the effective traffic intensity is only $\rho_{\text{eff}} = 1.08$. (A system with a traffic intensity of 1.08 can be considered to be critically loaded, and not overloaded, for practical purposes; see Garnett et al. (2002).) On the other hand, when $\rho = 1.1$ and the dependence is positive with $r = 0.4$, the system is effectively severely congested with $\rho_{\text{eff}} = 1.23$. These differences have significant economic consequences: When $\rho = 1.2$, approximately 16.7% of the customers are expected to abandon in the independent model (since a proportion $(1.2 - 1)/1.2 \approx 0.167$) of the arrivals abandons), but only about 7.4% (a proportion $(1.08 - 1)/1.08$) end up abandoning in my example with negative correlation. In contrast, when $\rho = 1.1$ roughly 9% of the customers are expected to abandon in the independent model, but 18.7% are expected to abandon in my example with positive correlation. I study the economic aspect of my results in §2.6 below in the context of optimal staffing.

2.5. Performance Analysis

Recall that the fluid model is fully characterized by the primitive data set $\mathcal{D} = (\lambda, s, f)$. In this section, I analyze the impact of each of the three components in $\mathcal{D}$ on the fluid system by fixing the other two components. In particular, for a given joint distribution f, in §2.5.1 I study the effect of changes to the arrival rate λ when s is fixed, and the effect of changing the staffing level s, when λ is fixed, on the throughput. Next, in §2.5.2 I quantify how the throughput is impacted by the dependence structure, employing the PQD order and Gaussian copula discussed in §§2.3.1 and 2.3.2. To this end, I fix λ and s and the two marginal densities f_S and f_T, and vary the joint distribution f.

However, I first prove that it is sufficient to know the value of the nominal traffic intensity, equivalently, the values of λ, s and μ, in order to determine whether the system is overloaded. (The system is considered to be overloaded if $\bar{w} > 0$.) I have already observed that negative dependence of S and T decreases the load of the system relative to the independent case. On the other hand, it is not immediately clear whether $\rho \leq 1$ implies that $\bar{w} = 0$ when S and T are positively dependent. Specifically, a self-sustained overload may exist in this case, because a large initial queue leads to a slow effective service rate, which in turn leads to having a large queue. The next proposition shows that the nominal traffic intensity determines whether the fluid model is overloaded. In particular, a stationary fluid system with negative dependence remains overloaded if $\rho > 1$, and overloads cannot be self-sustained when $\rho \leq 1$.

Proposition 2.2. *The following three statements are equivalent:*

i) *The nominal traffic intensity is strictly greater than one;* $\rho > 1$.

ii) *The effective traffic intensity is strictly greater than one;* $\rho_{\mathit{eff}} > 1$.

iii) *The offered wait is strictly greater than zero;* $\bar{w} > 0$.

Throughout this section I assume that $\rho > 1$.

Let

$$g(w) := \mathbb{E}[S|T = w]. \tag{2.10}$$

I refer to the function g as the Conditional Service Time (CST). Then an Increasing Conditional Service Time (ICST) implies a positive dependence, whereas a Decreasing Conditional Service Time (DCST) implies a negative dependence, between S and T. The independence between S and T implies a Constant Conditional Service Time (CCST).

In general, for a given bivariate random variable (S, T), the CST need not be a monotone function. In Appendix A.1.3 I provide natural sufficient conditions for Monotone Conditional Service Time (MCST), and link the monotonicity of the CST to PQD and Gaussian copula introduced in §2.3. In particular, Lemma A.4 states that, for $(S, T) \in \mathcal{G}$, $r > 0$ implies that (S, T) is PQD and has an ICST, whereas $r < 0$ implies that (S, T) is NQD and has a DCST. This monotonicity of the CST can be observed in Figure 2.4, which plots curves of the CST for different bivariates in $\mathcal{G}$. In this figure, the marginal service and patience times S and T are exponential random variables with means 1 and 2, respectively.

2.5.1. Impact of Arrival Rate and Service Capacity on Performance Measures

I now analyze the effects of the arrival rate λ and number of agents s on the throughput R. In a congested system with nonnegligible offered waits, the served customers are also the

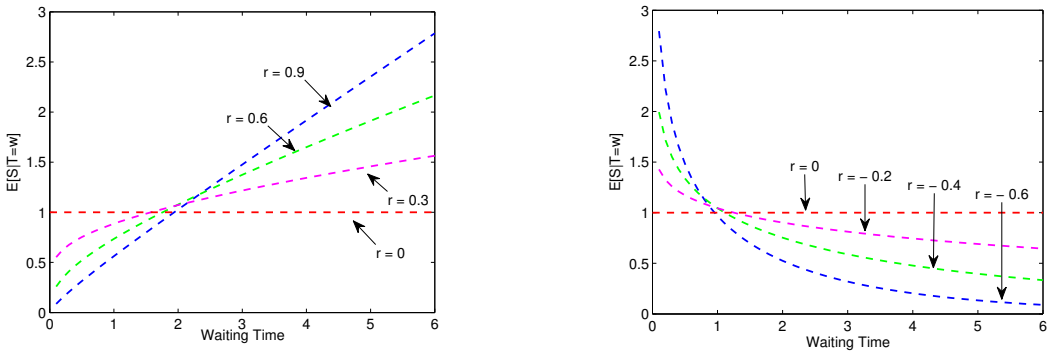


Figure 2.4. Conditional service time under different distributions generated by Gaussian copula. Positive dependence (left), $r > 0$ and ICST. Negative dependence (right), $r < 0$ and DCST. Independent case, $r = 0$ and CCST.

more patient customers. If S and T are positively dependent, served customers also tend to require relatively long service times, so that, as the arrival rate increases, the offered wait and, in turn, the effective mean service time, increase as well, so that throughput decreases. On the other hand, when the dependence is negative, served customers tend to require short service times. As the the arrival rate λ increases, the offered wait increases, leading to more abandonment, and therefore, higher effective service rate and throughput. In either case, as the next proposition shows, if f has an MCST, then the throughput R is a monotone function of λ. Specifically, for given s and f, let $R(\lambda)$ be the throughput when the arrival rate is λ. The assumption $\rho > 1$ implies that the domain of $R(\lambda)$ is $(s\mu, \infty)$.

Proposition 2.3. *$R(\lambda)$ is decreasing if f has an ICST and is increasing if f has a DCST.*

An important managerial insight that follows from Proposition 2.3 is that congestion does not necessarily lead to performance degradation. In particular, if f has a DCST, then

waiting "strains" the customers that have short patience times and long service times, thus increasing the effective service rate and the throughput. This self-selection of the customers can be exploited by appropriately staffing the system, as I will show in §2.6.

The following corollary follows immediately from Lemma A.4 and Proposition 2.3.

Corollary 2.1. *For $(S,T) \in \mathcal{G}$, $R(\lambda)$ is decreasing if $r > 0$, and $R(\lambda)$ is increasing if $r < 0$.*

I next consider the throughput as a function of the capacity when the arrival rate is fixed. To this end, let $R(s)$ denote the throughput as a function of the capacity s when λ and f are fixed. The assumption $\rho > 1$ implies that $0 \leq s < \lambda/\mu$, namely, the domain of $R(s)$ is $[0, \lambda/\mu)$.

Proposition 2.4. *$R(s)$ is convex increasing if f has an ICST and is concave increasing if f has a DCST. In particular, $R(s)$ is linear if f has a CCST.*

Unlike Proposition 2.3, in which the monotonicity of the throughput in λ depends on the dependence structure, the throughput is always increasing in s, regardless of the dependence, when λ is fixed. In the special case with independent service and patience times, the relation between the throughput and the capacity is linear. The structural properties of $R(s)$, stated in Proposition 2.4, facilitate the analysis of the staffing problem for revenue maximization in §2.6.

The intuition behind the fact that the throughput grows at a rate faster/slower than capacity s when f has an MCST, can be explained as follows. If f has an ICST, then as capacity s increases, the offered wait $\bar{w}$ decreases so that the effective service rate μ_{eff} increases. The throughput $s\mu_{\text{eff}}$ thus increases superlinearly in s. On the other hand, if f

has a DCST, the effective service rate μ_{eff} decreases with s and thus the throughput $s\mu_{\text{eff}}$ grows sublinearly in s.

For the Gaussian copula, I obtain the following corollary to Proposition 2.4.

Corollary 2.2. *For $(S,T) \in \mathcal{G}$, $R(s)$ is convex increasing if $r > 0$, and $R(s)$ is concave increasing if $r < 0$.*

2.5.2. Impact of Dependence between Service and Patience on Performance

I now consider how the strength of the dependence, as ranked by PQD order, impacts system performance. To this end, I fix the arrival rate λ and the number of agents s, as well as the marginals f_S and f_T. Let (S_1,T_1) and (S_2,T_2) denote two bivariate random variables both in a subset $\mathcal{P}(f_S,f_T)$ of $\mathcal{F}(f_S,f_T)$ whose elements can be ranked by PQD order (see §2.3.1). Let R_i, w_i and Q_i denote the throughput, offered wait and stationary queue, respectively, in the fluid model of a system with joint service time and patience (S_i,T_i), $i = 1,2$. The next result validates the intuition that the throughput is smaller under positive dependence and larger under negative dependence.

Proposition 2.5. *If $(S_1,T_1) \leq_{PQD} (S_2,T_2)$, then $R_1 \geq R_2$, $w_1 \leq w_2$ and $Q_1 \leq Q_2$.*

It is significant that the statement in Proposition 2.5 can be strengthened if one considers particular families of joint distributions with given marginals. In particular, if both bivariate random variables are generated via a Gaussian copula, then the inequalities in the statement are strict, as the next result shows.

Corollary 2.3. *If $(S_1,T_1),(S_2,T_2) \in \mathcal{G}(f_S,f_T)$ and $r_1 < r_2$, then $R_1 > R_2$, $w_1 < w_2$ and $Q_1 < Q_2$.*

2.5.3. Numerical Examples

I first demonstrate the statement of Proposition 2.3 and Corollary 2.1. In Table 2.3 I compare the throughput of different systems where the capacity s is fixed at 100 and the nominal traffic intensity ρ increases from 1.0 to 1.5 as the arrival rate λ varies. I also compare the throughput calculated by my fluid model with those observed by simulations. In the example, (S, T) is generated via a Gaussian copula, with S and T being exponentially distributed with means 1 and 2, respectively. The gaps between the fluid predictions and the simulated values of the throughput are also reported. It is readily seen that the throughput is increasing in λ when $r = -0.4$ (representing negative dependence), and is decreasing in λ when $r = 0.4$ (representing positive dependence). Moreover, the changes to the throughput as λ increases are substantial. I remind the readers that for any $\rho \geq 1$, the throughput is fixed at $s\mu = 100$ when S and T are independent.

I note that the gaps between the fluid estimates and the corresponding simulation experiments are the largest when the system is critically loaded ($\rho = 1$), because stochastic fluctuations, which are of lower order than the dynamics captured by the fluid model (see Remark 2.1), play a dominant role when the fluid estimate is zero for the system. I consider the staffing problem in the next section and propose a heuristic refinement that is based on diffusion approximations for critically loaded systems.

I next validate the result in Proposition 2.4. Figure 2.5 compares the throughputs obtained from simulations (discrete marks) and from fluid models (dashed line) under different capacities and joint distributions. I vary the capacities from 10 to 90 while keeping the arrival rate fixed at $\lambda = 100$. Service time S and patience time T are exponentially distributed with means 1 and 2, respectively, and the bivariate (S, T) is

Table 2.3. A comparison of throughputs under different nominal traffic intensities ($s = 100$)

		$r = -0.4$			$r = 0.4$		
		Throughput		Gap	Throughput		Gap
λ	ρ	Simulation	Fluid	Percentage	Simulation	Fluid	Percentage
100	1	98.01	100.00	2.03%	94.34	100.00	6.00%
105	1.05	101.66	103.20	1.52%	93.09	93.44	0.38%
110	1.1	104.82	106.00	1.13%	90.08	89.79	0.33%
120	1.2	110.26	110.96	0.63%	84.78	84.75	0.03%
130	1.3	114.89	115.37	0.41%	81.17	81.16	0.01%
150	1.5	122.75	123.08	0.27%	76.12	76.11	0.01%

generated via Gaussian copulas for different values of r. The convexity of $R(s)$ when $r > 0$ and the concavity of $R(s)$ when $r < 0$ are apparent. When service and patience times are independent, so that $r = 0$, $R(s)$ linearly increases in s (solid lines).

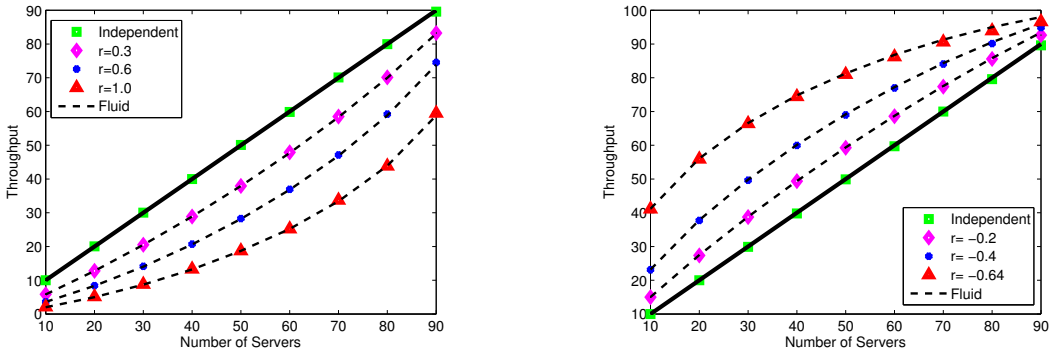
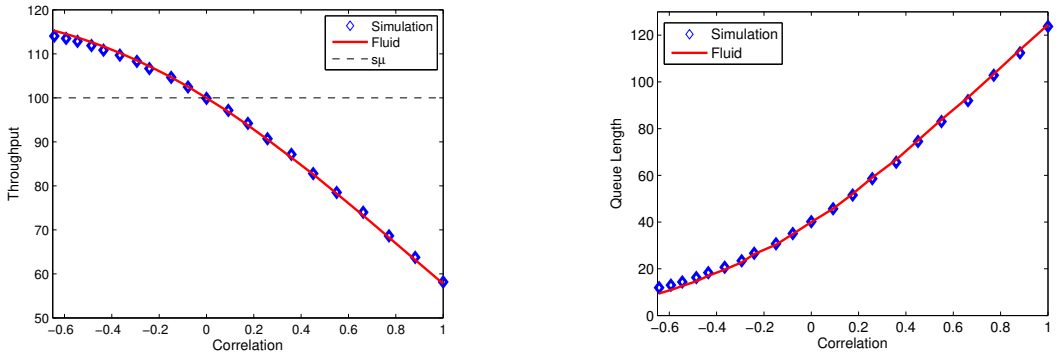


Figure 2.5. A comparison of throughputs under different capacities ($\lambda = 100$, s ranges from 10 to 90). Positive dependence (left figure): throughput convex increasing with s. Negative dependence (right figure): throughput concave increasing with s. The independent case (solid lines with squares): throughput is linear increasing with s.

Finally, I numerically validate the result in Proposition 2.5. I take S and T as in the former two examples, and again employ a Gaussian copula to generate their joint distribution. I fix $\lambda = 120$ and $s = 100$, and plot the throughput as a function of the

correlation coefficient r. Figure 2.6 reveals the significant impact of the dependence on the system performance. In particular, the throughput when $r = 1.0$ is only half of that under $r = -0.64$ (which is the minimal attainable correlation coefficient when the two marginals are exponentially distributed). The increase in the average queue length is even more salient: The fluid queue increases from 11.9, when $r = -0.64$, to 123.7 when $r = 1$.



Throughput Queue Length

Figure 2.6. A comparison of throughputs and queue lengths for different systems with service and patience times generated by Gaussian copulas.

2.6. Economics of Capacity Sizing

In this section I apply the results derived for the stationary fluid model to develop fluid-optimal solutions to a capacity-sizing problem under a linear cost structure. I start in §2.6.1 by considering the optimal staffing under the FIFO policy[‡]. It is significant that the analysis I apply was performed for overloaded systems having $\rho > 1$ (recall Proposition 2.2), but that it is sometimes optimal to staff the system so as to have it be critically loaded, namely, have $\rho = 1$; see Proposition 2.6 below. In the later case, my fluid model is too crude an approximation for the stochastic system (since the queue, and

[‡]The analysis in §2.6.1 is extended in Appendix A.1.3 to consider the optimal control policies.

thus the proportion of abandonment, are both null in the fluid model of critically loaded systems), and stochastic refinements must be considered. Thus, in §2.6.3 I propose a heuristic refinement based on existing approximations for critically loaded systems. The effectiveness of the fluid-based and the heuristic prescriptions are verified via simulations.

2.6.1. Capacity Sizing under FIFO Policy

I study the capacity-sizing problem when linear staffing and abandonment costs are incurred. Let c denote the unit cost of capacity and let p denote the penalty associated with an abandonment. For a given arrival rate λ, I consider the following cost optimization problem for the fluid system:

$$\min_{s\geq 0} C_\lambda(s) := cs + p\alpha_\lambda(s), \tag{2.11}$$

where $\alpha_\lambda(s)$ is the abandonment rate when the arrival rate is λ and capacity is set to s. The penalty for abandonment can be considered as the opportunity cost of a lost customer, or as the reputation cost resulting from customer dissatisfaction. Hence the cost function $C_\lambda(s)$ is a combination of the personnel cost incurred by capacity allocation and the customer-related cost induced by abandonments.

Equivalently to (2.11), I can maximize the profit function $\Pi_\lambda(s) := pR_\lambda(s) - cs$, where $R_\lambda(s)$ is the throughput when the arrival rate is λ and capacity is s. In the standard model (with independent service and patience), the throughput is $R_\lambda(s) = \min\{\lambda, s\mu\}$, so that $\Pi_\lambda(s) = p\min\{\lambda, s\mu\} - cs$. Clearly, an optimal solution to the problem $\min_{s\geq 0}\Pi_\lambda(s)$ cannot have the number of agents s be larger than the offered load λ/μ, for otherwise the cost of the "extra capacity" $s-\lambda/\mu$ can be eliminated without reducing the

throughput (the throughput stays λ as long as $s \geq \lambda/\mu$). In the independent model, the optimal capacity is trivial to compute because the profit-maximization problem reduces to maximizing $(p\mu - c)s$, which is positive if and only if $p\mu > c$. In the latter case, the optimal capacity is clearly $s^*_\lambda = \lambda/\mu$. See similar results in Whitt (2006b) and Ren and Zhou (2008).

When service times and patience are dependent, the throughput is determined by their joint distribution, in addition to the arrival rate and staffing, so that the optimal-staffing problem is no longer trivial. Nevertheless, similar to the independent case, it is easy to see that the optimal capacity s^*_λ must satisfy $s^*_\lambda \leq \lambda/\mu$, implying that (2.11) is equivalent to

$$\min_{0 \leq s \leq \lambda/\mu} C_\lambda(s) = cs + p\alpha_\lambda(s). \tag{2.12}$$

Note that $\alpha_\lambda(s) = \lambda F_T(\bar{w})$, where $\bar{w}$ solves (2.3), so that

$$C_\lambda(s) = cs + p\alpha_\lambda(s) = \lambda \left[c\phi(\bar{w}) + pF_T(\bar{w})\right] = \lambda \left[cF^c_T(\bar{w})a(\bar{w}) + pF_T(\bar{w})\right].$$

I can equivalently optimize over w and restate the optimization problem:

$$\min_{\bar{w} \geq 0} \bar{C}_\lambda(\bar{w}) := cF^c_T(\bar{w})a(\bar{w}) + pF_T(\bar{w}). \tag{2.13}$$

Differentiating $\bar{C}_\lambda(\bar{w})$ with respect to $\bar{w}$ gives $\bar{C}'_\lambda(\bar{w}) = f_T(\bar{w})\left(p - cg(\bar{w})\right)$, and setting the derivative to zero gives us the following first order condition: $g(\bar{w}) = p/c$. To interpret the latter equality, note that $p/g(\bar{w})$ represents the marginal revenue of adding capacity; in optimality, this marginal revenue must equal the marginal cost c of added capacity.

The derivation above gives rise to the following proposition. Let $g(\infty)$ denote the limit of $g(\bar{w})$ as $\bar{w} \to \infty$, whenever the limit exists.

Proposition 2.6. *Under FIFO,*

(i) If f has an ICST, then the critically loaded regime with capacity $s^*_\lambda = \lambda/\mu$ is fluid optimal if and only if $c < p\mu$. Otherwise if $c \geq p\mu$, then no capacity should be allocated.

(ii) If f has a DCST, then the overloaded regime is fluid optimal if and only if $g(\infty) < p/c < g(0)$. In this case, the optimal capacity is $s^*_\lambda = \lambda F_T(w^*)a(w^*)$, for $w^* := g^{-1}(p/c)$. Otherwise if $p/c \geq g(0)$, then the critically loaded regime with capacity $s^*_\lambda = \lambda/\mu$ is fluid optimal. If $p/c \leq g(\infty)$, then no capacity should be allocated.

I remark that the conditions in the second part of the proposition are always satisfied when (S, T) is generated by a Gaussian copulas with $r < 0$.

As was discussed above, when $p\mu < c$ then service is unprofitable in the independent model. The same is true for systems with positive dependence, because the throughput in such a system is no larger than the throughput $s\mu$ of the independent model. However, Proposition 2.6 shows that when f has a DCST, the effective service rate, and thus the throughput, can be sufficiently high to warrant service profitable *even when* $p\mu < c$.

Proposition 2.6 is concerned with the structure of the dependence in a given system. The next result considers the comparative statics focusing on the dependence measured by the PQD order. To state the result, recall the setting of Proposition 2.5. In particular, fix the arrival rate λ, capacity s, and the marginal densities f_S and f_T. Let (S_1, T_1) and (S_2, T_2) be two bivariate random variables in a set $\mathcal{P}(f_S, f_T) \subseteq \mathcal{F}(f_S, f_T)$ whose elements

can be ranked by PQD order. For $i = 1, 2$, let C_i^* denote the optimal cost when the service time and patience are distributed as S_i and T_i, respectively.

Corollary 2.4. *If* $(S_1, T_1) \leq_{PQD} (S_2, T_2)$, *then* $C_1^* \leq C_2^*$.

It follows from Corollary 2.4 that the optimal cost is monotone in the dependence strength. However, an analogous result for the optimal staffing does not necessarily hold, as will be seen in the numerical example presented in Table 2.4 below. Nevertheless, one intuitively expects that when abandonments are "too costly," namely, if the abandonment penalty p is sufficiently large relative to the staffing cost c, then a stronger dependence will also imply a larger optimal staffing level, because a stronger dependence implies increased abandonment for any given staffing level. This intuition is formalized in the next proposition. To state it, I need the following definition. Let h be a real-valued function. I say that h satisfies the *principle of permanence* at $z = 0$ when the following holds: if there exists a positive sequence $\{z_n : n \geq 1\}$ of distinct numbers such that $z_n \to 0$ as $n \to \infty$ and $h(z_n) = 0$ for all n, then $h(z) = 0$ in a neighborhood of $z = 0$. In particular, h cannot have infinitely many roots in any finite interval containing 0 unless it is identically equal to 0 over such interval.

Consider the setting of Corollary 2.4, and let $s_i^*(p/c)$ denote the optimal capacity as a function of the penalty-cost ratio p/c, when the service-time and patience are (S_i, T_i). Let $g_i(z)$ denote the corresponding conditional expectation, defined in (2.10), $i = 1, 2$.

Proposition 2.7. *Assume that (i)* $(S_1, T_1) \leq_{PQD} (S_2, T_2)$*; (ii)* f_i *has a DCST,* $i = 1, 2$*; (iii)* $h(z) := g_1(z) - g_2(z)$ *satisfies the principle of permanence at* $z = 0$. *Then there exists* M *satisfying* $0 < M < g_2(0)$ *such that* $s_1^*(p/c) \leq s_2^*(p/c)$ *for all* $p/c \in (M, g_2(0))$.

I note that Condition (iii) in Proposition 2.7 is a weak technical condition ensuring that g_1 and g_2 do not cross infinitely many times in the neighborhood of 0. Any of the following three conditions is sufficient for (iii) to hold: (1) $h(0) \neq 0$ (which typically holds); (2) if $h(0) = 0$, then $h'(0) \neq 0$; (3) h admits a Taylor series expansion at 0.

If the bivariates (S, T) are generated via a Gaussian copula, the monotonicity of the optimal staffing in Proposition 2.7 is strict, as stated in the following corollary.

Corollary 2.5. *If for* $(S_1, T_1), (S_2, T_2) \in \mathcal{G}(f_S, f_T)$ *it holds that* $r_1 < r_2 < 0$*, then there exists* $M > 0$ *such that* $s_1^*(p/c) < s_2^*(p/c)$ *for all* $p/c > M$*.*

2.6.2. Numerical Study

I now present numerical and simulation examples to demonstrate the accuracy and the limitations of the optimal fluid solution to Problem (2.11) described in Proposition 2.6. The system I consider has a Poisson arrival process with arrival rate $\lambda = 100$; the marginal service time and patience distribution are exponentially distributed with means 1 and 2, respectively; and the joint distributions of service and patience times are generated via Gaussian copulas with correlation coefficients ranging from -0.64 to 1. (Recall that for Gaussian copulas the correlation coefficient determines the joint distribution, and that $r = 0$ corresponds to the independent case. Moreover, $r = -0.64$ is the minimum attainable correlation coefficient for exponential marginals.) In the three examples, I fix $c = 1$ and vary the penalty p; in particular, I consider the values $p = 0.8$, $p = 1.25$ and $p = 3.5$. Note that, in the first case (with $p = 0.8$) service is not profitable in the independent and positively dependent models. On the other hand, $p = 3.5$ represents an extreme case of a high abandonment penalty.

In Table 2.4 I compare the fluid-optimal capacity and cost (shown in the 'Fluid Optimal' column) to the corresponding optimal values obtained from simulation experiments (these appear in the 'Simulation Optimal' column). The optimality gap between the fluid prescription and the true optimum is shown in the third column of the table. The simulation results are based on 10 independent runs; when a critically loaded regime is fluid optimal, each run lasts for 20,000 time units with the first 10,000 time units serving as the warm-up period. For overloaded systems, each run stops after 3,000 time units with the first 1,000 time units serving as the warm-up period. Before elaborating on the numerical results, I make the following quick observations: First, when $p = 0.8$ (so that $p\mu < c$), *operations can be profitable when the dependence is negative*, provided the staffing is done correctly, even though it is not profitable to operate when service and patience times are independent or positively dependent. I also observe that the optimal staffing is not monotone in the correlation r (and thus in the dependence strength) when $p = 0.8$, but is monotone for the other two cases with larger values of p; see Corollary 2.5. Finally, the optimality gap is relatively negligible in the overload regime, but the gap can be large when the system is critically loaded, in particular, when the dependence is strong and positive.

More specifically, when the service time and patience are negatively dependent, it is optimal to operate in the overload regime. In this regime, the fluid queue serves as a first-order approximation for the queue process, and the stochastic fluctuations about the fluid are of lower order, and so are negligible in large systems. As a result, the optimality gap between the optimal fluid prescription and the true optimum, as evaluated via the simulations, is negligible. However, there are considerable optimality gaps when the

Table 2.4. Optimal staffing of systems with dependencies ($\lambda = 100$)

Correlation	Fluid Optimal		Simulation Optimal		Optimality Gap	
r	Capacity	Cost	Capacity	Cost	Absolute	Percentage
$\mathbf{p/c = 0.8}$						
-0.64	19	55.18	19	55.18	0.00	0.0%
-0.6	20	58.29	20	58.29	0.00	0.0%
-0.4	22	69.78	21	69.78	0.00	0.0%
-0.2	14	78.81	14	78.81	0.00	0.0%
0 to 1	0	80.00	0	80.00	0.00	0.0%
$\mathbf{p/c = 1.25}$						
-0.64	36	71.62	36	71.62	0.00	0.0%
-0.6	39	75.25	38	75.23	0.02	0.0%
-0.4	55	88.63	54	88.58	0.04	0.1%
-0.2	79	98.02	78	97.93	0.09	0.1%
0	100	104.14	92	102.71	1.43	1.4%
0.2	100	105.40	97	105.20	0.20	0.2%
0.4	100	106.95	102	106.91	0.04	0.0%
0.6	100	109.06	102	108.12	0.94	0.9%
0.8	100	111.86	105	108.99	2.88	2.6%
1	100	115.40	106	109.47	5.94	5.4%
$\mathbf{p/c = 3.5}$						
-0.64	86	101.18	83	100.99	0.19	0.2%
-0.6	91	102.78	87	102.42	0.36	0.4%
-0.4	100	106.86	98	106.55	0.31	0.3%
-0.2	100	108.91	102	108.72	0.19	0.2%
0	100	111.59	104	110.22	1.38	1.3%
0.2	100	114.98	106	111.27	3.71	3.3%
0.4	100	119.46	108	112.17	7.30	6.5%
0.6	100	125.37	108	112.76	12.61	11.2%
0.8	100	133.22	109	113.32	19.90	17.6%
1	100	143.13	110	113.63	29.50	26.0%

dependence is positive, and the fluid-optimal solution for the staffing problem puts the system in the critically loaded regime. In this case, the stochastic fluctuations, which are not captured by the fluid model, become dominant. As should be expected, the optimality gap increases as the cost of abandonment and the strength of the dependence increase. In particular, when the dependence is strong ($r \geq 0.6$) and abandonment cost

is high ($p = 3.5$), the optimality gap is too substantial for the optimal fluid staffing to be considered a useful guideline.

Even though $p = 3.5$ represents an extreme case of a high abandonment penalty relative to the staffing cost, the results in Table 2.4 suggest that taking stochasticity into account can lead to substantial improvements in critically loaded systems, even more so than in the independent model. However, studying the optimal staffing problem in this setting requires a refined second-order (diffusion type) approximation to the system, which is beyond the scope of this chapter. I mention that extensive simulation experiments suggest that the safety capacity needed in order to achieve optimality in the critically loaded regime is of order $\sqrt{\lambda}$, which is consistent with diffusion approximations for many-server queueing systems without dependence (see Halfin and Whitt (1981) and Garnett et al. (2002)). In the next section I propose an algorithm to compute effective staffing recommendations for critically loaded systems with dependencies that are based on my characterization of the effective service rate combined with existing results for the independent model.

2.6.3. A Heuristic Stochastic Refinement for the Critically loaded Case

To refine the first-order staffing recommendation prescribed by the fluid model when the service time and patience are positively dependent, I propose the following algorithm, based on the diffusion approximation for the independent model (the Erlang A) in Garnett et al. (2002). Consider a system having Poisson arrivals with rate λ, exponential service time with rate μ, exponential patience time with rate θ and a given joint distribution for the service time and patience.

(i) Use the stationary diffusion approximation for the critically loaded Erlang-A in Garnett et al. (2002), and in particular, the formula for the proportion of abandonment in p. 218 of this reference: For a service system with s agents, define $\beta = \frac{s-\lambda/\mu}{\sqrt{\lambda/\mu}}$. Then the abandonment ratio can be approximated by

$$\mathbb{P}(\text{Ab}) \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\mu/\theta})}\right]\left[1 + \frac{h(\beta\sqrt{\mu/\theta})}{\sqrt{\mu/\theta}h(-\beta)}\right]^{-1}, \tag{2.14}$$

where h is the hazard function of the standard normal random variable. Approximate the abandonment rate $\alpha_\lambda(s) = \lambda \cdot \mathbb{P}(\text{Ab})$ and compute the optimal staffing level s_0 that solves (2.11) (without dependence). Let $\mathbb{P}^*(\text{Ab})$ denote the proportion of abandonment under s_0 in the Erlang-A model.

(ii) For the dependent model under consideration, compute the fluid waiting time w^* for which the proportion of abandonment is equal to $\mathbb{P}^*(\text{Ab})$ computed in (i), namely, for which $F_T(w^*) = \mathbb{P}^*(\text{Ab})$. Compute the effective service rate $\mu^*_{\text{eff}} = 1/a(w^*)$.

(iii) Employ the approximation in (2.14) once again, this time with service rate μ^*_{eff}, and compute the capacity s^* for which the proportion of abandonment is equal to $\mathbb{P}^*(\text{Ab})$.

Note that s^* computed in Step (iii) is of the form $s^* = \lambda + \beta^*\sqrt{\lambda}$, for some $\beta^* \in \mathbb{R}$. Then the proposed number of agents in the real system is $\lceil s^* \rceil$, namely, the smallest integer larger than s^*.

Numerical Example. Table 2.5 demonstrates the substantial improvements obtained by employing the procedure above. In this table, the capacity and resulting cost

obtained using my staffing algorithm is compared with the fluid prescriptions and the optimal values, which are estimated via simulations. Observe in particular, that the optimality gap in the cost reduces to 1.8% under my heuristic when $p = 3.5$ and $r = 1.0$, compared to 26% under the fluid prescription.

Table 2.5. Optimal staffing of systems with dependencies: simulations, fluid prescriptions and heuristic ($\lambda = 100$, $\mu = 1$, $\theta = 1/2$, $c = 1$, $p = 3.5$)

Correlation	Optimal	Fluid Model		Heuristic	
r	Capacity	Capacity	Cost Gap	Capacity	Cost Gap
-0.4	98	100	0.3%	101	0.5%
-0.2	102	100	0.2%	103	0.2%
0	104	100	1.2%	104	0.0%
0.2	106	100	3.3%	105	0.2%
0.4	108	100	6.5%	105	0.5%
0.6	108	100	11.2%	106	0.6%
0.8	109	100	17.6%	106	1.2%
1	110	100	26.0%	106	1.8%

2.7. Square-root Staffing under Dependent Service and Patience Times

The queueing literature has demonstrated that the square-root staffing rule in the many-server setting allows the service manager to to achieve two desirable goals: a high quality of service level in the sense that customers' queueing time is negligible and there is a strictly positive probability that customers do have to wait to get served; as well as a high efficiency in the capacity cost in that almost all service agents are busy all the time so that the utilization of service agents is close to 1. See e.g., Halfin and Whitt (1981) and Garnett et al. (2002). As a result, the square-root staffing regime is also termed as the Quality-and-Efficiency Driven (QED) regime by Borst et al. (2004).

In the presence of a dependence between service and patience times, however, I conjecture via extensive simulations that the square-root staffing may fail to achieve the

desirable QED performance in some circumstances. I find that for systems which have a positive dependence with $g(0) = 0$ and are staffed such that $\lambda^n/\mu = n + \beta\sqrt{n}$ for some fixed $\beta > 0$, the queue length can not be order of $O(\sqrt{n})$, an indicator suggesting the failure of the QED performance. In this case, the probability that an arriving customer will have to wait in queue is nearly 1. The intuition behind these facts is straightforward and similar to what I have explained in §2.5. The positive dependence puts more workload into the system compared to the independent model since customers who end up receiving service are those with high patience levels and thus longer-than-average service requirements.

Although I am unable to prove my argument rigorously, in this section I provide a heuristic to support my argument. I consider a sequence of systems having service and patience times distributed with a common distribution f, which does not vary with system size. The arrival rate λ^n and the number of service agents n in the n^{th} system are related such that $\lambda^n/\mu = n + \beta\sqrt{n}$ for some fixed $\beta \in (0, +\infty)$. Two performance metrics in the QED regime in the independent model are of interest. Garnett et al. (2002) show that under the square-root staffing, the waiting time $W^n \sim 1/\sqrt{n}$ and the queue length $Q^n \sim \sqrt{n}$ where the superscript n indicates the corresponding performance metrics in the n^{th} system. Assuming the same positive dependence along the sequence of systems, I will show by contradiction that the waiting time W^n cannot follow the same order as with no dependence. To derive the contradiction, suppose the waiting time W^n is still of order $1/\sqrt{n}$ in the n^{th} system with a positive dependence. Then it must hold that $\mathbb{E}[W^n] = \tilde{w}/\sqrt{n}$ for some $\tilde{w} \sim O(1)$. I will show in the following that the work inflow

into service must exceed the maximal processing capacity, hence a steady state cannot be sustained.

Recall the definition in (2.4) that $a(w) = \mathbb{E}[S|T > w]$ and $g(w) = \mathbb{E}[S|T = w]$. One can compute $a'(0) = f_T(0)\,(1/\mu - g(0))$. Thus, the effective service time in system can be approximated using a first-order Taylor expansion

$$\frac{1}{\mu_{\text{eff}}} \approx a\left(\mathbb{E}[W^n]\right) \approx \frac{1}{\mu} + a'(0)\mathbb{E}[W^n] = \frac{1}{\mu}\left[1 + \frac{f_T(0)(1-\mu g(0))\tilde{w}}{\sqrt{n}}\right].$$

Since the waiting time $\mathbb{E}[W^n] = \tilde{w}/\sqrt{n}$ and is close to 0 for large n, hence the abandonment ratio

$$P(\text{Ab}) \approx h_T(0)\mathbb{E}[W^n] = \frac{f_T(0)\tilde{w}}{\sqrt{n}}.$$

The workload into service is:

$$\begin{aligned}
\mathcal{T} &= \lambda^n \frac{1}{\mu_{\text{eff}}}\left(1 - P(\text{Ab})\right) = \left(n + \beta\sqrt{n}\right)\left(1 + f_T(0)(1-\mu g(0))\frac{\tilde{w}}{\sqrt{n}}\right)\left(1 - f_T(0)\frac{\tilde{w}}{\sqrt{n}}\right) \\
&= \left(n + \beta\sqrt{n}\right)\left(1 - \mu f_T(0)g(0)\frac{\tilde{w}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= n + (\beta - \mu f_T(0)g(0)\tilde{w})\sqrt{n} + o(\sqrt{n}).
\end{aligned}$$

In the case $g(0) = 0$ (for example, when the service and patience times are generated with Gaussian copulas with positive correlation coefficient), since $\beta > 0$, it necessarily follows that $\beta > f_T(0)g(0)\tilde{w}$ for all $\tilde{w} > 0$. Since each service agent is working at unit rate, the workload into service $\mathcal{T}$ exceeds the maximal capacity n. Hence the service flow described above cannot be sustained in a steady state in the QED regime.

Next I show that the above heuristic argument does not violate the predictions of Garnett et al. (2002) in the independent model. In Garnett et al. (2002),

$$\begin{aligned} P(\text{Ab}) &= \left[1 - \frac{h(-\beta\sqrt{\mu/\theta})}{h(-\beta\sqrt{\mu/\theta} + \sqrt{\theta/n\mu})}\right] \omega(\beta, \sqrt{\frac{\mu}{\theta}}) \\ &\approx \frac{h(-\beta\sqrt{\mu/\theta})'\sqrt{\theta/n\mu}}{h(-\beta\sqrt{\mu/\theta})}\, \omega(\beta, \sqrt{\frac{\mu}{\theta}}) \\ &= \left(h(-\beta\sqrt{\mu/\theta}) + \beta\sqrt{\mu/\theta}\right) \sqrt{\theta/n\mu}\, \omega(\beta, \sqrt{\frac{\mu}{\theta}}). \end{aligned}$$

If service and patience times are exponentially distributed with rates $\mu = 1$ and $\theta = 1/2$, respectively, then

$$P(\text{Ab}) = \left(h(-\sqrt{2}\beta) + \sqrt{2}\beta\right) \frac{1}{\sqrt{n}} \frac{h(\beta)}{\sqrt{2}h(\beta) + h(-\sqrt{2}\beta)} > \frac{\beta}{\sqrt{n}}.$$

Figure 2.7 plots $\sqrt{n}(P(Ab) - \beta)$ against β in the independent model. It can be readily seen that $P(Ab) > \beta$ always holds for all β. In fact, one can show $P(Ab) > \beta$ is true for all values of μ and θ. It then follows that the workload into service less than the maximal capacity in that $\mathcal{T} = \lambda^n/\mu(1 - P(\text{Ab})) = n - (P(Ab) - \beta)\sqrt{n} < n$, thus the steady state is stable and can be sustained.

The heuristic argument provides a necessary condition for the overloaded square-root staffing ($\beta > 0$) to achieve QED performance in the presence of a positive dependence: $g(0) > 0$, which is trivially satisfied in the independent model.

2.8. Summary

I considered a queueing model for large service systems in which the patience of customers depends on their individual service times. Since this dependency renders exact

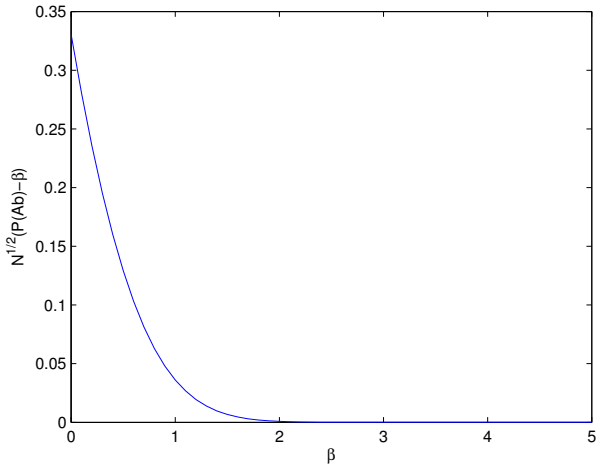


Figure 2.7. Independent case: $\sqrt{N}(P(Ab) - \beta)$ against β.

analysis intractable even if the marginal service time and patience are exponentially distributed, I utilized a stationary fluid model to approximate the system's steady state. That fluid model can be employed to provide accurate approximations of key performance measures of overloaded systems with any jointly continuous service-time and patience distribution, as is demonstrated via simulation experiments. Moreover, since the fluid model is characterized via the full joint distribution of service and patience times, it can be applied to obtain important qualitative results. In particular, I applied the fundamental PQD stochastic order and the Conditional Service Time (CST) to obtain structural results regarding the impact of the dependence on the fluid model. My qualitative results were shown to hold for the important family of Gaussian copulas, which is often employed in practice to analyze joint distributions due to its analytical tractability.

I then implemented the framework I developed to study an optimal staffing problem when staffing and abandonment costs are incurred. The fluid-optimal prescriptions were shown to be very close to the true optimum, as evaluated via simulations, in the overloaded

regime, but the optimality gap can be substantial when the fluid-optimal solution puts the system in the critically loaded regime. To handle that latter case, I proposed a simple algorithm to compute a square-root safety-staffing recommendation, based on a heuristic adjustment of an existing second-order refinement for the Erlang-A (independent) model, together with my characterization for the effective service rate. Numerical examples demonstrate that the proposed heuristic can decrease the optimality gap substantially, even for moderate positive dependencies, when the abandonment penalty (equivalently, the revenue from service) is relatively large.

Future Research There are many directions for related future research. One needs to develop efficient econometric methods to accurately estimate the joint distribution for the service and patience times from data. In doing so, one also needs to carefully address the censoring problem due to customer abandonments, e.g., see Brown et al. (2005). It also remains to formally develop second-order (diffusion-type) approximations for critically loaded systems. Finally, it remains to describe the transient fluid approximation and prove that both the transient and the stationary fluid models hold as weak limits for the stochastic system and its steady state, respectively, in the many-server heavy-traffic regime.

CHAPTER 3

Service Systems with Exogenous and Endogenous Dependencies: A Unified Fluid Model

3.1. Introduction

One of the most prevalent assumptions in the queueing literature is that the primitives of the model (the inter-arrival, service and patience times, for example) are independent of each other, as well as of the state and dynamics of the system. In practice, however, this independence assumption does not always hold. For example, the patience of a customer waiting for an agent to reply in a call center is likely to depend on the importance of the requested service, which in turn, is likely to be correlated with the length of the service; e.g., see Reich (2012). Similarly, the willingness to wait for a checkup in an Emergency Department (ED) is likely to depend on the acuity level of the waiting patient, which itself is correlated with the service (treatment) time that patient requires; e.g., see Lovett et al. (2014). In these two examples, the service requirement of each customer depends on his or her own (im)patience; in other settings, however, it is likely to depend on the customer's delay in queue. For example, Chan et al. (2016) provide empirical evidence that the hospitalization times at some Intensive-Care Units (ICUs) are positively dependent on the times that the patients waited for an available bed at those ICUs.

In this chapter, I study the queueing dynamics of two different models, each corresponding to one of the aforementioned dependence structures. When customers' service

times depend on their individual patience, I consider the dependence to be exogenously "brought into the system" by arriving customers, and refer to that model as *exogenous-dependence* model. When the service time of each customer depends on that customer's delay in queue, I treat each customer as arriving with a family of *conditional* service-time distributions (conditional on the delay), with the actual service-time distribution being determined by the realized delay of that customer. Note that, in this case, the actual service time of a customer is *endogenized* by the system dynamics, and I therefore refer to the latter model as the *endogenous-dependence* model. See §3.3 below for the specific modeling assumptions.

Note that the stochastic queueing dynamics of a system under either form of dependency (exogenous or endogenous) are intractable, because the number-in-system process is non-Markovian, even if the arrival process is Poisson and the conditional service time (conditioned on the delay) and patience time are both exponentially distributed. I therefore consider fluid approximations for the stochastic system, namely, a deterministic dynamic system that approximates the mean behavior of the corresponding stochastic system. I further note that, one can relatively easily observe a correlation between the service and waiting times of *served* customers, both the exogenous and endogenous model I study in this chapter can be used to explain such correlation observed from *censored* data. In particular, the service times and waiting times are observed for customers who did not abandon, but their actual patience times are *right censored* by the waiting times. Moreover, the patience times of *abandoned* customers can be observed, but their service times are not. Fortunately, I can capture both types of dependencies simultaneously via a *unified fluid model*, facilitating the study and comparisons between the two dependence

structures. The relation between the two dependencies in stochastic queueing systems is further discussed in Chapter 4, where a unified model for the stochastic systems is developed to capture the two dependencies.

The unified fluid model I propose is developed for non-stationary systems, namely, for systems with time-varying arrival rates. However, for the qualitative study of the fluid model, I focus on the important case of piecewise-constant arrival rate functions. Indeed, control and staffing decisions in practice are often made by dividing the day into time intervals over which the arrival rates are nearly constant, and then employing steady-state analysis over those time intervals; e.g., see Gans et al. (2003, §3.2). Such analysis relies on having the system converge to stationarity, rendering the study of transient behavior imperative. In particular, it is important to determine whether the system under consideration possesses a unique stationary behavior when the arrival rate is fixed, and whether it necessarily converges to the equilibrium in such a case. As my fluid model demonstrates, the answer to both questions is not always affirmative for the dependence models.

In particular, I characterize a sufficient condition for the fluid model to possess a unique equilibrium (stationary point), which always holds in the exogenous-dependence case, but may not hold in the endogenous-dependence case in which the service times depend on the delay. I provide examples for fluid models to have two equilibria, including both underload and overload equilibrium points. In the latter case, the system may be stuck in an overload equilibrium, even though it has sufficient potential service capacity to serve all the arrivals, a phenomenon referred to as *congestion collapse* in the queueing literature; see Perry and Whitt (2015, 2016). Further, I use numerical examples for

the transient fluid models to demonstrate that the rate of convergence to equilibrium is highly affected by the dependence structure. Therefore, employing the "standard model", in which the service times are assumed to be independent of the patience times and of the state of the system, when one of these dependencies is present, may be harmful.

Contribution. To summarize, my contribution is threefold.

(I) I develop a unified fluid model to approximate the stochastic dynamics of systems under both types of dependence structures simultaneously. I prove that, under minor regularity conditions, there exists a unique solution to the fluid equations, for any given initial condition. I propose an efficient algorithm to numerically solve the fluid-model equations.

(II) I provide a sufficient condition for an ordering of the fluid trajectories. In particular, if two systems only differ by the dependence structure (i.e., they have the same patience-time distribution, number of agents and arrival process), and both are initialized at the same initial condition, then the fluid model of one system will dominate the other all the time. In turn, such a result suggests the ordering of the equilibria (assuming the convergence to stationarity of the fluid model holds when the arrival rates are fixed). It may also suggest that the time to stationarity is a function of the specific joint distribution of service and patience times in the exogenous model, and of the conditional service time distribution in the endogenous model. Further, numerical examples show that the fluid model may oscillate towards a stationary point, implying that the time to stationarity may be relatively long.

(III) I characterize a sufficient condition for the existence of a unique stationary point for the fluid model with fixed arrival rate, which always holds for the exogenous model, and provide examples of bi-stable fluid models (for endogenous dependence), namely, fluid models possessing two stationary points. Further, I demonstrate that a bi-stable fluid model can have both an underload and an overload equilibrium. In this latter case, simulation experiments show that stochastic fluctuations may push a *nominally* underloaded system into an overload equilibrium, causing the system to experience congestion collapse for long time periods.

3.2. Related Literature

The exogenous-dependence model has been studied in Chapter 2 and in Bassamboo and Randhawa (2015). Chapter 2 analyzes the impact of such dependence on various performance measures and staffing decisions. The results show that even moderate dependence has significant impacts on system performance, so ignoring the dependence is harmful. Bassamboo and Randhawa (2015) characterize the optimal scheduling policies for a service system serving a homogeneous class of customers with dependent service and patience times. Both Chapter 2 and Bassamboo and Randhawa (2015) focus on the steady-state analysis of these queueing systems. This chapter, on the other hand, studies the transient behavior of these systems. My results demonstrate the uniqueness of the equilibrium point for the exogenous-dependence fluid model with fixed arrival rate and lend support to the steady-state analysis in Chapter 2 and Bassamboo and Randhawa (2015).

Related to the endogenous-dependence model in this chapter, Chan et al. (2016) and Vries et al. (2017) empirically report a dependence between the service times of served customers and their delay times in queue in different service contexts. Chan et al. (2016) show that patients' long waiting times for admissions to ICUs lead to increased hospitalization times in ICUs. Vries et al. (2017) (personal communication) find an opposite effect of delays in queue by analyzing data from a popular restaurant chain in India. The authors show that the time that customers spend dining is negatively dependent on the time they spend waiting for a table.

A growing literature on service operations studies the behavioral aspect of service agents who may exhibit speed-up or slow-down effect in response to the system's state. For example, Staats and Gino (2012) and Tan and Netessine (2014) find evidence of the agents' speed-up effect in reaction to a high system load in bank loan applications and restaurants, respectively. In a series of papers, Kc and Terwiesch (2009, 2012), Kuntz et al. (2014), Chan et al. (2014) and Batt and Terwiesch (2016) also report the speed-up effect in various healthcare service settings. On the other hand, the slow-down behavior of servers in a congested environment is also observed in Dietz (2011) and its implication on system performance is investigated in Dong et al. (2015). Recently, Armony et al. (2015) and Batt and Terwiesch (2016) show that service agents tend to slow down when system load remains high over an extended period, a phenomenon referred to as *overwork* in Kc and Terwiesch (2009). Delasay et al. (2016) design algorithms to analytically compute the performance measures when service agents experiencing overwork tend to slow down.

In terms of the throughput (number of service completions per unit time), a similar speed-up or slow-down effect also appears as a feature in my model. The service times

of customers in service may be longer or shorter than the average depending on their actual delays in queue, which in turn, depends on the congestion level in the system. However, it is significant that the driver of the relevant speed-up or slow-down effects in my model stems from customers rather than service agents. My model assumes that a customer's service time only depends on his patience or waiting time. In particular, the service time of a customer does not change once that customer is admitted into service. By contrast, when service agents may instantly adjust their service rates, the service time of a customer in service changes instantly while he is being served; e.g., see Dong et al. (2015). My modeling framework is similar to Chan et al. (2016). Motivated by their empirical evidence, Chan et al. (2016) analyze an $M/M(f)/n$ queueing model (with no abandonment) in which a customer's service time is exponentially distributed with mean which increases with the delay he experiences according to a given inflation function f (the notation $M(f)$ for the service time). Upper bounds for the steady-state workload in systems are developed, and are shown to be fairly accurate for small systems or systems with low utilization. I, on the other hand, consider the transient dynamics of large systems.

3.3. Model

I start by introducing two stochastic queueing systems with an exogenous and endogenous dependence, respectively. I then develop a deterministic fluid model to approximate the associated fluid-scaled processes derived in the stochastic models.

3.3.1. Two Stochastic Systems

I consider a multi-server queueing system with s statistically identical agents. Customers arrive to the system with time-varying arrival rate $s\lambda(t), t \geq 0$. New arrivals enter service immediately if there is an available agent and are delayed in queue if all agents are busy. Each customer has a finite patience for waiting to be served, and will abandon the queue if his waiting time exceeds that patience. A key feature of the underlying queueing model is that a customer in service requires a service time that depends on his patience time or waiting time in queue.

More specifically, let S_i and T_i be customer i's service and patience times, respectively. I assume that customers' patience times $\{T_i : i \geq 1\}$ are independent of all other random variables in the model and are independently and identically distributed (I.I.D.) with cumulative distribution function (cdf) F_T and probability density function (pdf) f_T. Let $F_T^c := 1 - F_T$ be the complementary cdf (ccdf) of the patience time. Let Z_i be the *offered wait* of customer i, which represents the virtual waiting time of that customer if he had infinite patience. It follows immediately that customer i's actual waiting time $W_i = \min\{T_i, Z_i\}$. To capture the dependence between customers' service, patience and waiting times, I assume a parametric form of the cdf of customer i's *conditional service time* conditioned on his waiting time:

$$\Psi_z(x) := \mathbb{P}(S_i \leq x | T_i > W_i, W_i = z) = \mathbb{P}(S_i \leq x | T_i > Z_i, Z_i = z).$$

For expositional convenience, I treat the parameter z as a single argument and write $\Psi(z, x) \equiv \Psi_z(x)$. Define the complementary cdf $\bar{\Psi}(z, x) := 1 - \Psi(z, x)$. Assuming $\Psi(z, x)$

is differentiable in x for all $z \geq 0$, the pdf of the conditional service time exists and satisfies

$$\psi(z,x) := \frac{\partial \Psi(z,x)}{\partial x}. \tag{3.1}$$

Define the hazard rate of the conditional service time as follows:

$$h(z,x) := \begin{cases} \frac{\psi(z,x)}{\bar{\Psi}(z,x)} & \text{if } x \in \mathcal{S}(z), \\ 0 & \text{if } x \notin \mathcal{S}(z), \end{cases} \tag{3.2}$$

where $\mathcal{S}(z) := \{x : \bar{\Psi}(z,x) > 0\}$ is the support of the conditional pdf $\psi(z,\cdot)$.

Exogenous Dependence: In the exogenous-dependence system, I assume that the service and patience time of each arriving customer are dependent and are drawn from a common bivariate joint distribution; e.g., Bassamboo and Randhawa (2015) and Chapter 2. Specifically, I assume $\{(S_i, T_i) : i \geq 1\}$ are I.I.D. bivariate random variables, all having the same continuous joint density f, with marginal densities f_S and f_T. Furthermore, I assume customer i's service-and-patience (S_i, T_i) is independent of the system's state, in particular, of his offered wait Z_i.

To derive the distribution of the conditional service time, note that

$$\begin{aligned} \Psi(z,x) &\overset{(1)}{=} \mathbb{P}(S_i \leq s | T_i > Z_i, Z_i = z) \\ &= \mathbb{P}(S_i \leq s | T_i > z, Z_i = z) \\ &\overset{(2)}{=} \mathbb{P}(S_i \leq s | T_i > z), \end{aligned}$$

where (1) follows from the definition of Ψ and (2) follows from the fact that (S_i, T_i) is independent of Z_i. Since (S_i, T_i) has joint density f, the distributions of the conditional service time satisfy

$$\Psi(z,x) = \frac{\int_z^\infty \int_0^x f(x,y)dxdy}{\int_z^\infty \int_0^\infty f(x,y)dxdy} \quad \text{and} \quad \psi(z,x) = \frac{\int_z^\infty f(x,y)dy}{\int_z^\infty \int_0^\infty f(x,y)dxdy}. \tag{3.3}$$

Endogenous Dependence: In the endogenous-dependence system, I assume that each customer' service time changes in response to his waiting time in queue. To capture this feature, I assume that a customer's service time only depends on his offered wait. Formally,

$$\Psi(z,x) = \mathbb{P}(S_i \leq s | T_i > Z_i, Z_i = z) = \mathbb{P}(S_i \leq s | Z_i = z).$$

It then follows that the service time of a customer in service only depends on his actual waiting time.

3.3.2. A Unified Fluid Model

Since the dependence renders the exact analysis intractable, I develop a fluid model to approximate the stochastic systems. Even though I do not prove limit theorems in this chapter, it is helpful to think of the fluid model as a weak law of large numbers for a sequence of stochastic systems. More formally, consider a sequence of stochastic systems as described above indexed by the number of agents s. Assume that the arrival rate to system s is $\lambda_s(t) := s\lambda(t) + o(s)$ (where $o(s)$ denotes a function that increases slower than s, i.e., $o(s)/s \to 0$ as $s \to \infty$), but that the *primitive model data* $\mathcal{D} := (\lambda, f_T, \psi)$ is fixed along the sequence. Letting Q_s denote the process of queue length in the s system, I

conjecture that Q_s/s converges in distribution to a deterministic limit, and that a similar result holds for the service process. The rest of this section is devoted to deriving the corresponding deterministic limits.

Let $B(t, y)$ denote the amount of fluid in service at time t that have been in service for at most y time units. Let $Q(t, y)$ denote the amount of fluid in queue at time t that have been in queue for at most y time units. Let $B(t) := B(t, \infty)$ and $Q(t) := Q(t, \infty)$ be the amount of fluid in service and in queue at time t, respectively. Let $X(t) := Q(t) + B(t)$ be the total amount of fluid in system at time t. Let $w(t)$ be the head-of-line waiting time, namely, the elapsed waiting time of the fluid at the head of the queue at time t. Let $v(t)$ be the *offered wait*, i.e., the virtual waiting time of an infinitely patient fluid arriving at time t. Suppose $B(t, y)$ and $Q(t, y)$ are integrable with continuous densities b and q:

$$B(t, y) = \int_0^y b(t, x)dx, \; Q(t, y) = \int_0^y q(t, x)dx.$$

Then $b(t, x)$ and $q(t, x)$ represent the amount of fluid at time t that has been in service and in queue for exactly x time units, respectively.

As in Liu and Whitt (2011a), I analyze the fluid model by considering alternating time intervals over which the system is either *underloaded* (UL) or *overloaded* (OL). To state these concepts, let $\sigma(t)$ denote the total rate of service completions at time t. An interval starting at time 0 is OL if (i) $Q(0) > 0$ or (ii) $Q(0) = 0, B(0) = 1$ and $\lambda(0) > \sigma(0)$. The OL interval ends at the termination time

$$T_{OL} = \inf\{t \geq 0 | Q(t) = 0 \text{ and } \lambda(t) < \sigma(t)\}.$$

An interval starting at time 0 is UL if (i) $B(0) < 1$ or (ii) $Q(0) = 0, B(0) = 1$ and $\lambda(0) \leq \sigma(0)$. The UL interval ends at the termination time

$$T_{UL} = \inf\{t \geq 0 | B(t) = 1 \text{ and } \lambda(t) > \sigma(t)\}.$$

When the system is in an UL interval, the new arrivals are served immediately upon arrival and experience no delay in queue. The conditional service-time distribution is the same for all the fluid that arrive during this interval. If the initial fluid in service is not delayed as well, then the fluid model in this UL interval is essentially the same as a model with fluid having a homogeneous service-time distribution. Hence, the analysis of the independent fluid model in UL intervals considered in Liu and Whitt (2012) carries over to this case.

In an OL interval, the waiting time w is strictly positive. Fluid in service with uncompleted service remains in service, leading to the evolution equation of fluid in service:

$$b(t+u, x+u) = b(t,x)\frac{\bar{\Psi}(w(t-x), x+u)}{\bar{\Psi}(w(t-x), x)} \text{ for } 0 < x < t. \tag{3.4}$$

To understand (3.4), note that the fluid density $b(t,x)$ enters service at time $t-x$, hence its waiting time is $w(t-x)$. The total service rate of the system at time t is

$$\sigma(t) = \int_0^t b(t,x)h(w(t-x), x)dx + c(t), \tag{3.5}$$

where $c(t)$ is the total service rate of the initial fluid at time t, which I define as follows. To state it, let $w_B(x)$ be the waiting time of $b(0,x)$, the initial fluid density that has been in service for time x. Define $\mathcal{S}_B(0) := \{x : b(0,x) > 0\}$ to be the set of elapsed service

times of nonzero fluid density in service at time 0. By the definition of $\mathcal{S}_B(0)$, it follows that that $\bar{\Psi}(w_B(x), x) > 0$ for all $x \in \mathcal{S}_B(0)$. Then the total service rate of the initial fluid that remains in service at time t can be computed as

$$c(t) := \int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x), t+x)}{\bar{\Psi}(w_B(x), x)} dx.$$

Fluid in queue that does not abandon and does not move into service, remains in queue, leading to the evolution equation of fluid in queue:

$$q(t+u, x+u) = q(t,x)\frac{F_T^c(x+u)}{F_T^c(x)}. \tag{3.6}$$

The total abandonment rate of fluid in queue is then

$$\alpha(t) = \int_0^\infty q(t,x)h_T(x)dx, \tag{3.7}$$

where $h_T(\cdot)$ is the hazard rate of patience time.

To fully characterize the fluid model, I specify boundary conditions of $b(t)$ and $q(t)$:

$$\begin{aligned} &q(t,0) = \lambda(t) \quad \text{if } Q(t) > 0, \\ &b(t,0) = \lambda(t) \quad \text{if } B(t) < 1, \\ &b(t,0) = \sigma(t) \wedge \lambda(t),\ q(t,0) = \lambda(t) - (\sigma(t) \wedge \lambda(t)) \quad \text{if } B(t) = 1, Q(t) = 0, \end{aligned} \tag{3.8}$$

where $x \wedge y = \min\{x, y\}$.

The waiting time function w is required to compute the conditional service-time distribution in (3.4) and (3.5). To determine w, I relate it to the offered wait v in the following

lemma. Recall the offered wait $v(t)$ is the virtual waiting time of infinitely patient fluid that arrives at time t.

Lemma 3.1. *In an OL interval, the following holds:*

$$w(t + v(t)) = v(t). \tag{3.9}$$

The proof of Lemma 3.1 follows from Liu and Whitt (2012, Proposition 5). I next characterize v so that w can be computed via (3.9). Consider a virtual arrival of fluid with infinite patience at time t observing a queue length $Q(t)$. Turn off the arrivals after time t and define $k(u)$ to be the clearing queue process at time $t + u$ with $u \geq 0$, which starts at $k(0) = Q(t)$. Then

$$v(t) := \inf_{u\geq 0} \left\{ k(u) = 0 \middle| k(0) = Q(t), k'(u) = -\beta(t + u) - \sigma(t + u) \right\}, \tag{3.10}$$

where

$$\beta(t + u) = \int_u^{w(t+u)} q(t + u, x)h_T(x)dx. \tag{3.11}$$

is the total abandonment rate of the clearing queue at time $t + u$.

Equivalently, one can characterize v as follows (see Liu and Whitt (2012)):

$$v(t) = \inf_{u\geq 0}\{E(t + u) - E(t) + A_t(u) \geq Q(t)\},$$

where

$$E(t) = \int_0^t \sigma(u)du,\ A_t(u) = \int_t^{t+u} \alpha_t(x)dx \text{ and } \alpha_t(u) = \int_{u-t}^{\infty} q(u, x)h_T(x)dx,\ u \geq t.$$

3.4. Analysis

In this section, I show that the fluid model with time-varying arrivals admits a unique solution under mild regularity conditions. I study the stationary behavior of the fluid model when the arrival process is stationary, which reveals a fundamental difference between the exogenous and endogenous dependencies.

3.4.1. Solution to the Fluid Model

For the fluid model to admit a unique solution, I make several technical assumption on the smoothness of the model primitives. First, I assume the arrival rate function $\lambda(\cdot)$ is continuous and strictly positive. Second, I assume the initial conditions are continuous. The initial fluid content in service $B(0,\cdot)$ and in queue $Q(0,\cdot)$ are twice differentiable functions with derivatives $b(0,\cdot), q(0,\cdot)$. Furthermore, if $Q(0) > 0$, then $b(0,0) = q(0, w(0))$ and $q(0,0) = \lambda(0)$; if $Q(0) = 0$, then $b(0,0) = \lambda(0)$. This assumption holds true if the system starts running in the distant past. Third, I assume the conditional service-time distribution is smooth: $f_T(\cdot)$ is continuous and $\psi(z,\cdot)$ is continuous for all $z \geq 0$. Further, it holds that

$$\sup_{0\leq x\leq x_0, 0\leq z\leq z_0} \psi(z,x) < \infty \quad \text{and} \quad \sup_{0\leq x\leq x_0, 0\leq z\leq z_0} \Big|\frac{\partial \psi(z,x)}{\partial z}\Big| < \infty \text{ for all } x_0, z_0 > 0.$$

Next, I impose a bound on the service rate followed from the initial service content. I assume the total service rate of the initial fluid that remains in service at time t is uniformly bounded for all $t > 0$. Specifically, the tail of initial density in service is

bounded relative to the initial conditional service-time distribution:

$$C(t) := \sup_{0\le u\le t} c(u) = \sup_{0\le u\le t} \int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x), t+x)}{\bar{\Psi}(w_B(x), x)} dx < \infty \text{ for all } t > 0.$$

I give some important lemmas, which are useful in proving the existence and uniqueness of a solution to the fluid model. I start by rewriting (3.5) in an OL interval.

Lemma 3.2. *In an OL interval starting at time 0,*

$$\sigma(t) = \int_0^t \sigma(t-u)\psi(w(t-x), x)dx + c(t). \tag{3.12}$$

In an OL interval, w is strictly positive. I can compute the derivative of w by assuming that the initial waiting time is continuous and the initial queue density is strictly positive if there is a queue at time 0. Formally, $w_B(x)$ is continuous for all x and $q_{\inf}(0) := \inf_{0\le u\le w(0)}\{q(0,u)\} > 0$ if $Q(0) > 0$.

Lemma 3.3. *In an OL interval starting at time 0, the following hold:*

$$w'(t+) = 1 - \frac{\sigma(t-)}{\tilde{q}(t, w(t+))}, \tag{3.13}$$

$$w(t) = \int_0^t \left[1 - \frac{\sigma(x)}{\tilde{q}(x, w(x))}\right] dx + w(0), \tag{3.14}$$

where

$$\tilde{q}(t,x) = \lambda(t-x)F_T^c(x)\mathbf{1}_{[x\le t]} + q(0, x-t)\frac{F_T^c(x)}{F_T^c(x-t)}\mathbf{1}_{[t<x]}. \tag{3.15}$$

The proof of Lemma 3.3 is similar to Liu and Whitt (2012, Proposition 3).

Observe that the fluid model is fully characterized by the primitive model data $\mathcal{D} := (\lambda, f_T, \psi)$. The intricacy of the fluid model lies in the nested w and σ in (3.12) and (3.14) in an OL interval. The next proposition claims that there is a unique solution $(b, q, \sigma, \alpha, Q, B, W, V)$ that solves the fluid equations.

Proposition 3.1. *The fluid model* (3.4)-(3.14) *with model data* $\mathcal{D} = \{\lambda, f_T, \psi\}$ *has a unique solution* $(b, q, \sigma, \alpha, Q, B, w, v)$.

I am interested in comparing trajectories of fluid models with different model data. Indeed, with all else being equal in the model data, the trajectories of two fluid models can be ordered if the hazard rates of the conditional service times in the two models are ordered. To state the result, consider two fluid models with common arrival rate, patience-time distribution and initial conditions. I use subscript i to specify the conditional hazard rate and performance functions associated with model i, $i = 1, 2$. Define the order of functions to be pointwise order for all arguments and the order of vectors to be pointwise order for all coordinates.

Proposition 3.2 (trajectory comparison)**.** *Consider two fluid models with common arrival rate* $\lambda(\cdot)$ *and patience-time distribution* $f_T(\cdot)$. *The initial conditions* $(b(0), q(0), w_B)$ *are also the same for both models. If the hazard rates of the conditional service times can be ordered as follows,*

$$\sup_{z,x\geq 0} h_1(z, x) \leq \inf_{z,x\geq 0} h_2(z, x), \tag{3.16}$$

then $(w_1, v_1, q_1, Q_1, B_1, X_1) \geq (w_2, v_2, q_2, Q_2, B_2, X_2)$.

It is important that the ordering (3.16) is satisfied for an important class of model data. In particular, when the dependence is exogenous and the bivariate random variables representing the service and patience times are generated by Gaussian copulas with the marginal of service time being exponential, the ordering (3.16) holds when two models with and without a dependence are compared. See §3.4.3.1 for more details.

3.4.2. Stationary Arrivals

I identify the stationary point of the fluid model when the arrival process is stationary, namely, $\lambda(t) \equiv \lambda$ for all $t \geq 0$. To gain qualitative results, I impose a condition on the workload associated with fluid in queue to guarantee the existence and uniqueness of a stationary point. Define the average conditional service rate, conditioned on the waiting time z,

$$\mu(z) := \frac{1}{\mathbb{E}[S_i | T_i > Z_i, Z_i = z]} = \frac{1}{\int_0^\infty x\psi(z,x)dx}. \tag{3.17}$$

Define the *work evolution function*

$$\phi(z) := F_T^c(z)/\mu(z), \tag{3.18}$$

which represents the workload required by unit fluid that remains in the system after waiting for w time units in queue.

Condition 3.1. *$\phi(z)$ is strictly decreasing in z.*

Condition 3.1 clearly holds when $\mu(\cdot)$ is nondecreasing (implying a negative dependence between service and waiting times). In particular, when service and waiting times

are independent, $\mu(\cdot)$ is constant so that Condition 3.1 also holds. When $\mu(\cdot)$ is decreasing (implying a positive dependence between service and waiting times), Condition 3.1 does not hold in general. However, I show in §3.4.3.1 that Condition 3.1 is always satisfied when the dependence is exogenous even if $\mu(\cdot)$ may be decreasing.

Proposition 3.3. *Let $\lambda(t) \equiv \lambda$ for all $t \geq 0$. Under Condition 3.1, for any $\lambda > 0$, the fluid model with primitive model data $\mathcal{D} = (\lambda, f_T, \psi)$ has a unique stationary point characterized by the vector $(b, q, \sigma, \alpha, Q, B, w, v)$ depending on the relative value of λ and $\mu(0)$.*

(a) Underloaded and critically loaded systems:

If $\lambda \leq \mu(0)$, then

$$B = \lambda/\mu(0),\ \sigma = \lambda,\ w = v = \alpha = 0,\ Q = q = 0 \quad \textit{and} \quad b(x) = \lambda\bar{\Psi}(0, x).$$

(b) Overloaded systems:

If $\lambda > \mu(0)$, then

$$w = v = w^*,\ \sigma = \mu(w^*),\ \alpha = \lambda - \sigma,\ B = 1,$$

$$b(x) = \sigma\bar{\Psi}(w^*, x),\ x \geq 0,$$

$$q(x) = \lambda F_T^c(x),\ 0 \leq x \leq w^* \quad \textit{and} \quad q(x) = 0,\ x > w^*,$$

where w^* is the solution to

$$\lambda\phi(w) = 1. \tag{3.19}$$

The total queue content is

$$Q = \int_0^{w^*} q(x)dx = \lambda \int_0^{w^*} F_T^c(x)dx. \tag{3.20}$$

I remark that Condition 3.1 is a sufficient and necessary condition to guarantee a unique equilibrium for the fluid model for any *arbitrary* stationary arrival rate λ. However, for a system with *given* arrival rate λ, Condition 3.1 is a sufficient condition for the existence of a unique stationary point. Other sufficient conditions that can lead to a unique equilibrium for a system with fixed stationary arrival rate are stated in the following proposition. When these sufficient conditions are violated, multiple stable equilibria may exist for the fluid model, as I demonstrate using simulations in §3.5.2.1.

Proposition 3.4. *Suppose* $\lim_{z\to\infty} \phi(z) = 0$.

1. *For an overloaded system with stationary arrival rate* $\lambda > \mu(0)$*, if either of the following holds, then there exists a unique stationary point for the fluid model:*
   (i) $\phi(z)$ *is unimodal;*
   (ii) $1/\phi(z) = \mu(z)/F_T^c(z)$ *is convex in* z.
2. *For an underloaded or critically loaded system with stationary arrival rate* $\lambda \leq \mu(0)$*, the stationary point for the fluid model is unique if and only* $\phi(z) < 1/\lambda$ *for all* $z \geq 0$.

3.4.3. Implications of Two Dependencies

In this section, I discuss the implications of the two dependence structures in the context of the fluid model. I consider the exogenous dependence in §3.4.3.1 and employ a bivariate dependence concept to compare trajectories for systems differing from one another only by

the dependence. I consider the endogenous dependence in §3.4.3.2. By further assuming that the conditional service time is exponentially distributed, I prove the convergence to stationarity under certain initial conditions.

3.4.3.1. Exogenous Dependence: Dependent Service and Patience Time. In the exogenous-dependence model, each arriving customer is endowed with an exogenous service and patience time which are dependent. The following lemma shows that Condition 3.1 holds when the dependence is exogenous. The proof of the lemma can be found in Chapter 2.

Lemma 3.4. *If S and T have a joint density f, then Condition 3.1 holds.*

Lemma 3.4 highlights a fundamental difference between exogenous and endogenous dependencies. When the dependence is exogenous, Lemma 3.4 shows that Condition 3.1 necessarily holds. Therefore, when the arrival process is stationary, Proposition 3.3 and Lemma 3.4 imply that the fluid model under an exogenous dependence has a unique stationary point, as characterized in Chapter 2 and Bassamboo and Randhawa (2015). However, when the dependence is endogenous, Condition 3.1 may be violated and multiple stable equilibria may exist. I leave the discussion on the identification of the two dependencies in a *censored* dataset to Section 4.1.

Next, I study the impact of the exogenous dependence on the evolution of performance functions. I am interested in comparing trajectories of systems differing from one another only by the dependence between the service and patience times. To this end, I fix the marginal distributions of the service and patience times and only vary the dependence of the two. To invoke Proposition 3.2, one needs an ordering of the hazard rate of conditional

service time. To gain qualitative results, I consider the following bivariate dependence concept (Block et al. (1985)). Two random variables X_1 and X_2 are said to be Positive Dependent through Stochastic ordering (PDS) if $\mathbb{P}(X_1 > x_1 | X_2 = x_2)$ is increasing in x_2 for all x_1. Similarly, X_1 and X_2 are said to be Negative Dependent through Stochastic ordering (NDS) if $\mathbb{P}(X_1 > x_1 | X_2 = x_2)$ is decreasing in x_2 for all x_1. When the service time has an *marginal* exponential distribution, the following proposition characterizes the qualitative structure of the conditional hazard rate using the PDS or NDS concept.

Proposition 3.5. *Consider bivariate random variables (S, T). Suppose S is exponentially distributed with rate μ and T has a general distribution.*

(1) If (S, T) is PDS, then $\sup_{z,x\geq 0} h(z, x) \leq \mu$.

(2) If (S, T) is NDS, then $\inf_{z,x\geq 0} h(z, x) \geq \mu$.

It is important that the set of PDS or NDS bivariate distributions with given marginals is non-empty. In particular, if bivariate random variables (S, T) are generated by Gaussian copulas (e.g., Wu et al. (2017, Appendix A)), then (S, T) must be either PDS or NDS depending on the correlation coefficient of S and T.

Lemma 3.5. *Consider bivariate random variables (S, T) generated by Gaussian copulas. Let r be the correlation coefficient of S and T. Then (S, T) is PDS if $r > 0$ and NDS if $r < 0$.*

The proof of Lemma 3.5 can be found in Chapter 2. Combining Proposition 3.5 and Lemma 3.5, I immediately have the following corollary.

Corollary 3.1. *Consider two fluid models with common arrival rate λ and initial condition. In model i ($i = 1, 2$), the service time S_i is exponentially distributed and (S_i, T_i) are generated by Gaussian copulas. Let r_i be the correlation coefficient of (S_i, T_i). If $r_1 \geq 0 \geq r_2$, then $(w_1, v_1, q_1, Q_1, B_1, X_1) \geq (w_2, v_2, q_2, Q_2, B_2, X_2)$.*

The ordering of the trajectories for different fluid models also suggests the ordering of the equilibria given that the convergence to stationarity holds. This argument provides an alternative to proving the ordering of equilibria for different exogenous-dependence models in Chapter 2.

3.4.3.2. Endogenous Dependence: Exponential Conditional Service Time. Recall that in the endogenous-dependence model, a customer's service time only depends on his offered wait. To gain qualitative results, I further assume that a customer with offered wait z requires a service time which is exponentially distributed with rate $\mu(z)$. In general, for an arbitrary rate function $\mu(\cdot)$, one cannot find a joint density of service and patience times in the exogenous-dependence model that induces $\mu(\cdot)$ via (4.1). In particular, Lemma 3.4 shows that if $\phi(z) = F_T^c(z)/\mu(z)$ is nondecreasing, then μ can never be induced by an exogenous-dependence model.

Proposition 3.1 claims the existence and uniqueness of a solution to the fluid equations. Proposition 3.3 identifies the unique stationary point for the fluid model with stationary arrivals when Condition 3.1 is valid. It is natural to conjecture that the transient fluid model converges to the unique stationary point. To prove this conjecture is nontrivial in the exogenous-dependence model mainly because a general joint density f of service and patience times does not provide good structures on the hazard rate of the conditional service time. However, in the endogenous-dependence model, the assumption that the

conditional service time has an exponential distribution allows us to prove this conjecture under certain initial conditions. Under the exponential assumption, the conditional service-time distribution has relatively simple forms:

$$\Psi(z,x) = 1 - e^{-\mu(z)x},\ \psi(z,x) = \mu(z)e^{-\mu(z)x} \text{ and } h(z,x) = \mu(z).$$

I further assume $\mu(\cdot)$ is decreasing.

Assumption 3.1. *$\mu(z)$ is differentiable and decreasing in z.*

A decreasing conditional service rate $\mu(\cdot)$ implies a positive dependence between the service and waiting times, namely, a customer who waits longer in queue tends to require a longer service time. I remark that when $\mu(\cdot)$ is strictly decreasing, the exponential conditional service-time distribution cannot be induced by an exogenous-dependence model via (4.1) even if $\phi(z) = F_T^c(z)/\mu(z)$ is decreasing. In other words, no exogenous-dependence model gives the same system dynamics as those under the specific endogenous dependence I consider in this section.

Recall $w_B(x)$ is the waiting time of the initial density $b(0,x)$. For notational convenience, I define $w(-x) := w_B(x)$ for $x \in \mathcal{S}_B(0)$.

Proposition 3.6. *Suppose $\lambda(t) \equiv \lambda > \mu(0)$ for all $t \geq 0$. Let Condition 3.1 and Assumption 3.1 hold true. Let $w^*(\lambda)$ be the equilibrium waiting time solving* (3.19) *when the arrival rate is λ. If one of the following holds,*

(1) The system is OL at time 0 *and*

$$\max_{-\max \mathcal{S}_B(0) < t \leq v(0)} \{w(t)\} = v(0) \leq w^*(\lambda),\ w'(v(0)+) \geq 0.$$

(2) The system is OL at time 0 *and*

$$\min_{-\max \mathcal{S}_B(0)<t\leq v(0)} \{w(t)\} = v(0) \geq w^*(\lambda),\ w'(v(0)+) \leq 0.$$

(3) The system is UL at time 0 *and* $w_B(x) = 0$ *for all* $x \in \mathcal{S}_B(0)$.

then it holds that

$$\lim_{t\to\infty} w(t) = w^*(\lambda).$$

If the fluid model starts empty, then condition (iii) in Proposition 3.6 holds. Hence, a fluid model initialized empty necessarily converges to the stationary point given that it is unique. Another immediate result of Proposition 3.6 is that, when the arrival rate λ is piecewise constant and the system remains overloaded all the time, one can observe transitions between steady states associated with different arrival rates.

Corollary 3.2. *Let Condition 3.1 and Assumption 3.1 hold true. Suppose the fluid model is at the stationary point associated with arrival rate* $\lambda_1 > \mu(0)$ *at time* 0. *If* $\lambda(t) = \lambda_2 > \mu(0)$ *for* $t \geq 0$, *then* $\lim_{t\to\infty} w(t) = w^*(\lambda_2)$.

3.5. Numerical Study

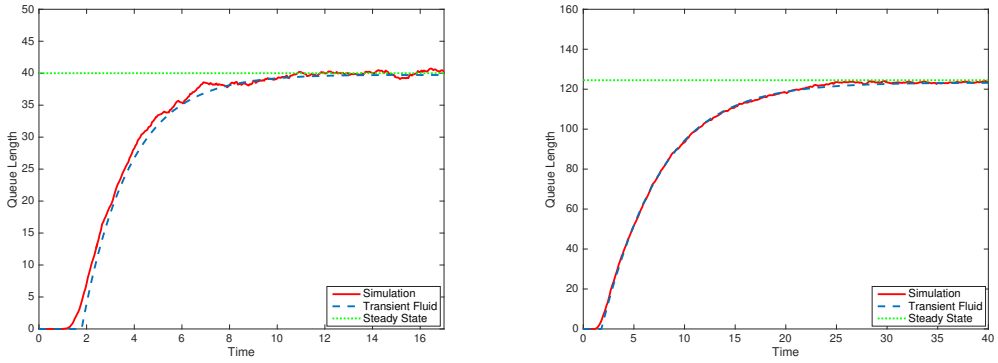
I conduct simulation experiments to validate the accuracy of the fluid model in approximating the stochastic queueing systems. I first solve the fluid equations using a discretization algorithm; details are relegated to Appendix B.1. I then simulate systems with an exogenous and endogenous dependencies, respectively, and compare the simulation results to the fluid estimates.

3.5.1. Exogenous Dependence

I first simulate systems with an exogenous dependence. Specifically, I simulate three queueing systems with different dependencies between the service and patience times. All systems start empty and have 100 service agents and Poisson arrivals of customers with rate 120. The service and patience times of arriving customers are exponentially distributed with rates 1 and 1/2, respectively. I use Gaussian copulas to generate the bivariate random variable representing the service and patience times. In the current study, the correlation coefficient between the service and patience times is 1 (-0.64) for the system with a positive (negative) dependence, which is the maximal (minimal) attainable correlation coefficient for two exponential random variables. See the discussion on bivariate generation using Gaussian copula in Chapter 2.

For each simulated system, I take averages over 500 independent runs and use the queue length function $Q(t)$ as the performance metric to demonstrate the convergence. I compare the queue length estimates computed from the fluid model with those observed by simulations. Results are plotted in Figure 3.1 (systems with no dependence and a positive dependence) and the left panel of Figure 3.2 (system with a negative dependence).

I find that the fluid model is effectively approximating *overloaded* systems with a positive dependence or no dependence between the service and patience times. However, I observe substantial gaps between the fluid estimates and simulation results when the dependence is negative. This is because the nominal traffic intensity $\rho := \lambda/\mu(0) = 1.2$ in the current study does not reflect the actual load of the system when a dependence



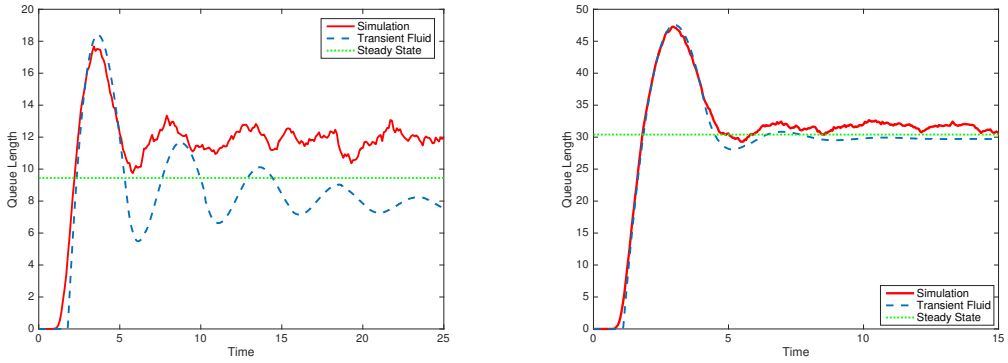
Independence Positive dependence

Figure 3.1. Queue length process of systems under different dependencies between service and patience times. Each system has $s = 100$ agents, Poisson arrivals with rate $\lambda_s = 120$, service-time distribution exp(1), patience-time distribution exp(1/2). The joint distributions of service and patience times are generated via Guassian copulas.

exists. Instead, the *effective* traffic intensity in steady state can be defined via

$$\rho_{\text{eff}} := \frac{\lambda_s}{s\mu(w^*)} = \frac{\lambda}{\mu(w^*)},$$

where w^* is the unique equilibrium waiting time solving (3.19). In the expression above, $\mu(w^*)$ is the *effective service rate* in steady state, corresponding to the average service rate of customers in service. For the system with a negative dependence, I can compute $\rho_{\text{eff}} \approx 1.04$. The stochastic system in steady-state can be considered to be critically loaded (see Garnett et al. (2002)) and I expect the effect of stochastic fluctuations to be significant. In the system that generates the right panel of Figure 3.2, I increase the arrival rate to 150 while fixing other system parameters. In this latter case, it can be computed that $\rho_{\text{eff}} \approx 1.11$. The stochastic system is effectively overloaded in steady state and the accuracy of the fluid estimates improves substantially.



$\lambda_s = 120$ $\lambda_s = 150$

Figure 3.2. Queue length process of systems under *negative* dependence between service and patience Time. Each system has $s = 100$ agents, Poisson arrivals with rate λ_s, service-time distribution exp(1), patience-time distribution exp(1/2). The joint distributions of service and patience times are generated via Guassian copulas.

Interestingly, even if the dependencies in all three system are exogenous, they may lead to very different patterns of system dynamics. When the dependence is positive and the system is initialized empty, the transient queue converges to the stationary queue monotonically. A similar pattern is observed in the independent model. However, the time to stationarity under a positive dependence is roughly twice as much as that in the independent model. (This observation is made clear in Figure 3.7 below.) Furthermore, with the number of servers and arrival rate being equal, the steady-state queue length under a positive dependence is roughly three times as much as that without a dependence. A detailed discussion on the impact of an exogenous dependence on various steady-state performance measures can be found in Chapter 2.

When the dependence is negative, I observe oscillations in the system dynamics, a pattern very different from those with a positive dependence and no dependence. The intuition behind the observed oscillations can be explained as follows. When the queue

length lands above the steady state, the fluid in queue has to wait longer than that in steady state to enter service. Those who get served require shorter service times due to the negative dependence, implying an increase in the total service rate. The queue is cleared at a higher rate until it drops below the steady state. A similar reasoning explains why the queue length is pulled up after it falls below the steady state. Finally, I remark that it takes less time to converge to stationarity under a negative dependence compared to under no dependence.

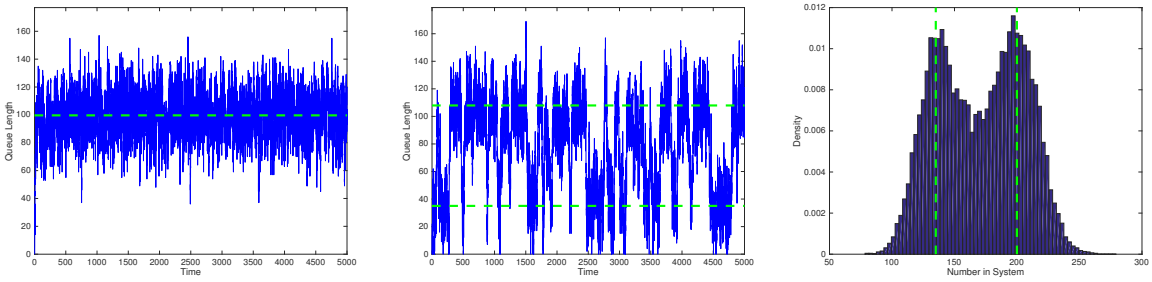
3.5.2. Endogenous Dependence

To simulate systems with an endogenous dependence, I assume the conditional service time is exponentially distributed. A decreasing service rate function $\mu(\cdot)$ implies a positive dependence while an increasing $\mu(\cdot)$ implies a negative dependence. I use simulations to demonstrate the potential bistability of equilibria when Condition 3.1 is violated. I also validate the accuracy of the fluid model when the equilibrium is unique.

3.5.2.1. Bistability of Equilibria. When the service rate function $\mu(\cdot)$ is increasing (implying a negative dependence), Condition 3.1 holds trivially. Recall that Condition 3.1 is a sufficient and necessary condition to guarantee a unique equilibrium for the fluid model for any arbitrary stationary arrival rate λ. For a given arrival rate, the unique equilibrium of the fluid model is prescribed in Proposition 3.3. However, Condition 3.1 may be violated under a positive dependence with a decreasing $\mu(\cdot)$. As such, there may exist multiple equilibria for the fluid model, which I demonstrate using simulations below. Similar results are observed in Dong et al. (2015).

The first two figures in Figure 3.3 give sample paths of the queue length process for two *nominally overloaded* systems with different service rate function $\mu(\cdot)$ satisfying $\lambda > \mu(0)$. I also plot the estimates of the stationary points computed from the fluid model (dashed horizontal line). Both systems have $s = 100$ agents and Poisson arrivals of customers with rate $\lambda = 115$. The patience time is exponentially distributed with rate $1/2$. Let $\mu(z) = 0.6+0.4\exp(-1.8\cdot z)$ in the first system, where one can show both of the conditions for the overloaded case in Proposition 3.4 hold so that the uniqueness of the equilibrium follows. However, in the second system where I set $\mu(z) = 0.6+0.4\exp(-1.8\cdot z^3)$, none of the sufficient conditions in Propositions 3.3 and 3.4 is satisfied and as a result, I observe two stationary points in the middle panel of Figure 3.3. The stochastic system stays in each equilibrium for a substantial amount of time until the stochastic fluctuation in the system triggers the transition from one equilibrium to the other. The right panel of Figure 3.3 plots the stationary distribution of the number in system process obtained from simulation results. The two equilibria computed by the fluid model (dashed vertical line) match the two peaks in the stationary density.

For *nominally underloaded* systems with $\lambda < \mu(0)$, bistability of equilibria may also be observed when the condition in Proposition 3.4 for the underloaded case is violated. Figure 3.4 plots sample paths of the number-in-system process for two nominally underloaded system with different service rate function $\mu(\cdot)$. Both systems have $s = 100$ agents and Poisson arrivals of customers with rate 90. The patience time of arriving customers is exponentially distributed with rate $1/2$. The first system with $\mu(z) = 0.6+0.4\exp(-1.8\cdot z)$ has one unique stationary point, similar to the overloaded case with the same system parameters except for the arrival rate. The second system with service rate function

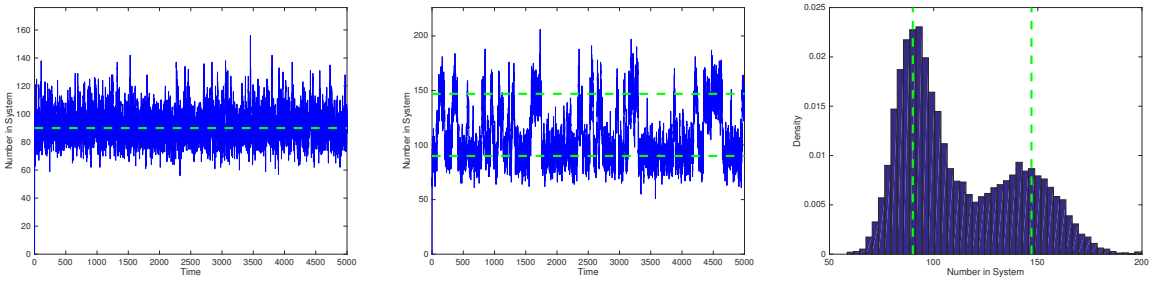


$\mu(z) = 0.6 + 0.4\exp(-1.8 \cdot z)$ $\mu(z) = 0.6 + 0.4\exp(-1.8 \cdot z^3)$ Probability Density

Figure 3.3. Sample paths of queue length process of *overloaded* systems under *decreasing* conditional service rate functions. Each System has $s = 100$ agents, Poisson arrivals with rate $\lambda_s = 115$, patience-time distribution $\exp(1/2)$.

$\mu(z) = 0.6 + 0.4\exp(-3 \cdot z)$ (the middle panel) has two stationary points, an underload one with no queue and an overload one with a nonempty queue. The stationary distribution of the number-in-system process is plotted in the right panel. Note that for the second system, the system may experience congestion collapse, namely, it may settle at an overload equilibrium with a nonempty queue on the fluid scale, even though it is nominally underloaded with $\lambda < \mu(0)$ and has sufficient potential service capacity to serve all the arrivals.

3.5.2.2. Convergence to Stationarity. I conduct simulation experiments to verify the accuracy of the fluid model given that the conditional service time is exponentially distributed. I simulate four overloaded systems, all having one unique equilibrium on the fluid scale. The number of service agents $s = 100$ is the same for all systems. I set the arrival rate $\lambda_s = 115$ for systems with a positive dependence (decreasing $\mu(\cdot)$) and $\lambda_s = 130$ for systems with a negative dependence (increasing $\mu(\cdot)$). The patience time is



$\mu(z) = 0.6 + 0.4\exp(-1.8 \cdot z)$ $\mu(z) = 0.6 + 0.4\exp(-3 \cdot z)$ Probability Density

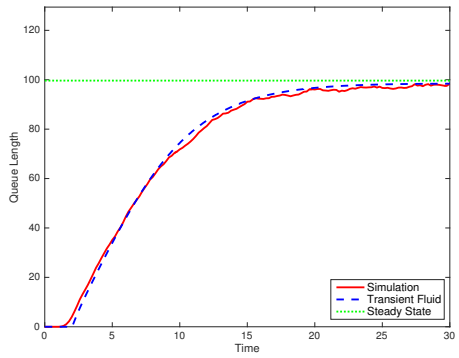
Figure 3.4. Sample paths of number in system process of *underloaded* systems under *decreasing* conditional service rate functions. Each system has $s = 100$ agents, Poisson arrivals with rate $\lambda_s = 90$, patience-time distribution $\exp(1/2)$.

exponentially distributed with rate 1/2. Results of the queue length process are plotted in Figures 3.5 and 3.6.

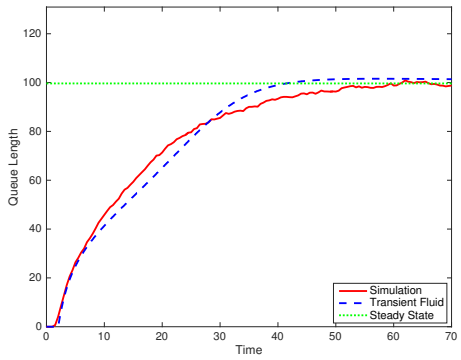
Figure 3.5 depicts the queue-length process for two systems with a decreasing $\mu(\cdot)$ and arrival rate $\lambda_s = 115$. Although Condition 3.1 holds for neither of the two systems, both of them have one unique equilibrium. (Recall Condition 3.1 is a sufficient condition for the uniqueness of equilibrium.) Note that the time to stationarity in the second system is very long.

Figure 3.6 plots the queue-length process for two systems with an increasing $\mu(\cdot)$. I let λ_S be 130 for the systems to be effectively overloaded with ρ_{eff} defined in §3.5.1 substantially greater than 1. Similar to the case with an exogenous negative dependence in §3.5.1, I observe oscillations in the first system where $\mu(z) = 0.6 + 0.4\exp(-1.8 \cdot z)$. However, in the second system where $\mu(z) = 0.6 + 0.4\exp(-1.8 \cdot z^2)$, the transient queue length converges monotonically to the stationary point. The different patterns of

system dynamics observed under different increasing service rate functions highlight a complicated interaction between the underlying dependence and system performance.

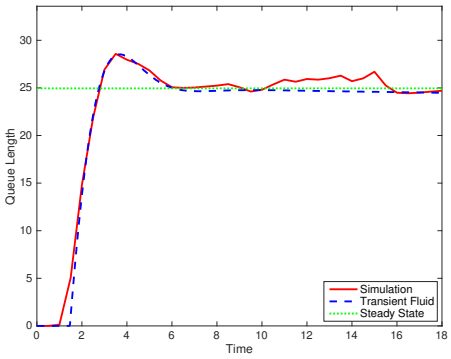


$\mu(z) = 0.6 + 0.4\exp(-1.8 \cdot z)$

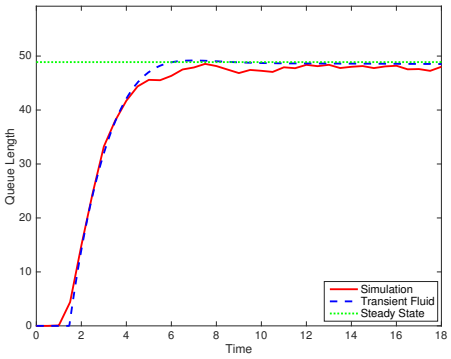


$\mu(z) = 0.6 + 0.4\exp(-1.8 \cdot z^2)$

Figure 3.5. Queue length process of *overloaded* systems under *decreasing* conditional service rate functions. Each system has $s = 100$ agents, Poisson arrivals with rate $\lambda_s = 115$, patience-time distribution $\exp(1/2)$.



$\mu(z) = 0.6 + 0.4\exp(1.8 \cdot z)$



$\mu(z) = 0.6 + 0.4\exp(1.8 \cdot z^3)$

Figure 3.6. Queue length process of *overloaded* systems under *increasing* conditional service rate functions. Each system has $s = 100$ agents, Poisson arrivals with rate $\lambda_s = 130$, patience-time distribution $\exp(1/2)$.

3.5.3. Trajectory Comparison

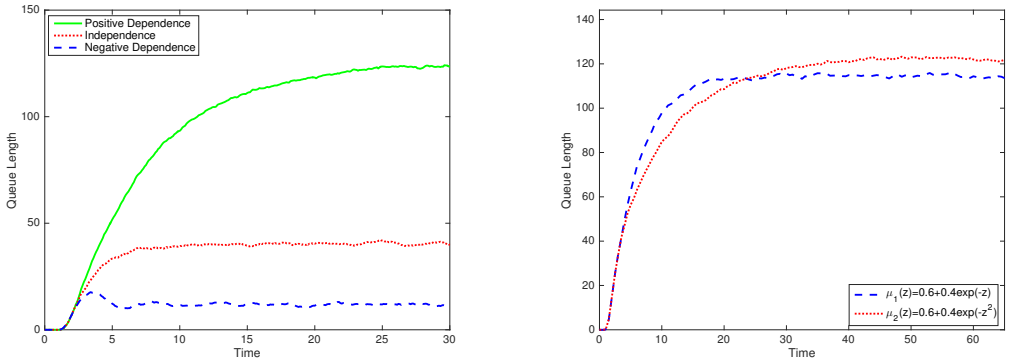
I use simulations to validate the trajectory comparison result in Proposition 3.2. I first consider exogenous dependencies which lead to the ordering of conditional hazard rates in (3.16), as discussed in Proposition 3.5. I then give a counterexample in the context of §3.4.3.2 to demonstrate the violation of the trajectory comparison result in Proposition 3.2 when the ordering in (3.16) fails. Results of the queue length process are plotted in Figure 3.7.

In the left panel of Figure 3.7, I plot the queue-length processes of three systems simulated in §3.5.1 together. Recall that all systems have $s = 100$ agents, Poisson arrivals with rate $\lambda_s = 120$ and service and patience times generated by Gaussian copulas with marginal distributions being exponential with rates 1 and 1/2, respectively. Corollary 3.1 implies the trajectory comparison result holds across systems with a positive, negative and no dependence, as is indeed observed in the left panel of Figure 3.7.

It can be readily seen that the time to stationarity is much longer under positive dependence than under no dependence. This observation suggests caution when one performs steady-state analysis over separate time intervals for a system with time-varying arrival rates. In particular, if the transient system does not converge to the equilibrium sufficiently fast (relative to the change in the arrival rate), then it is imperative to account for the entire transient dynamics in the staffing and control decisions. Restricting the analysis to the steady-state, in this case, is harmful.

In the right panel of Figure 3.7, I simulate two systems with an endogenous dependence but different service rate functions $\mu(\cdot)$. Both systems have $s = 100$ agents and Poisson arrivals with rate $\lambda_s = 130$. The patience-time distribution is exponential with rate 1/2.

The conditional service rate functions in the two systems are set $\mu_1(z) = 0.6+0.4\exp(-z)$ and $\mu_2(z) = 0.6 + 0.4\exp(-z^2)$, respectively. The ordering in (3.16) is violated in this case and the trajectory comparison result in Proposition 3.2 is invalid. Indeed, I observe a crossing of the two queue-length processes.



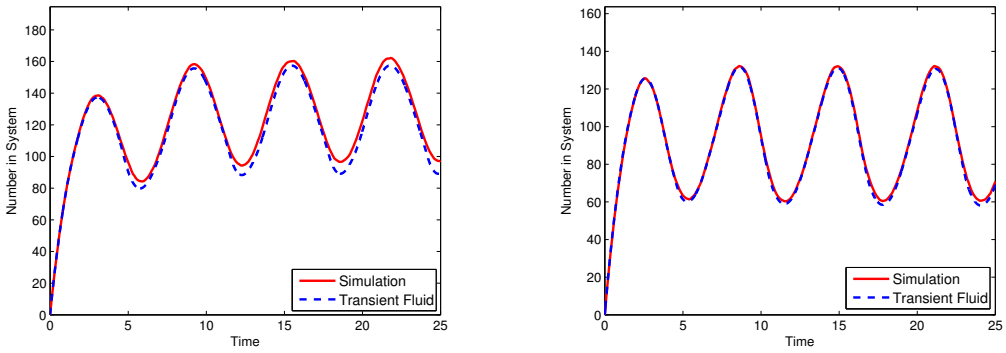
Exogenous dependence, $\lambda_s = 120$ Endogenous dependence, $\lambda_s = 130$

Figure 3.7. Trajetory comparison. Each system has $s = 100$ agents, Poisson arrivals with rate λ_s, patience-time distribution exp(1/2).

3.5.4. Time-varying Arrivals

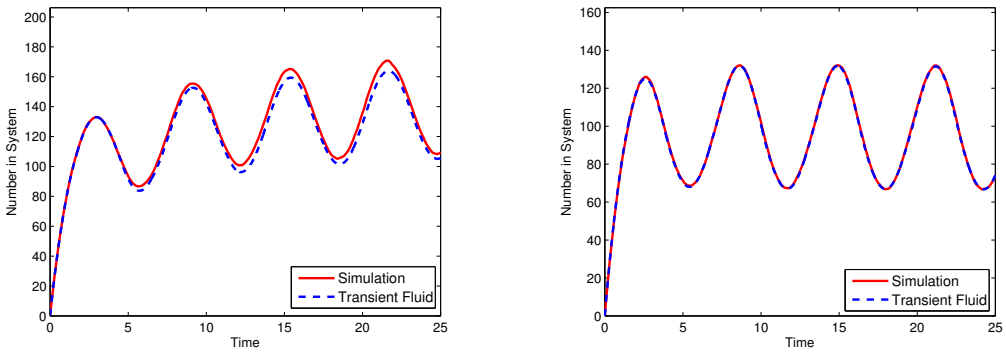
I use simulation to demonstrate that the accuracy of the fluid model is robust to time-varying arrivals. The systems have $s = 100$ agents and non-homogeneous Poission arrivals with periodic arrival rate $\lambda_s(t) = s \cdot (1 + 0.4\sin(t))$. The patience-time distribution is exponential with rate 1/2. I plot the number-in-system processes in Figures 3.8 and 3.9. Figure 3.8 corresponds to systems with an exogenous dependence, where the service and patience times generated by Gaussian copulas. Figure 3.9 corresponds to systems with an endogenous dependence, where the conditional service time is exponentially distributed. Both systems with positive and negative dependencies are simulated. I find that the fluid

model provides accurate approximations for the stochastic systems under time-varying arrivals and different dependence structures.



Positive Dependence | Negative Dependence

Figure 3.8. Number-in-system process with time-varying arrivals. Each system has $s = 100$ agents, Poisson arrivals with time-varying rate $\lambda_s(t) = s \cdot (1 + 0.4\sin(t))$, service-time distribution exp(1), patience-time distribution exp(1/2). Exogenous dependence, the joint distributions of service time and patience time are generated via Guassian copulas.



$\mu(z) = 0.6 + 0.4\exp(-1.8z)$ | $\mu(z) = 0.6 + 0.4\exp(1.8z)$

Figure 3.9. Number-in-system process with time-varying arrivals. Each system has $s = 100$ agents, Poisson arrivals with time-varying rate $\lambda_s(t) = s \cdot (1 + 0.4\sin(t))$, patience-time distribution exp(1/2). Endogenous dependence.

CHAPTER 4

Future Work

In this chapter, I comment on two directions for future research which are closely related to the previous chapters in this thesis. Despite the difference revealed in Chapter 3 between the exogenous and endogenous dependencies regarding their stationary behaviors, I first give a detailed analysis of the fundamental relation between the two seemingly different dependencies. Implications to the empirical estimation on a *censored* dataset are also discussed. In the second direction, I consider the pricing implication of another dependence structure in service systems: a customer's value of service acquisition depends on his own service requirement. Utilizing a queueing-game framework, I characterize the optimal pricing scheme and routing policy for a revenue-maximizing service provider.

4.1. More on Exogenous and Endogenous Dependencies

Following the setting in Chapter 3 where two relevant but seemingly different dependencies are introduced, I further discuss the relation between these two dependencies in this section. Recall in Chapter 3, I consider a multi-server queueing system with s statistically identical agents. Customers arrive to the system according to a stochastic process $A(t)$ which is right continuous with left limits. New arrivals enter service immediately if there is an available agent and are delayed in queue if all agents are busy. I assume that each customer has a finite patience for waiting to be served, and will abandon the queue if his waiting time exceeds that patience. A key feature of the underlying queueing model

is that a customer in service requires a service time that depends on his patience time or waiting time in queue.

More specifically, let S_i and T_i be customer i's service and patience times, respectively. I assume that customers' patience times $\{T_i : i \geq 1\}$ are independent of all other random variables in the model and are I.I.D. with cumulative distribution function (cdf) F_T and probability density function (pdf) f_T. Let $F_T^c := 1 - F_T$ be the complementary cdf (ccdf) of the patience time. Let V_i be the *offered wait* of customer i, which is the virtual waiting time of the customer if he had infinite patience. It then follows that customer i's actual waiting time $W_i = \min\{T_i, V_i\}$. To capture the dependence between customers' service, patience and waiting times, I introduce the following parametrized cdf of the *conditional service time* of served customers:

$$\Psi(v; x) := \mathbb{P}(S_i \leq x | T_i > W_i, W_i = v) = \mathbb{P}(S_i \leq x | T_i > V_i, V_i = v).$$

Assuming the conditional cdf $\Psi(v; x)$ is differentiable in x for all $v \geq 0$, the pdf of the conditional service time exists and satisfies

$$\psi(v; x) := \frac{\partial \Psi(v; x)}{\partial x}.$$

Exogenous Dependence. When the dependence is exogenous, I assume that each customer's patience level depends on his individual service requirement. See e.g., Bassamboo and Randhawa (2015) and Chapter 2. Each arriving customer is endowed with an exogenous service and patience time which are dependent. I assume the bivariate random variables representing the service and patience times are I.I.D. across different customers. Specifically, I assume $\{(S_i, T_i) : i \geq 1\}$ are I.I.D. bivariate random variables, all having

the same continuous joint density f, with marginal densities f_S and f_T. Furthermore, I assume (S_i, T_i) are independent of V_i.

To derive the distribution of the conditional service time, note that

$$\begin{aligned}\Psi(v;x) &\overset{(1)}{=} \mathbb{P}(S_i \leq s | T_i > V_i, V_i = v) \\ &= \mathbb{P}(S_i \leq s | T_i > v, V_i = v) \\ &\overset{(2)}{=} \mathbb{P}(S_i \leq s | T_i > v),\end{aligned}$$

where (1) follows from the definition of Ψ and (2) follows from the independence of (S_i, T_i) and V_i. Since (S_i, T_i) has joint density f, the distributions of the conditional service time satisfy

$$\Psi(v;x) = \frac{\int_v^\infty \int_0^x f(x,y)dxdy}{\int_v^\infty \int_0^\infty f(x,y)dxdy} \quad \text{and} \quad \psi(v;x) = \frac{\int_v^\infty f(x,y)dy}{\int_v^\infty \int_0^\infty f(x,y)dxdy}. \tag{4.1}$$

Let $\mathcal{D}_{ex}(A, s) = \{A, s, f\}$ be the *primitive model data* of a system with three components: an arrival process A, capacity s and an exogenous dependence characterized by the joint distribution of service and patience times f.

Endogenous Dependence. When the dependence is endogenous, each customer' service time changes in response to his waiting time in queue. To capture this feature, I assume that a customer's service time depends on his offered wait. It then follows that the service times of *served* customers depend on their actual waiting times. Formally,

$$\Psi(v;x) = \mathbb{P}(S_i \leq s | T_i > V_i, V_i = v) = \mathbb{P}(S_i \leq s | V_i = v) := \Xi(s, v).$$

Let $\mathcal{D}_{en}(A, s) = \{A, s, \xi, f_T\}$ be the *primitive model data* of a system with four components: an arrival process A, capacity s, patience-time distribution f_T and an endogenous dependence characterized by the conditional service-time distribution ξ.

4.1.1. Relation between Two Dependencies

Equivalence Class. For a given arrival process $A(t)$ and capacity s, an equivalence class is defined as a set of model data characterized by distributional primitives such that the system dynamics of the stochastic queueing systems with any model data in this equivalence class have the same law. To formally state the concept, let $\mathcal{L}(\mathcal{D}(A, s))$ be the law of the offered wait process of a system with an arrival process A, capacity s and distributional model data $\mathcal{D}$ describing the dependence in the system.

Definition 4.1. *Two systems with model data $\mathcal{D}_1$ and $\mathcal{D}_2$ are equivalent if*

$$\mathcal{L}(\mathcal{D}_1(A, s)) \stackrel{d}{=} \mathcal{L}(\mathcal{D}_2(A, s)) \quad \textit{for all } A \textit{ and } s,$$

given the same initial condition of the queueing systems, where $\stackrel{d}{=}$ represents equality in distribution.

By Definition 4.1, if two systems are equivalent, then their offered processes must have the same law. Further, it can be shown that all the system dynamics of the two systems must have the same law. Therefore, if two different dependencies give rise to two systems which are equivalent, and if we are only interested in the distributions of performance functions, then we don't really have to distinguish the exact form of dependence in the system.

Proposition 4.1. *Two systems with model data specified as follows are equivalent:*

$$\mathcal{D}_{ex}(A,s) = (A,s,f) \text{ and } \mathcal{D}_{en}(A,s) = (A,s,\xi,f_T),$$

where f_T is the patience-time distribution and $\xi(v;x) = \partial\mathbb{P}(S \le x|T > v)/\partial x$ derived from f. To be specific,

$$f_T(v) = \int_0^\infty f(x,v)dx \text{ and } \xi(v;x) = \frac{\int_v^\infty f(x,y)dy}{\int_v^\infty f_T(y)dy}. \tag{4.2}$$

Proof. To develop intuition, I first give the proof in the single-server case. I use superscript *ex* and *en* to distinguish between the exogenous and endogenous models. Let α_n denote the inter-arrival time between the $n-1^{st}$ and n^{th} customer. For customer n, in the exogenous model, he arrives with a service and patience time pair (S_n, T_n) which is randomly drawn from the joint distribution f. Following Baccelli et al. (1984, Eq 2.1), the offered wait process can be characterized using a recursion

$$V_{n+1}^{ex} = [V_n^{ex} + \mathbf{1}_{\{T_n^{ex} > V_n^{ex}\}} S_n^{ex} - \alpha_{n+1}^{ex}]^+. \tag{4.3}$$

Customer n enters service if and only if his patience T_n exceeds his virtual waiting time V_n, in which case his service requirement is counted in the service process of the service agent. Otherwise, he abandons the queue and his service requirement is removed from the service process. Note that in (4.3), the effective workload of customer n into the service process $\mathbf{1}_{\{T_n^{ex} > V_i^{ex}\}} S_i^{ex}$ has a probability mass $P(T_n^{ex} < V_n^{ex})$ at 0. Suppose V_n^{ex} has

a distribution $f_{V_n}^{ex}$, then for any $v > 0$, it holds that

$$\begin{aligned}
&P(V_n^{ex} + \mathbf{1}_{\{T_n^{ex} > V_n^{ex}\}} S_n^{ex} > v) \\
=&P(V_n^{ex} > v) + P(S_n^{ex} > v - V_n^{ex}, T_n^{ex} > V_n^{ex}, V_n^{ex} \leq v) \\
\overset{(1)}{=}&P(V_n^{ex} > v) + \int_{u=0}^{v} P(S_n^{ex} > v - u, T_n^{ex} > u) f_{V_n}^{ex}(u) du \\
\overset{(2)}{=}&P(V_n^{ex} > v) + \int_{u=0}^{v} \int_{y=u}^{\infty} \int_{x=v-u}^{\infty} f(x,y) dx dy \; f_{V_n}^{ex}(u) du,
\end{aligned}$$

where (1) follows because (S_n^{ex}, T_n^{ex}) is independent of V_n^{ex} and (2) follows because (S_n^{ex}, T_n^{ex}) is drawn from the joint distribution f.

In the endogenous model,

$$V_{n+1}^{en} = [V_n^{en} + \mathbf{1}_{\{T_n^{en} > V_n^{en}\}} S_n^{en} - \alpha_{n+1}^{en}]^+. \tag{4.4}$$

The effective service time of customer n into the service process $\mathbf{1}_{\{T_n^{en} > V_n^{en}\}} S_n^{en}$ has a probability mass $P(T_n^{en} < V_n^{en})$ at 0. Suppose V_n^{en} has a distribution $f_{V_n}^{en}$, then for any

$v > 0$, it holds that

$$
\begin{aligned}
&P(V_n^{en} + \mathbf{1}_{\{T_n^{en} > V_i^{en}\}} S_n^{en} > v) \\
=&P(V_n^{en} > v) + P(S_n^{en} > v - V_n^{en}, T_n^{en} > V_n^{en}, V_n^{en} \leq v) \\
=&P(V_n^{en} > v) + \int_{u=0}^{v} P(S_n^{en} > v - u, T_n^{en} > v | V_n^{en} = u) f_{V_n}^{en}(u) du \\
=&P(V_n^{en} > v) + \int_{u=0}^{v} P(S_n^{en} > v - u | V_n^{en} = u) P(T_n^{en} > u)\ f_{V_n}^{en}(u) du, \\
\overset{(3)}{=}&P(V_n^{en} > v) + \int_{u=0}^{v} \int_{x=v-u}^{\infty} \int_{y=u}^{\infty} \frac{f(x,y)}{\int_{y=u}^{\infty} f_T(y) dy} dy dx\ P(T_n^{en} > u) f_{V_n}^{en}(u) du \\
=&P(V_n^{en} > v) + \int_{u=0}^{v} \int_{y=u}^{\infty} \int_{x=v-u}^{\infty} f(x,y) dx dy\ f_{V_n}^{en}(u) du,
\end{aligned}
$$

where (3) follows from the definition of ξ. Note that α_{i+1} is independent of $\{V_n, S_n, T_n\}$ in both models and $\alpha_{n+1}^{ex} \overset{d}{=} \alpha_{n+1}^{en}$. If $V_0^{ex} \overset{d}{=} V_0^{en}$, then an induction argument proves $V_n^{ex} \overset{d}{=} V_n^{en}$ for all $n \geq 1$.

The intuition in the single-server case can be easily transferred to the many-server case except that a more subtle recursive representation of workload vector is required. See e.g., Moyal (2017). For a system with s servers, define an s-dimensional workload vector at the arrival time of the n^{th} customer $V_n = (V_n(1), V_n(2), \ldots, V_n(s))$. This vector is ranked in the increasing order such that $V_n(1) \leq V_n(2) \leq \ldots \leq V_n(s)$. To interpret this workload vector, consider another $s-1$ *virtual* customers arriving at the same time of customer n. Index these s customers arriving simultaneously by $\{1, 2 \ldots, s\}$ with the *real* customer n being indexed by 1. For $1 \leq i \leq s$, let $V_n(i)$ represent the virtual waiting time of customer i if he has infinite patience. Then $V_n(1)$ corresponds to the offered wait of the real customer n. Under this representation, the workload vector V_n is updated when

the $n+1^{st}$ arrives:

$$\begin{cases} V_{n+1}(i) = \left[\left(V_n(i) \vee (V_n(1) + \mathbf{1}_{\{V_n(1)<T_n\}} S_n)\right) \wedge V_n(i+1) - \alpha_{n+1}\right]^+, \ 1 \le i \le s-1 \\ V_{n+1}(s) = \left[V_n(s) \wedge (V_n(1) + \mathbf{1}_{\{V_n(1)<T_n\}} S_n) - \alpha_{n+1}\right]^+. \end{cases}$$

In the exogenous model,

$$\begin{aligned} & P(V_n^{ex}(1) + \mathbf{1}_{\{V_n(1)<T_n\}} S_n > v_1, V_n^{ex}(2) > v_2, \ldots, V_n^{ex}(s) > v_s) \\ &= P(V_n^{ex}(1) > v_1, \ldots, V_n^{ex}(s) > v_s) \\ &+ \int_{u=0}^{v_1} P(T_n^{ex} > u, S_n^{ex} > v_1 - u, V_n^{ex}(2) > v_2, \ldots, V_n^{ex}(s) > v_s | V_n(1)^{ex} = u) f_{V_n(1)}^{ex}(u) du \\ &= P(V_n^{ex}(1) > v_1, \ldots, V_n^{ex}(s) > v_s) \\ &+ \int_{u=0}^{v_1} \int_{y=u}^{\infty} \int_{x=v_1-u}^{\infty} f(x,y) dx dy \ P(V_n^{ex}(2) > v_2, \ldots, V_n^{ex}(s) > v_s | V_n(1)^{ex} = u) f_{V_n(1)}^{ex}(u) du. \end{aligned}$$

In the endogenous model,

$$\begin{aligned} & P(V_n^{en}(1) + \mathbf{1}_{\{V_n(1)<T_n\}} S_n > v_1, V_n^{en}(2) > v_2, \ldots, V_n^{en}(s) > v_s) \\ &= P(V_n^{en}(1) > v_1, \ldots, V_n^{en}(s) > v_s) \\ &+ \int_{u=0}^{v_1} P(T_n^{en} > u, S_n^{en} > v_1 - u, V_n^{en}(2) > v_2, \ldots, V_n^{en}(s) > v_s | V_n(1)^{en} = u) f_{V_n(1)}^{en}(u) du \\ &\overset{(4)}{=} P(V_n^{en}(1) > v_1, \ldots, V_n^{en}(s) > v_s) \\ &+ \int_{u=0}^{v_1} \int_{y=u}^{\infty} \int_{x=v_1-u}^{\infty} f(x,y) dx dy \ P(V_n^{en}(2) > v_2, \ldots, V_n^{en}(s) > v_s | V_n(1)^{en} = u) f_{V_n(1)}^{en}(u) du, \end{aligned}$$

where (4) follows from the same computation of the endogenous model in the single-server case. If $V_0^{ex} \overset{d}{=} V_0^{en}$, then an induction argument proves $V_n^{ex} \overset{d}{=} V_n^{en}$ for all $n \ge 1$. Q.E.D.

Proposition 4.2. *A system with an endogenous model data* $\mathcal{D}_{en}(\lambda, s) = (\lambda, s, \xi, f_T)$ *is equivalent to a system with an exogenous model data if and only if* $(1 - \Xi(v; x))F_T^c(v)$ *is decreasing in both* v *and* x*, where* Ξ *is the corresponding cdf of the conditional service time,* $\Xi(v; x) = \int_0^x \xi(v; x)dx$*.*

Proof. If an endogenous model has an exogenous counterpart and there exists a two-dimensional density f such that (4.2) holds. It follows immediately that

$$(1 - \Xi(v; x))F_T^c(v) = \int_{y=v}^{\infty}\int_{x}^{\infty} f(x, y)dxdy \tag{4.5}$$

is decreasing in both v and x.

On the other hand, suppose $(1 - \Xi(v; x))F_T^c(v)$ is decreasing in both v and x. Note that

$$\lim_{v\to\infty}\lim_{x\to\infty}(1 - \Xi(v; x))F_T^c(v) = 0 \quad \text{and} \quad \lim_{v\to 0}\lim_{x\to 0}(1 - \Xi(v; x))F_T^c(v) = 1.$$

This implies $(1 - \Xi(v; x))F_T^c(v)$ can be represented as a two-dimensional ccdf. Assuming the derivate exists, the joint density f in the exogenous model can be computed via

$$f(x, v) = \frac{\partial^2(1 - \Xi(v; x))F_T^c(v)}{\partial x \partial v}.$$

It can be verified that such f in the exogenous model will induce f_T and ψ in the endogenous model via (4.2). Invoking Proposition 4.1, the exogenous model with density f has the same system dynamics as in the endogenous model specified by f_T and ψ. Q.E.D.

Proposition 4.2 gives a sufficient and necessary condition to distinguish the two dependencies. It is a stronger than Condition 1 in Chapter 3 developed to evaluate the

uniqueness of equilibrium for the fluid approximation of the stochastic system. Proposition 3.3 states that, for a fluid model with stationary arrivals, the equilibrium is unique if and only if $\mathbb{E}[S|T > V, V = v]\mathbb{P}(T > v)$ is decreasing in v. This latter condition trivially holds when $(1-\Xi(v;x))F_T^c(v)$ is decreasing in both v and x. It then follows from Proposition 4.1 and Proposition 4.2 that a fluid model with an exogenous dependence must have a unique equilibrium.

Proposition 4.2 also implies that the possible system dynamics given by exogenous models form a subset of those given by endogenous models. Although the two dependencies correspond to very different mechanisms in the stochastic system regarding how customers' service times are determined, they can, in fact, give rise to the same system performance when the condition in Proposition 4.2 is satisfied.

Proposition 4.2 has very important implication on the empirical identification of dependence. The estimation of an exogenous dependence characterized by the joint distribution of the service and patience times are, in general, very difficult given that the dataset is necessarily censored, as I remark at the end of Chapter 2. The relation between the exogenous and endogenous dependence, as demonstrated in Proposition 4.2, indicates that it is only necessary to estimate the primitives of its endogenous representation, even if the dependence is physically exogenous in the service system. It is significant that the estimation of the endogenous model is much easier even if the dataset is censored, which boils down to estimating the marginal distribution of the patience time and the conditional service-time distribution (conditioned on the waiting time). I elaborate below.

First, one can use the Kaplan-Meier estimator to extract the patience-time distribution since waiting times are observed for both *abandoned* and *served* customers. Let n denote

the number of customers observed in the sample, among whom J customers abandon and the other $N - J$ customers are served. One can rank the waiting times observed from abandoned customers in the increasing order such that $0 = t_0 < t_1 < t_2 < \ldots < t_J < t_{J+1} = \infty$. The patience-time distribution can be estimated as follows:

$$\hat{F}_T(x) = \prod_{t_j \leq x} \left(1 - \frac{\#\text{ customers who abandon at } t_j}{\#\text{ customers who have not abandoned by } t_j-}\right)$$

The conditional service-time distribution can be estimated by analyzing the service times observed from served customers. Specifically, let a_i indicate whether that customer is served. It equals 1 if that customers gets served and 0 if that customer abandons. If $a_i = 1$ so that customer i is served, let s_i and w_i be the service and waiting times observed from customer i. The conditional service-time distribution can be estimated as follows:

$$\hat{\Psi}(v; x) = \frac{\sum_{i=1}^{n} \mathbf{1}\{a_i = 1, w_i = v, s_i \leq x\}}{\sum_{i=1}^{n} \mathbf{1}\{a_i = 1, w_i = v\}}.$$

4.1.2. Performance Analysis: A Fluid Approach

Since the exact analysis of performance measures under either dependence is intractable, I develop a stationary fluid model (similar to Chapter 2) to approximate the steady-state performance measures. The following analysis generalizes some of the results in Chapter 2 derived for systems with an exogenous dependence.

Define

$$a(v) := \mathbb{E}[S|T > V, V = v] = \int_0^\infty x\psi(v; x)dx$$

to be the average conditional service time of a customer given that he has not abandoned after waiting v time units in queue. I assume $a(v)\mathbb{P}(T > v)$ is strictly decreasing in v.

Assumption 4.1. *$a(v)\mathbb{P}(T > v)$ is strictly decreasing in v for all v.*

Chapter 2 shows that Assumption 4.1 holds when the dependence is exogenous. To see this, recall $g(v) := \mathbb{E}[S|T = v]$ is the average service time of a customer whose patience time is v, which we refer to as the Conditional Service Time in Chapter 2. Differentiating $a(v)\mathbb{P}(T > v)$ gives the derivative $-f_T(v)[a(v)-a'(v)/h_T(v)] = -f_T(v)g(v) < 0$, where the equality follows because Reich (2012, Equation (4.5)) shows that $a(v) - a'(v)/h_T(v) = g(v)$, so that Assumption 4.1 must hold. Now, if the dependence is endogenous and further, if a is decreasing (implying a negative dependence between service and waiting times), Assumption 4.1 also holds. However, if a is increasing (implying a positive dependence between service and waiting times), Assumption 4.1 may be violated.

Assumption 4.1 can be used to derive the uniqueness of an equilibrium for the fluid model when the arrival process is stationary. Indeed, Proposition 3.3 shows this. It follows immediately that the equilibrium is unique when the dependence is exogenous. However, Chapter 3 uses numerical examples to demonstrate that multiple equilibria may exist when the dependence is endogenous and Assumption 4.1 fails. In the following, I always assume that Assumption 4.1 holds for any dependence under consideration so that a unique equilibrium exists. The steady-state performance measures are derived based on that unique equilibrium.

Proposition 4.2 indicates that the law of system dynamics only depends on the arrival rate of customers λ, number of agents s, patience-time distribution f_T and the distribution of the conditional service time ψ. In other words, the primitive model data $\mathcal{D} := \{\lambda, s, f_T, \psi\}$ fully determines the dynamics of the fluid model. I analyze the impact of each of the components in $\mathcal{D}$ on the stationary fluid system by fixing the other three

components. I use the throughput R as the performance measure to demonstrate the comparative statics.

Let $R(\lambda)$ be the throughput when the arrival rate is λ and other data components are fixed. The following result generalizes Proposition 2.3.

Proposition 4.3. *$R(\lambda)$ is decreasing if a is increasing and is decreasing if a is decreasing.*

Let $R(s)$ be the throughput when the capacity is s and other data components are fixed.

Proposition 4.4. *If f_T has an increasing hazard rate, then $R(s)$ is convex increasing if a is concave increasing and concave increasing if a is convex decreasing.*

The conditions in Proposition 4.4 are different from those in Proposition 2.7. In Proposition 2.7 where an exogenous dependence is concerned, the structure of the throughput only depends on a simple condition on the monotonicity of g. Proposition 4.4 applies to a more general dependence but requires the patience-time distribution to have an increasing hazard rate.

Fix the arrival rate λ, capacity s and distribution of the conditional service time ψ. Let T_1 and T_2 denote two random variables which can be ranked by the first-order stochastic dominance. Let R_i be the throughput in the fluid model of a system with patience time distributed as T_i.

Proposition 4.5. *Suppose $T_1 \leq_{st} T_2$. It follows that $R_1 \geq R_2$ if a is increasing and that $R_1 \leq R_2$ if a is decreasing.*

Proposition 4.5 demonstrates the impact of the patience-time distribution, which is not considered in Chapter 2. The proposition shows that a high patience level implies a low throughput given a positive dependence. This is because as customers become more patient, they are less likely to abandon and it takes longer to clear the queue. Customers must wait longer to get served and tend to require longer service times due to the positive dependence, leading to a lower throughput. Similarly, the throughput is higher when the patience level is higher but the dependence is negative.

Fix the arrival rate λ, capacity s and patience-time distribution f_T. Let a_1 and a_2 be two functions representing the average conditional service times. Let R_i be the throughput in the fluid model of a system with conditional service time a_i.

Proposition 4.6. *If $a_1(v) \leq a_2(v)$ for all $v > 0$, then $R_1 \geq R_2$.*

Proposition 4.6 generalizes Proposition 2.5, which shows that the throughput is decreasing in the strength of the dependence as ranked by PQD order. To see how Proposition 4.6 implies Proposition 2.5, note that the ordering of the average conditional service time follows from the PQD order, as is shown in Shaked and Shanthikumar (2007, p. 389).

4.2. On the Dependence between Service Value and Service Requirement

In the second direction for future research, I consider a dependence between the service value and service requirement in service systems and its implication on the revenue management of these systems.

In a wide variety of services, the value a customer receives from service may depend on his individual service requirement. For example, the value of visiting a grocery store is

likely to depend on the number of items a customer wishes to purchase, and that number is positively correlated with the time spent in the store, picking up items and checking out. Similarly, the value of visiting a doctor and receiving medical care is likely to be high if a patient's illness is acute, which in turn, is likely to imply a long treatment time. The dependence between service value and service requirement has been supported by empirical evidence. Analyzing a dataset from a call center, Hu et al. (2016) find that a caller's service reward tends to be larger when a call center agent spends more time talking to that caller. The authors argue that the longer one's service time is, the more complicated and urgent the service request is to the caller, and the larger reward is generated from service.

In this section, I consider a queueing model which explicitly captures the dependence described above. My goal is to study the impact of such dependence on the provider's pricing decision and revenue performance. In my model, customers are rational and self-interested: they decide whether to queue for service based on their individual valuations, waiting cost and price of service. The dependence between the valuation and service requirement considered in my model has significant implications on the system performance. In particular, it implies that the service-time distribution of customers who join the queue is in general different than that of the entire customer population. This property distinguishes my work from the existing literature, which, in general, assumes identical service-time distributions for customers who join and who balk (see e.g., Anand et al. (2011)).

There has been a rich literature on service operations that studies the joining/balking decisions of rational customers based on self-interest. Most of the existing research treats

the system primitives to be independent, in particular, it is assumed that a customer's valuation for service is independent of his service requirement. My work contributes to the literature by explicitly modeling and analyzing a dependence between a customer's valuation and service requirement.

Relevant to the dependence considered in this section, Hopp et al. (2007) and Alizamir et al. (2013) use dynamic models to study discretionary services, in which the service rate can be adjusted at the provider's discretion. The value of the service increases with service quality, which in turn increases with service time. In a similar spirit, Anand et al. (2011) consider the interaction between the service value and service time using a static queueing game. All these papers assume service speeds are adjustable and the key research question addressed in these papers is to determine the optimal trade-off between providing high-valued services and serving more customers. My model, on the other hand, studies service settings where customers' service requirements are exogenous and cannot be easily controlled by the provider.

Recently, there has been research on consumer-driven dependencies between system primitives. Nazerzadeh and Randhawa (2015) assume that a customer's sensitivity to delays increases with his valuation and the authors find that offering two service grades is asymptotically optimal in revenue maximization. Gurvich et al. (2016) extend their model to study the social welfare optimization. Afèche and Pavlin (2016) make a similar assumption on customer valuation and delay sensitivity. The authors study the optimal price and delay menus for revenue maximization under asymmetric information. It is worth noting that all the references above assume a special correlation structure between the primitives under study. In particular, a customer's delay sensitivity is a deterministic

function of his service value. By contrast, the interaction between the service value and service requirement in my model is captured using a stochastic dependence concept, and is more reflective of the heterogeneity in customer preferences.

4.2.1. Model

I consider a monopoly service provider selling service to a market of customers. I model the service process using an invisible $M/G/1$ queueing system. Potential customers arrive for service in a Poisson stream with an exogenous rate Λ, which I refer to as the *market size* or *demand* interchangeably. Each arriving customer has a finite random valuation for service and decides to join service if the utility of obtaining service is nonnegative. A key feature of my model is that the valuation of an arriving customer depends on that customer's individual service requirement, although the bivariate random variables representing the valuation and service requirement are independent across different customers.

Specifically, letting V_i and S_i denote the valuation and service requirement of customer i, respectively, I assume that $\{(V_i, S_i) : i \geq 1\}$ are independent and identically distributed (I.I.D.) bivariate random variables, all having the same continuous joint density f and marginal densities f_V and f_S for valuation V and service requirement S, respectively. Suppose f_V has support in $[\underline{v}, \bar{v}]$ while the support of f_S is the entire positive half of the real line. I further assume that $\mathbb{E}[S^2] < \infty$, so that unconditional service time has a finite expectation. I refer to $\mu := 1/\mathbb{E}[S] > 0$ as the *nominal service capacity*, because μ would be the service rate if all customers choose to join service.

Each customer incurs a cost c for each time unit waiting in queue, which I refer to as the *delay sensitivity* and is assumed to be homogeneous across customers. In the benchmark

case, the provider charges a uniform price p for service, which is processed in a First Come First Served (FCFS) manner. I will relax this assumption and consider priority-based pricing schemes in §**??**. Upon arrival, each customer decides whether to acquire service (joining) or quit service (balking) based on his individual valuation, expected waiting cost and price charged for service. To formally characterize a customer's decision, let W be the random variable representing the equilibrium delay time in queue of a joining customer.. A customer joins service if his utility of obtaining service is positive:

$$U := V - p - c\mathbb{E}[W] \geq 0.$$

Otherwise, he balks. I assume that joining customers do not renege and that balking customers do not retry.

In my model, a customer's waiting cost is measured by his waiting time in queue. This assumption is supported by various psychological studies which claim that customers usually perceive the time in queue to be unproductive and thus undesirable. The time in service, however, produces value and is often not associated with a mental cost. This fact is emphasized in my model with dependence, especially when a customer's valuation is positively dependent on his service requirement. In this case, the time in service is valued more than it is if the valuation and service requirement were independent.

Since the price p and waiting cost $c\mathbb{E}[W]$ are the same across customers, there exists a valuation threshold v such that customers join the service if their valuation exceed v

and balk otherwise. This implies

$$\begin{cases} v - c\mathbb{E}[W(v)] - p = 0 & \text{if } v > \underline{v}, \\ v - c\mathbb{E}[W(v)] - p \geq 0 & \text{if } v = \underline{v}, \end{cases} \tag{4.6}$$

where $W(v)$ is the equilibrium delay given that the valuation threshold is v. Consider the customers who join service. By Poisson thinning, they enter service in a Poisson stream with rate $\lambda = \Lambda \bar{F}_V(v)$. If V and S are dependent, the service-time distribution of these customers also depends on v. I refer to the service time of joining customers as the *effective service time*, which I represent using a generic random variable S_e. It is the service time of customers whose valuation is greater than the threshold v, which, in general, has a different distribution than the nominal service time S. This feature distinguishes my model from the literature, which typically assumes an identical distribution for S_e and S. See e.g., Anand et al. (2011), Cachon and Feldman (2011) and Huang et al. (2013).

It is worth noting that when the valuation and service requirement are independent, models with waiting cost measured by the time in queue and the time in system (queueing time plus service time) are qualitatively the same because the distribution of the effective service time does not depend on customers' joining decisions. However, in the presence of a dependence, the effective service time depends on the valuation threshold, exposing an intricate difference between the two measurements of waiting costs.

4.2.1.1. Measures of Dependence. While there are various ways to model dependence in system primitives, a commonly used approach in the prior studies to capturing dependence between two primitives is to assume that one primitive is a deterministic function of the other. For example, Afèche and Pavlin (2016) and Nazerzadeh and Randhawa (2015)

assume that a customer's delays sensitivity is an increasing function of his valuation. However, such a modeling approach is restrictive in that it is unable to capture a stochastic dependence between the primitives, which is important to reflect the heterogeneity in customer preferences. By assuming a joint distribution for the valuation and service requirement, my model allows for any arbitrary stochastic dependence to be captured. The dependence studied in Afèche and Pavlin (2016) and Nazerzadeh and Randhawa (2015) can be be considered as a special case of my general modeling framework.

In this section, I introduce two measures of dependence which I will use later. First, I will identify the service provider's optimal pricing strategy for a given joint density f. To gain qualitative insights, I will impose structural assumptions on f by assuming that the conditional expected service time (conditioned on the valuation) is monotone; I elaborate in §4.2.1.1. Second, to compare the optimal revenues for systems with different dependencies, I consider the set of all bivariate distributions with the same marginal densities f_V and f_S, which I denote by $\mathcal{F}(f_V, f_S)$ and is nonempty. See §2.3.

Measuring Dependence via Bivariate Dependence Orders. The first dependence concept is the PQD order defined in Chapter 2, where it was employed to rank the strength of dependence between customers' service and patience times in the many-server setting.

Conditional Expected Service Time. In addition to the PQD order, dependence between V and S can also be captured by the conditional expected service time given a customer's service value. In particular, let

$$\begin{aligned} a(v) &= \mathbb{E}[S|V > v],\ b(v) = \mathbb{E}[S^2|V > v], \\ g(v) &= \mathbb{E}[S|V = v],\ h(v) = \mathbb{E}[S^2|V = v]. \end{aligned} \tag{4.7}$$

For the notational convenience later, I also define

$$A(v) = \bar{F}_V(v)a(v),\ B(v) = \bar{F}_V(v)b(v). \tag{4.8}$$

I refer to the function g as the Conditional Service Time (CST) and h as the Second-moment of Conditional Service Time (SCST). An increasing CST and SCST imply a positive dependence, whereas a decreasing CST and SCST imply a negative dependence, between V and S. The independence between V and S implies a constant CST and a constant SCST.

In general, for a given bivariate random variable (V, S), the CST and SCST need not be monotone. The following lemma provides natural sufficient conditions for monotone CST and SCST, and link the monotonicity of the CST and SCST to PQD.

Lemma 4.1. *If $\mathbb{P}(S > s|V = v)$ is increasing in v, then (V, S) is PQD and has an increasing CST and SCST. If $\mathbb{P}(S > s|V = v)$ is decreasing in v, then (V, S) is NQD and has a decreasing CST and SCST.*

4.2.1.2. Service Price. The service provider collects revenue from customers served. In the benchmark model, the provider determines a uniform price p to charge for service. The effective service-time distribution of customers who receive service depends on the valuation of these customers, which in turn depends on the provider's pricing decision. Therefore, it is important that one can show there exists a unique subgame equilibrium under each price p charged by the provider.

To show a unique subgame equilibrium exists, I first characterize the expected delay $\mathbb{E}[W(v)]$ given a valuation threshold v. Using the Pollaczek-Khinchine formula, the

expected delay (in queue) $\tilde{W}$ for an $M/G/1$ queue with arrival rate $\tilde{\lambda}$ and service time $\tilde{S}$ can be computed by $\mathbb{E}[\tilde{W}] = \tilde{\lambda}\mathbb{E}[\tilde{S}^2]/2(1-\tilde{\lambda}\mathbb{E}[\tilde{S}])$. For a system with a valuation threshold v, it follows that $\tilde{\lambda} = \Lambda\bar{F}_V(v)$, $\mathbb{E}[\tilde{S}] = \mathbb{E}[S_e] = \mathbb{E}[S|V > v] = a(v)$ and $\mathbb{E}[\tilde{S}^2] = \mathbb{E}[S_e^2] = \mathbb{E}[S^2|V > v] = b(v)$. Hence the expected delay can be computed via

$$\mathbb{E}[W(v)] = \frac{\Lambda\bar{F}_V(v)b(v)}{2(1-\Lambda\bar{F}_V(v)a(v))} = \frac{\Lambda B(v)}{2(1-\Lambda A(v))}, \tag{4.9}$$

provided the system is stable: $\Lambda A(v) < 1$, where A and B are defined in (4.8).

Proposition 4.7. *Fix Λ and f, then for each price p, there exists a unique valuation threshold v. Moreover, the valuation threshold v is increasing in p and the expected delay $\mathbb{E}[W(v)]$ is decreasing in p.*

Proposition 4.7 shows that the waiting time decreases with the service price, implying that a higher admission fee can be used to reduce congestion. Note that when V and S are positively dependent, as the price increases, the valuation threshold also increases. This implies customers who join service must have higher valuations. When the dependence is positive, these customers require longer service times. Since a high price reduces congestion, Proposition 4.7 shows that the decrease in the effective arrival rate dominates the increase in the effective service time. Mathematically, the average of the effective service time $a(v) = \mathbb{E}(S|V > v)$ might increase to infinity as $v \to \bar{v}$ due to a positive dependence. The assumption that $\mathbb{E}[S] < \infty$, however, ensures that $A(v) = \bar{F}_V(v)a(v)$ is strictly decreasing and converges to 0 as $v \to \bar{v}$. Similarly, $B(v)$ is also strictly decreasing

and converges to 0 as $v \to \bar{v}$. It then follows from (4.9) that the expected delay $\mathbb{E}[W(v)]$ is decreasing in v, thus decreasing in price.

Proposition 4.7 shows that each price charged by the provider induces a unique subgame perfect equilibrium. Using a backward induction, the provider seeks the optimal price p^* to maximize her revenue:

$$\begin{aligned} &\max_{p\geq 0, v\in[\underline{v},\bar{v}]} && p\Lambda \bar{F}_V(v) \\ &\text{s.t.} && v - p - c\mathbb{E}[W(v)] \geq 0, \\ & && \Lambda A(v) < 1. \end{aligned} \tag{4.10}$$

The first constraint in (4.10) corresponds to the individual rationality of joining customers, i.e., the utility of obtaining service for the customer whose valuation equals the valuation threshold is nonnegative. The second constraint ensures that the queueing system is stable.

The first constraint must be binding for the optimal price p^*. In other words, the customer whose valuation equals the valuation threshold gains zero utility. To see this, suppose this constraint is nonbinding, then the provider could be strictly better-off by increasing the price by a sufficiently amount while maintaining the same valuation threshold v and throughput $\Lambda \bar{F}_V(v)$. Since the constraint is binding, I can represent the price as a function of the valuation threshold v via $p = v - c\mathbb{E}[W(v)]$. I can equivalently optimize over the valuation threshold v and restate the optimization problem:

$$\max_{v\in\mathcal{V}} \ (v - \mathbb{E}[W(v)])\Lambda \bar{F}_V(v), \tag{4.11}$$

where $\mathcal{V}$ is the feasible region of the valuation threshold such that the system is stable, i.e.,

$$\mathcal{V} := \{v \in [\underline{v}, \bar{v}] : \Lambda A(v) < 1\}.$$

I assume that there is a unique v^* which solves (4.11) and the optimal price can be uniquely computed by $p^* = v^* - c\mathbb{E}[D(v^*)]$.

Let $\mathcal{D} := \{\Lambda, c, f\}$ be the primitive model data, a set containing all the system primitives. It endogenizes the optimal price of the provider, and in turn, determines the joining and balking decisions of the customer population. This implies that the queueing game is fully characterized by $\mathcal{D}$.

4.2.2. Revenue Maximization

In this section, I consider the provider's revenue maximization problem with and without a dependence in the service system. By studying the independent and dependent models separately, I show that some traditional wisdom in the independent model may fail to hold in the dependent model. I also demonstrate how the strength of the dependence, as ranked by PQD order, impacts the optimal revenue.

4.2.2.1. Benchmark: Independent Model. I first consider the benchmark model in which a customer's valuation is *independent* of his service requirement. Under this assumption, the effective and nominal service times have the same distribution. I first identify the optimal price of the provider, assuming the service time has an exponential distribution. Next, I relax the distributional assumption of service time and study the impact of the market size and delay sensitivity on the optimal revenue.

If a customer's valuation is independent of his service requirement, then the functions defined in (4.7) are all constants, i.e., f has a constant CST and SCST. If I further assume that the service time of potential customers is exponentially distributed with rate μ, then $a(v) = g(v) = 1/\mu$ and $b(v) = h(v) = 2/\mu^2$ for all $v \geq 0$. The expected delay in this case is

$$\mathbb{E}[W(v)] = \frac{\Lambda \bar{F}_V(v)/\mu^2}{1 - \Lambda \bar{F}_V(v)/\mu} = \frac{\Lambda}{\mu} \frac{\bar{F}_V(v)}{\mu - \Lambda \bar{F}_V(v)}.$$

The feasible region $\mathcal{V} = \{v \in [\underline{v}, \bar{v}] | \bar{F}_V(v) < \min\{1, \mu/\Lambda\}\}$. The revenue of the provider is

$$R(v) = \left(v - \frac{c\Lambda}{\mu} \frac{\bar{F}_V(v)}{\mu - \Lambda \bar{F}_V(v)} \right) \Lambda \bar{F}_V(v).$$

I characterize the provider's optimal pricing decision in the following proposition.

Proposition 4.8. *Suppose valuation V has an increasing hazard rate $h_V(\cdot)$ and service time S is exponentially distributed with rate μ. Further, if either of the following conditions holds:*

(i) $\Lambda \geq \mu$.

(ii) $\Lambda < \mu$ *and* $\underline{v} \leq \frac{1}{f_V(\underline{v})} + \frac{c(2\Lambda\mu - \Lambda^2)}{\mu(\mu - \Lambda)^2}$.

then $R(v)$ is quasiconcave in v and the optimal valuation threshold $v^* = argmax_v R(v)$ uniquely solves

$$v = \frac{c\Lambda \bar{F}_V(v)}{\mu(\mu - \Lambda \bar{F}_V(v))} + \frac{c\Lambda \bar{F}_V(v)}{(\mu - \Lambda \bar{F}_V(v))^2} + \frac{1}{h_V(v)}. \tag{4.12}$$

The optimal price $p^* = v^* - c\Lambda \bar{F}_V(v^*)/\mu(\mu - \Lambda \bar{F}_V(v^*))$.

The proof of Proposition 4.8 follows from Cachon and Feldman (2011, Proposition 1). Proposition 4.8 shows that $R(v)$ is quasiconcave in v, hence solving the first order condition (4.12) gives the optimal v^*. To ensure the existence of a solution to (4.12), Proposition 4.8 gives two conditions: Either the market size is greater than the nominal capacity so that the market can only be partially covered; or the lowest valuation in the population is small enough.

I next consider the impact of the market size on the provider's optimal revenue, allowing for a general distribution for the service time. The following proposition claims that the provider's optimal revenue increases with the market size Λ when V and S are independent.

Proposition 4.9. *If V and S are independent, the optimal revenue R^* is strictly increasing in the market size Λ.*

Proposition 4.9 generalizes a similar result in Huang et al. (2013, Proposition 7). To derive their result, Huang et al. (2013) make specific distributional assumptions on the valuation (Gumbel) and service time (exponential). My result only requires the service time to have a finite second moment and does not pose any distributional assumption on the valuation.

Proposition 4.9 shows that the service provider always benefits from a large market size. As the market size grows, the provider can be strictly better off by adjusting prices. The impact of the market size on the optimal price is characterized in the following corollary.

Corollary 4.1. *If the conditions in Proposition 4.8 are satisfied, then the optimal price p^* and the optimal effective arrival rate λ^* are strictly increasing in the market size Λ.*

By Corollary 4.1, the provider enjoys a high margin of price as well as a high throughput in a large market. However, as I will discuss in §4.2.3, the benefit of a large market size may vanish when the valuation depends on the service requirement. In particular, in a system with a positive dependence and a small service capacity, the provider may prefer a small market size.

By Proposition 4.9, the optimal revenue R^* is strictly increasing in the market size Λ. I can characterize the limit of R^* as the market size Λ grows indefinitely.

Proposition 4.10. *Suppose $\bar{v} < \infty$. Then*

$$\begin{aligned}
&\lim_{\Lambda\to\infty} v^* = \bar{v},\\
&\lim_{\Lambda\to\infty} p^* = \bar{v} + \frac{c}{\mu} - \sqrt{c(\bar{v} + \frac{c}{\mu})/\mu},\\
&\lim_{\Lambda\to\infty} R^* = \left(\bar{v} + \frac{c}{\mu} - \sqrt{c(\bar{v} + \frac{c}{\mu})/\mu}\right)\left(\mu - \sqrt{c\mu/(\bar{v} + \frac{c}{\mu})}\right).
\end{aligned}$$

The intuition behind Proposition 4.10 is straightforward. When the market size Λ is sufficiently large, the valuation threshold v needs to stay very close to $\bar{v}$ to ensure system stability. The joining customers are almost homogeneous in their valuation. With this observation, I can apply the result in Anand et al. (2011, Proposition 1) to determine the optimal price.

The impact of the delay sensitivity on the optimal revenue will be discussed in Proposition 4.14 of §4.2.2.2, where I show that the optimal revenue strictly decreases with the delay sensitivity regardless of the dependence in the system. The intuition is straightforward. As the delay sensitivity decreases, the provider can extract more surplus from customers as the waiting cost decreases.

4.2.2.2. Dependent Customer Valuation and Service Requirement. In this section, I consider the model with dependent valuation and service requirement. My analysis in §4.2.1.2 shows that the queueing game is fully characterized by the primitive model data $\mathcal{D} = \{\Lambda, c, f\}$. I will study the impact of each of the components in $\mathcal{D}$ on the provider's optimal revenue by fixing other components. I first analyze how the revenue is impacted by different dependence structures, employing the PQD order discussed in Chapter 2. Next, for a given joint distribution f, I utilize the monotone CST and SCST introduced in §4.2.1.1 to study the effect of changes to the market size Λ. Finally, I discuss the impact of the delay sensitivity.

Recall that given a valuation threshold v, the expected equilibrium delay $\mathbb{E}[W(v)] = \Lambda B(v)/2(1-\Lambda A(v))$, where A and B are defined in (4.8). The feasible region for v is $\mathcal{V} = \{v | A(v) < 1/\Lambda, v - \mathbb{E}[W(v)] \geq 0\}$. The provider's revenue is

$$R(v) = \left(v - \frac{c\Lambda B(v)}{2(1-\Lambda A(v)))}\right)\Lambda \bar{F}_V(v).$$

One can easily compute $A'(v) = -g(v)f_V(v)$ and $B'(v) = -h(v)f_V(v)$. Differentiating R gives:

$$R'(v)/\Lambda = \bar{F}_V(v) - f_V(v)v + \frac{c\Lambda f_V(v)\bar{F}_V(v)}{2}\left[\frac{h(v)+b(v)}{1-\Lambda A(v)} + \frac{\Lambda B(v)g(v)}{(1-\Lambda A(v))^2}\right].$$

I show how the strength of the dependence, as ranked by PQD order, impacts the provider's revenue. To this end, I fix the market size Λ, delay sensitivity c as well as the marginals f_V and f_S. Let (V_1, S_1) and (V_2, S_2) denote two bivariate random variables both in a subset $\mathcal{P}(f_V, f_S)$ of $\mathcal{F}(f_V, f_S)$ whose elements can be ranked by PQD order (see §4.2.1.1). Let R_i denote the optimal revenue, corresponding to a system with joint valuation and service requirement $(V_i, S_i), i = 1, 2$. The next result claims that the optimal revenue decreases with the strength of the dependence.

Proposition 4.11. *If* $(V_1, S_1) \leq_{PQD} (V_2, S_2)$*, then* $R_1 \geq R_2$.

Proposition 4.11 gives the structural result on how different dependencies between customer valuation and service requirement impact the provider's revenue. To explain the intuition, note that if the market is not fully covered, served customers are those with higher valuations, which implies a potential high margin of price. If V and S are positively dependent, joining customers tend to require longer-than-average service times, so that the effective service rate is lower than the nominal service rate and the throughput is lower than it is in the independent model. Since the provider's revenue is determined by the price and throughput, the overall effect of the positive dependence on the revenue is not straightforward. Proposition 4.11 shows that a positive dependence always hurts the provider's revenue, indicating that the loss in throughput due to the positive dependence dominates the potential to charge a high price. Moreover, the numerical study in §4.2.3 shows that the optimal price for a system with a positive dependence can even be lower than that for a system without dependence. To see why, note that the congestion may get worse in the system with a positive dependence because joining customers require longer

service times. The optimal price drops when the increased waiting cost in the dependent model dominates the increase in the valuation threshold of joining customers.

On the other hand, when the dependence is negative, customers with higher valuations stay in service and they tend to require shorter service times. The provider benefits from the negative dependence as she is able to charge a high price while maintain a high throughput.

Proposition 4.11 rationalizes the operational incentives of offering an express line in addition to the regular lines in many service systems. For example, a common practice in grocery stores is to set up self check-out counters or an express line for customers with less than 10 items. These customers require relatively shorter service times and there is a good chance that they will leave the store if they have to wait in regular lines for a long time. A similar practice is adopted in call centers that provide service for commercial banks. A special line is often set up for new card activations, which in general, take less time than other regular services.

Proposition 4.11 also demonstrates how the strength of the dependence affects the provider's revenue. It provides incentives for the service provider to manipulate the dependence. In service settings where customer valuation is positively dependent on the service requirement, a low degree of dependence is desirable. The goal is to relate a customer's valuation to the nature of his service request, rather than the time he spends in service. To achieve this, the provider may choose to provide standard services. In many call centers, the high valuation of a caller is driven by an urgent service request, such as reporting an electricity outrage or a lost card. For these urgent requests, the time a

call center agent spends with the caller may not have a profound impact on the caller's valuation once the caller's request is resolved.

Proposition 4.11 makes a contrast to Anand et al. (2011). Anand et al. (2011) study customer intensive services which also highlight the interaction between the service value and service time. The authors capture the degree of the interaction using a term named as customer intensity, which is defined as the growth factor in the service value with respect to the average service time. In Anand et al. (2011), a high degree of customer intensity benefits the service provider as it allows for a high margin of price. By contrast, my result shows that a high degree of dependence between customer valuation and service requirement, in fact, hurts the provider's revenue. The contrast between the results in my model and in Anand et al. (2011) follows from the difference in the modeling of the dependence between service value and service time. Anand et al. (2011) models the dependence by assuming the service value is a prespecified function of the service speed, with the latter being optimized by the provider. My model, on the other hand, considers a stochastic dependence between the service value and service requirement. The strength of the (stochastic) dependence is measured by PQD order.

Proposition 4.9 shows that the service provider can exploit a large market size in the independent model. I am interested in whether a large market size also benefits the provider's revenue in the presence of a dependence. To gains insights, I assume an MCST and MSCST for the joint distribution of the valuation and service requirement. I find that the benefit of a large market size remains valid when the dependence is negative, as the next proposition shows.

Proposition 4.12. *For a fixed f, if a and b are decreasing, then the provider's optimal revenue R^* is strictly increasing in the market size Λ.*

Note that if I take the CST a and SCST b to be constants, then Proposition 4.12 reduces to Proposition 4.9 in the independent model. The intuition behind Proposition 4.12 can be explained as follows. When V and S are negatively dependent, joining customers tend to require shorter service times. As the market size Λ increases, the valuation threshold also increases. This implies shorter effective service times due to the negative dependence, which further implies a higher price and a higher throughput.

It is worth noting that the benefit of a large market size Λ may not carry over to the case with a positive dependence, where joining customers tend to bring long service requirements. The increase in market size Λ leads to a higher valuation threshold, which implies a longer effective service time. Two opposing effects appear due to the positive dependence: an increase in the valuation threshold and a decrease in the throughput. Which effect is dominant depends on the the system parameters. I will give a numerical example in §4.2.3 to illustrate that the benefit of the a large market size may fail under a positive dependence and certain system parameters. This violation can also be easily observed when the limiting case is considered in which the market size grows indefinitely. Although the optimal revenue need not be monotone with the market size Λ, the limiting result on the optimal revenue as the market size grows still holds, as the next proposition shows. The intuition is similar to that of Proposition 4.10.

Proposition 4.13. *Suppose $\bar{v} < \infty$ and the following limits exist and are strictly positive:*

$$\lim_{v \to \bar{v}} g(v) = \bar{g} > 0 \quad \textit{and} \quad \lim_{v \to \bar{v}} h(v) = \bar{h} > 0.$$

Then the limit of $R^*(\Lambda)$ as $\Lambda \to \infty$ also exists. Furthermore, $\lim_{\Lambda \to \infty} R^*(\Lambda) > 0$ if and only if $\bar{g} < \infty$ and $\bar{h} < \infty$. In this case,

$$\lim_{\Lambda \to \infty} R^*(\Lambda) = \left[\bar{v} - \frac{c\lambda^* \bar{h}}{2(1 - \lambda^* \bar{g})} \right] \lambda^*,$$

where λ^* uniquely solves

$$\bar{v} - \frac{c\lambda \bar{h}}{2(1 - \lambda \bar{g})} - \frac{c\lambda \bar{h}}{2(1 - \lambda \bar{g})^2} = 0.$$

Proposition 4.13 shows that the optimal revenue can be very low if either of $g(v)$ and $h(v)$ becomes very large as $v \to \bar{v}$, which is never the case in the independent model as $g(v)$ and $h(v)$ are constants for all v .

Next I analyze the impact of the delay sensitivity c. As the delay sensitivity decreases, the waiting cost decreases and the provider is able to extract more surplus from joining customers and thus generate more revenue.

Proposition 4.14. *For fixed Λ and f, the provider's optimal revenue R^* is strictly decreasing in the delay sensitivity c.*

Proposition 4.14 shows that the provider's optimal revenue strictly decreases with the delay sensitivity, regardless of the exact form of the dependence in the system. In particular, the provider's optimal revenue R^* decreases in the delay sensitivity c when there is no dependence.

4.2.3. A Numerical Study

In this section, we give a numerical example to compare the provider's optimal revenues across systems with and without dependencies. I let the service value V be uniformly distributed in $[0, 1]$ and the service requirement S be exponentially distributed with rate μ. In models with dependent primitives, we let $V = 1 - e^{-\mu S}$ to capture a positive dependence and $V = e^{-\mu S}$ to capture a negative dependence. The correlation coefficients between V and S in these two cases are the largest and smallest, respectively, among all attainable correlation coefficients for bivariate random variables with the same marginal distributions (See Dhaene et al. (2002)).

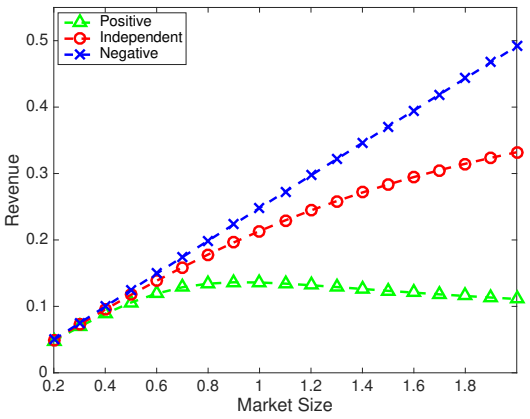
By measuring the time unit by average service time, we can always normalize μ to 1 without loss of generality. I first fix μ and vary the market size Λ to study the effect of varying market sizes. The nominal traffic intensity $\rho := \lambda/\mu$ ranges from 0.2 to 2. I next vary the delay sensitivity $c \in \{0.1, 0.01\}$ to expose the effect of waiting cost.

Proposition 4.8 derives the optimal price in the independent model which uniquely solves (4.12). One can show that the optimal price is also unique in the model with a negative dependence and can be solved through an equation similar to (4.12). However, for the case with a positive dependence, the explicit formula for the optimal price is not attainable and we can only compute it using an exhaustive search. In Figures 4.1 and 4.2, we present results of optimal revenue R and optimal price p for systems with and without dependencies.

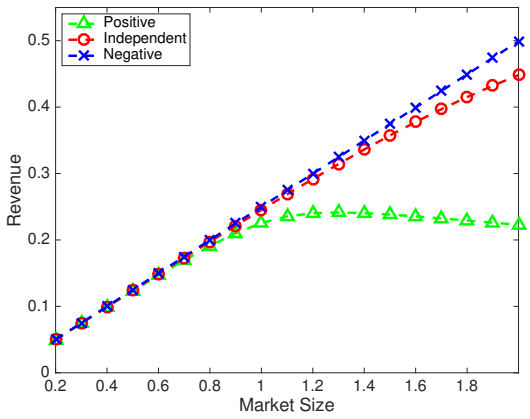
I make two important observations from Figure 4.1. First, as Proposition 4.12 shows, the optimal revenue for systems with a negative dependence or without a dependence increases with the market size. However, such result fails to hold in the presence of a

positive dependence. For example, when the waiting cost is relatively high such that $c = 0.1$, the optimal revenue starts to decrease after the market size exceeds 0.9. Second, Proposition 4.11 predicts a revenue loss due to a positive dependence and a revenue gain due to a negative dependence, compared to the system with no dependence. This numerical example shows that the revenue loss or gain due to a dependence could be substantial when the market size is large. To see why, note that the revenue loss due to the positive dependence follows from the fact that the loss in throughput dominates the potential high margin of price. When the market size goes large, the system becomes more congested and joining customers must have higher service value and longer average service times, implying a further loss in throughput. Since the effect of a reduced throughput dominates, the revenue loss due to the positive dependence exacerbates as the market size grows large.

As we show in Figure 4.2, the relative value of the optimal prices for systems with and without dependencies is sensitive to the system parameters. I find that when the delay sensitivity c is large and market size Λ is small, the optimal price under a positive dependence can be lower than that under no dependence. This implies even the potential of a high margin of price may not be feasible when the dependence is positive. In this case, joining customers with high service values bring in more service requirement, which may lead to a longer waiting time. The effect of waiting cost is amplified when the delay sensitivity c is large and may even dominate the potential of a high margin of price that comes with high-valued joining customers.

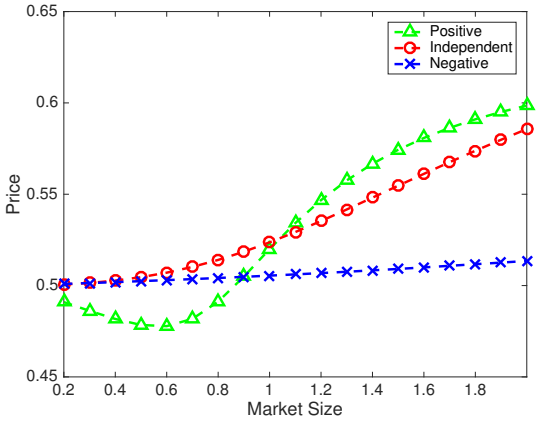


$c = 0.1$

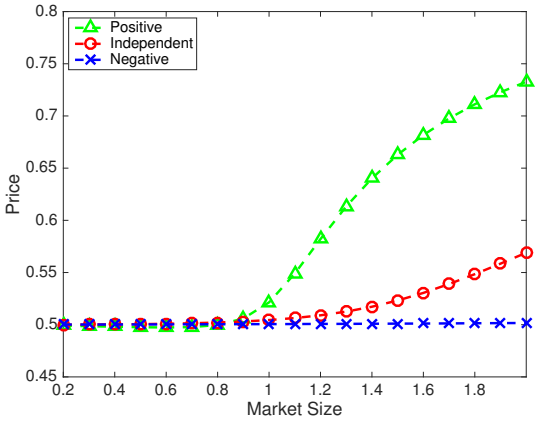


$c = 0.01$

Figure 4.1. Optimal Revenue under Single Pricing



$c = 0.1$



$c = 0.01$

Figure 4.2. Optimal Price under Single Pricing

4.2.4. Service-Based (SB) Pricing

So far we have considered the provider's revenue maximization problem when the same price price is charged to all customers. I have discussed in §4.2.2.2 about the loss of the provider's revenue due to a positive dependence between the service value and service requirement. In this section, we consider how the provider should respond to the positive

dependence to improve his revenue performance. I propose a service-based (SB) pricing scheme as a solution to overcoming the disadvantage of the positive dependence. An SB pricing scheme, loosely speaking, prices a customer's service based on his realized service time. I demonstrate that the SB pricing scheme will only work when the dependence exists. I further show that SB pricing can be used to exploit the dependence and generate more revenue than if the dependence were not there.

The distribution of service value V can be estimated through marketing surveys or lab experiments while the distribution of service time S can be estimated from historic data collected from the service process. The optimal revenue in the independent model represents the provider's predicted revenue when he has estimated the marginals of V and S but treated them as independent primitives. However, estimating the *marginal* distributions of the service value and service requirement is insufficient to determine the optimal prices if the two primitives are in fact dependent. The *joint* distribution of the two primitives should be estimated. Suppose the service provider has estimated the joint distribution f of the two primitives. He could design an SB pricing scheme with the information of the joint distribution f, which we discuss below.

4.2.4.1. Service-based Pricing. I propose a service-based pricing scheme in response to the positive dependence. The literature has demonstrated the role of SB pricing schemes in price discrimination among customers with heterogeneous service requirements; e.g., Mendelson and Whang (1990) and Van Mieghem (2000). In a service system with a positive dependence between the service value and service requirement, customers with higher service values join service and they require longer-than-average service times. The main drawback of single pricing is that each customer is charged the same price even

if that customer requires a very long service time. As a potential fix, the provider may offer an SB pricing scheme which is contingent on each customer's realized service time. Mathematically, an SB pricing function $p : \mathbb{R}+ \mapsto \mathbb{R}_+$, is a mapping that translates a customer's realized service time to a price the customer has to pay. In other words, a customer with a realized service time s will be charged a price $p(s)$ for his service.

I next analyze the how the SB function p affects customers' decisions on whether to join service. The service provider first announces the SB pricing function p. Each customer estimates the price he will have to pay based on the knowledge of his service and then decides whether he should join service. The service-time distribution of each customer can be computed conditioned on that customer's service value v,

$$f_{S|V}(s|v) = \frac{f(s,v)}{f_V(v)}.$$

With the announced SB pricing function $p(\cdot)$, the expected price q charged to a customer with a service value v is

$$q(v) = \mathbb{E}[p(S)|V = v] = \int_0^\infty p(s) f_{S|V}(s|v) ds.$$

Note that the SB pricing scheme can be used to charge different prices across customers only when there exists a dependence between the service value and service requirement. Indeed, if the dependence does not exist, then the conditional service-time distribution $f_{S|V}(s|v) = f_S(s)$ does not depend on v, implying the expected price function q is a constant. Thus, SB pricing is equivalent to single pricing without the dependence. On the other hand, q will vary among customers with different service values when a dependence exists. For example, if the dependence is positive such that $\mathbb{P}(S \geq s|V = v)$ is increasing

in v, one can easily show that if the SB pricing function p is strictly increasing, then so is q.

The provider also announces the expected delay $\hat{w}$, which reflects the provider's belief of the delay time in equilibrium. A customer will join service if his utility of joining service $U(v) = v - q(v) - c\hat{w} \geq 0$. I assume a customer who is indifferent between joining and balking always joins. Define

$$\mathcal{I}(\hat{w}) := \{v : v - q(v) - c\hat{w} \geq 0\},$$

which represents the set of joining customers indexed by their service values. The resulting expected delay is

$$w = \frac{\Lambda \int_{\mathcal{I}(\hat{w})} E[S^2|V = v]\, f_V(v)dv}{2(1 - \Lambda \int_{\mathcal{I}(\hat{w})} E[S|V = v]\, f_V(v)dv)}. \tag{4.13}$$

In equilibrium, the provider's announced delay must be consistent with the actual delay, so that $w = \hat{w}$.

Lemma 4.2. *Given any expected price function $q(\cdot)$, there exists a unique waiting time w.*

I denote the expected waiting time $w := w(q)$ given the uniqueness of q for any expected price function q. Hence, the provider's revenue-optimization problem under SB

pricing scheme is

$$\max_{p(\cdot)} \; R(p) = \int_{\mathcal{I}(w(q))} q(v) f_{S|V}(v) dv \tag{4.14}$$

$$\text{s.t.} \quad q(v) = \int_0^\infty p(s) f_{S|V}(s|v) ds. \tag{4.15}$$

For tractability, we consider a relaxed problem of (4.14). In particular, we will remove the constraint (4.15) and optimize the provider's revenue over q,

$$\max_{q(\cdot)} \; R(q) = \int_{\mathcal{I}(w(q))} q(v) f_{S|V}(v) dv. \tag{4.16}$$

The feasible region of q expands in the relaxed problem (4.16) since we have removed the constraint (4.15) describing how q is generated. In general, for any arbitrary expected price function $q(v)$ and conditional distribution of the service time $f_{S|V}(s|v)$, there is no guarantee that we can find an SB pricing function p such that the constraint (4.15) is satisfied. The optimal solution to the relaxed problem (4.16), thus, is an upper bound of the optimal constrained problem (4.14).

Note that in (4.16), the optimal q squeezes the surplus of joining customers to zero. In other words, the provider is able to use SB pricing as an instrument to extract the full surplus of joining customers. The revenue-optimization problem is equivalent to the welfare-optimization problem in this case. The equivalence between the revenue-optimization and welfare-optimization is widely known for unobservable queues when customers' service values are homogeneous; e.g., Hassin and Haviv (2003, Chapter 3) and the references therein. However, the two optimization problems are, in general, different when customers have heterogeneous service values. Assuming a dependence between the

service value and service requirement, we can reestablish the equivalence between the two problems by allowing for an SB pricing scheme.

To gain insights, we further reduce the feasible region of q to consider threshold policies. I restrict our analysis to the pricing policies such that there exists a service value threshold $\tilde{v}$ and customers elect to join service if their service value is greater than $\tilde{v}$ and balk otherwise. Under threshold policies, the relaxed problem (4.16) can be reduced to an optimization problem which is optimized over the service value threshold $\tilde{v}$, similar to the independent case. Specifically, the expected delay $w(\tilde{v})$ under threshold policies can be computed by

$$w(\tilde{v}) = \frac{\Lambda \int_{\tilde{v}}^{\bar{v}} E[S^2|V=v]\, f_V(v)dv}{2\left(1-\Lambda \int_{\tilde{v}}^{\bar{v}} E[S|V=v]\, f_V(v)dv\right)} = \frac{\Lambda \int_{\tilde{v}}^{\bar{v}} \int_0^{\infty} s^2 f(s,v)dsdv}{2\left(1-\Lambda \int_{\tilde{v}}^{\bar{v}} \int_0^{\infty} s f(s,v)dsdv\right)}. \tag{4.17}$$

The optimal $q^*(v)$ extracts the full surplus of joining customers so that $q^*(v) = v - cw$. Hence,

$$\max_{\tilde{v}} \; R(\tilde{v}) = \int_{\tilde{v}}^{\bar{v}} (v - cw(\tilde{v}))f_V(v)dv. \tag{4.18}$$

Comparing the optimal revenue generated under the SB threshold pricing policies with the independent model, we find that a positive dependence, although undesirable under single pricing, can help to generate more revenue under SB pricing. To state the result, consider two systems with service value and service requirement distributed as (V_1, S_1) and (V_2, S_2), respectively, with fixed marginals. Suppose V_1 and S_1 are independent and V_2 and S_2 are positively dependent characterized by an increasing CST and SCST.

Proposition 4.15. *Under SB pricing, the optimal revenue of the positive dependent system 2 under the relaxed problem* (4.18) *is strictly larger than that of the independent system 1 under the revenue-optimization problem* (4.14) *if either of the two conditions holds:*

(1) When the market size Λ is sufficiently small;

(2) When the delay sensitivity c is sufficiently small and the provider's capacity can cover the entire market, i.e., $\Lambda < \mu$.

SB pricing is useful because it allows for different prices charged to different customers depending on customers' service values. In particular, it can be used to charge a customer a higher price for a longer service time. In the relaxed problem (4.18), it can be further employed to extract the full surplus of joining customers. As a result, Proposition 4.15 shows that when the market size Λ or the delay sensitivity c is sufficiently small, the provider can always do better in the relaxed problem (4.18) compared to the case of no dependence. To complete the analysis of the last component in the primitive data $\mathcal{D}$, the following proposition is concerned with identifying the condition on the strength of dependence to have a similar statement as above.

Proposition 4.16. *Suppose $(V_2, S_2) \in \mathcal{G}(f_V, f_S)$ with correlation coefficient $r_2 > 0$. If r_2 is sufficiently small, then under SB pricing, the optimal revenue of the positive dependent system 2 under the relaxed problem* (4.18) *is strictly larger than that of the independent system 1 under the revenue-optimization problem* (4.14) *.*

The optimal $\tilde{v}^*$ that solves (4.18) gives rise to an expected pricing function $q^*(v) = v - cw(\tilde{v^*})$. Again, we may not find an SB pricing function p which induces q^* via

(4.15) under any arbitrary conditional service-time distribution $f_{S|V}$. However, there always exists an SB pricing function p that induces q^* if we consider the following special dependence between V and S. I say the service value and service requirement are *co-monotonic* (Dhaene et al. (2002)) if $(V, S) \stackrel{d}{=} (F_V^{-1}(U), F_S^{-1}(U))$, where $\stackrel{d}{=}$ stands for equality in distribution and U is a standard uniform random variable independent of V and S.

Corollary 4.2. *Suppose* $(V_1, S_1) \stackrel{d}{=} (F_V^{-1}(U), F_S^{-1}(U))$ *where* U *is a uniform random variable independent of* V_1 *and* S_1. *Under revenue-optimization problem* (4.14) *with SB pricing, the optimal revenue of the positive dependent system 2 is strictly larger than that of the independent system 1 if either of the two conditions holds:*

(1) When the market size Λ *is sufficiently small;*

(2) When the delay sensitivity c *is sufficiently small and the provider's capacity can cover the entire market, i.e.,* $\Lambda < \mu$.

I can explicitly design an SB pricing function in the case when V and S are co-monotonic. Let $\tilde{v}^*$ be the optimal service value threshold solving (4.18) and let $\tilde{s}^* := F_S^{-1}F_V(\tilde{v}^*)$. A feasible SB pricing function is given by

$$p(s) = \begin{cases} F_V^{-1}(F_S(s)) - cw(\tilde{v}^*) & \text{for } s \geq \tilde{s}^* \\ \tilde{v}^* - cw(\tilde{v}^*) & \text{for } s < \tilde{s}^*, \end{cases} \tag{4.19}$$

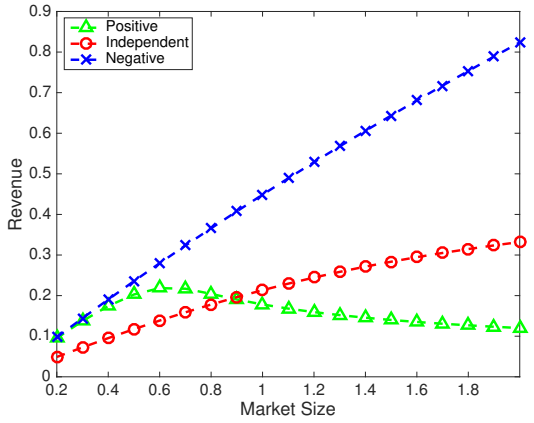
where $w(\tilde{v}^*)$ is the expected waiting time computed by (4.17). To explain the intuition of the pricing function above, note that when V and S are co-monotonic, a customer with service value v has a deterministic service time $F_S^{-1}(F_V(v))$. Under the SB pricing

function defined in (4.19), only customers with service value higher than $\tilde{v}^*$ join service and their surplus is fully extracted by the service provider. Hence, the outcome of the pricing function is consistent with the provider's strategy and customers' decisions described in (4.18).

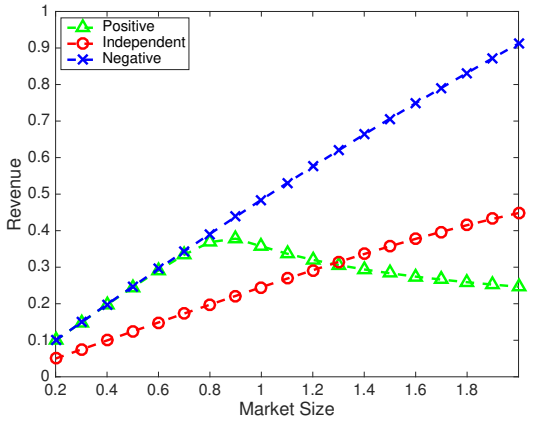
I conduct a numerical study to verify the insights derived from the analytical result. Following the setting in §4.2.3, I let the service value V be uniformly distributed in $[0, 1]$ and the service requirement S be exponentially distributed with rate μ. In models with dependent primitives, I let $V = 1 - e^{-\mu S}$ to capture a positive dependence and $V = e^{-\mu S}$ to capture a negative dependence.

I normalize μ to 1 and vary the market size Λ from 0.2 to 2 with increment 0.02 to study the effect of varying market sizes. I also vary the delay sensitivity c from 0.01 to 0.6 with increment 0.01 to expose the effect of waiting cost. I plot some of the numerical results in Figures 4.3 and 4.4. Across the 5,460 combinations of parameters, the revenue improvement using SB pricing compared to single pricing is 32.0% under a positive dependence and 71.9% under a negative dependence.

Figure 4.3 shows that how the market size shapes the optimal revenue for systems with and without a dependence under SB pricing. When the market size is small, the optimal revenue generated by the SB threshold policies under a positive dependence is larger than that under no dependence. In particular, when the delay sensitivity $c = 0.01$ is small, the optimal revenue under the positive dependence outperforms that under no dependence till the market size grows to approximately 1.2. In this case, if the market can be fully covered by the provider's service capacity, i.e., $\Lambda < \mu$, the positive dependence indeed benefits the provider's revenue performance when SB pricing is allowed. However, when

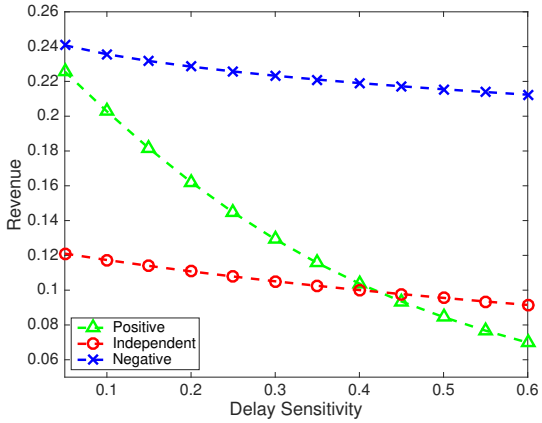


$c = 0.1$

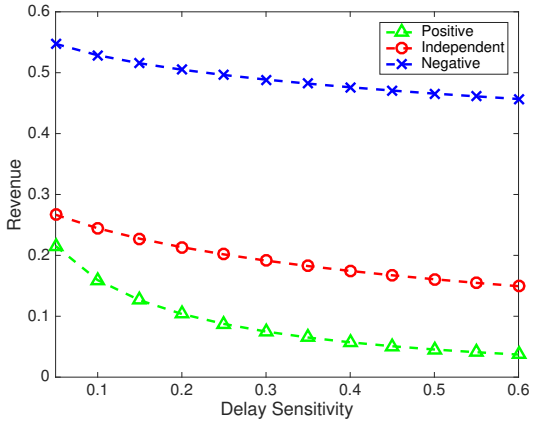


$c = 0.01$

Figure 4.3. Optimal Revenue under SB Threshold Policy



$\Lambda = 0.5$



$\Lambda = 1.2$

Figure 4.4. Optimal Revenue under SB Threshold Policy

the market size is large, the extra revenue extracted from high-valued customers under SB pricing is not sufficient to compensate for a reduced throughput due to a positive dependence. As a result, the optimal revenue under a positive dependence drops below that under no dependence as the market size grows large. Furthermore, we observe that when the market size is small, the difference between the optimal revenue achieved under

a positive and a negative dependence is also relatively small. To explain, note that most of the market can be served when the market size is small, implying a relatively modest difference between the workload and waiting cost in systems with a positive and a negative dependence. Note that the optimal revenue under SB pricing is equal to the social welfare in the relaxed problem (4.18) and that the social welfare is roughly the same when most of the customers join service.

Figure 4.4 shows the impact of delay sensitivity on the optimal revenue when the market size is fixed. I find that when the market size is moderate $\Lambda = 0.5$, the optimal revenue generated by SB pricing under a positive dependence can outperform that under no dependence if the delay sensitivity is smaller than 0.4. By contrast, when the market size is large $\Lambda = 1.2$, even if the delay sensitivity is small $c = 0.05$, the positive dependence can hurt the provider's revenue. I conclude that the benefit of a positive dependence under SB pricing is the most substantial when the market size and delay sensitivity are both small, i.e., when the queueing effect is not very significant.

4.2.4.2. A Heuristic for General SB Pricing. In Propositions 4.15 and 4.16, we compare the optimal revenue of a system with dependence in the relaxed problem (4.18) and a system with no dependence in the original revenue-optimization problem (4.14). I have discussed that the solution to the relaxed problem (4.18) may not always be derived by an SB function via (4.15). Even if the solution to (4.18) can be derived by a SB pricing function, it can be difficult to find that SB pricing function. One of the exceptions is the case with the service value and service requirement being co-monotonic, in which we are able to give an explicit formula for the pricing function. Motivated by the pricing

function in that special case, we propose a heuristic SB pricing function under a general dependence between the service value and service requirement.

The SB pricing formula (4.19) takes advantage of the co-monotonicity between the service value and service requirement. Optimizing over the service value threshold, the provider serves customers with service value higher than that threshold and fully extracts their surplus. Now, with a general *positive* dependence between the service value and service requirement, we revise the pricing function (4.19) and let

$$p(s) = F_V^{-1}(F_S(s)) - c\tilde{w} \quad \text{for all } s, \tag{4.20}$$

where we require $\tilde{w}$ to be an equilibrium expected waiting time that solves (4.13) given the pricing function (4.20). The existence and uniqueness of such $\tilde{w}$ is guaranteed by Lemma 4.2. In general, customers' decisions under the pricing function (4.20) may not exhibit a threshold structure with respect to their service value. Therefore, the expected waiting time must be computed via (4.13) rather than (4.17). The pricing function (4.20) can be interpreted as a two-part tariff. The first term $F_V^{-1}(F_S(s))$ is the regular price a customer pays for his service time s, which is motivated by the case when the service value and service time are co-monotonic. The second term $-c\tilde{w}$ is the price adjustment accounting for the actual waiting time in the system.

I discuss the performance of the pricing function (4.20). For tractability, we consider a special class of dependencies between the service value and service requirement, as characterized by Gaussian copulas. See Appendix A.1.1 for details. In particular, a positive dependence between V and S is captured by assuming $(V, S) \in \mathcal{G}(f_V, f_S)$ with

correlation coefficient $r > 0$. The expected price a customer with service value v pays is

$$q(v) = \int_0^\infty F_V^{-1}(F_S(s)) f_{S|V}(s|v) ds - c\tilde{w}. \tag{4.21}$$

Under the dependence generated by Gaussian copulas, we can show that the expected price $q(v)$ increases with v.

Lemma 4.3. *Assume $(V, S) \in \mathcal{G}(f_V, f_S)$ with correlation coefficient $r > 0$. Then $q(v)$ defined in (4.21) is increasing in v.*

I next characterize customers' decisions under the pricing function (4.20). A customer with service value v joins service if $v - q(v) - c\tilde{w} \geq 0$, or equivalently,

$$v \geq \int_0^\infty F_V^{-1}(F_S(s)) f_{S|V}(s|v) ds.$$

Assuming $(V, S) \in \mathcal{G}(f_V, f_S)$ with V being uniformly distributed, we can show that the pricing function (4.20) indeed induces a threshold structure in customers' decisions.

Lemma 4.4. *Assume $(V, S) \in \mathcal{G}(f_V, f_S)$ with correlation coefficient $r > 0$. Further, if the marginal distribution of V is uniform, then the price function (4.20) induces a threshold policy of customers' decisions. In particular, a customer joins service if and only if his service value $v \geq \mathbb{E}[V]$.*

Combining Lemmas 4.3 and 4.4, we immediately have the following result.

Proposition 4.17. *Assume $(V, S) \in \mathcal{G}(f_V, f_S)$ with correlation coefficient $r > 0$ and V is uniformly distributed in $[0, \bar{v}]$. The revenue generated by the pricing function (4.20) is strictly larger than that in the independent model if either of the two conditions holds:*

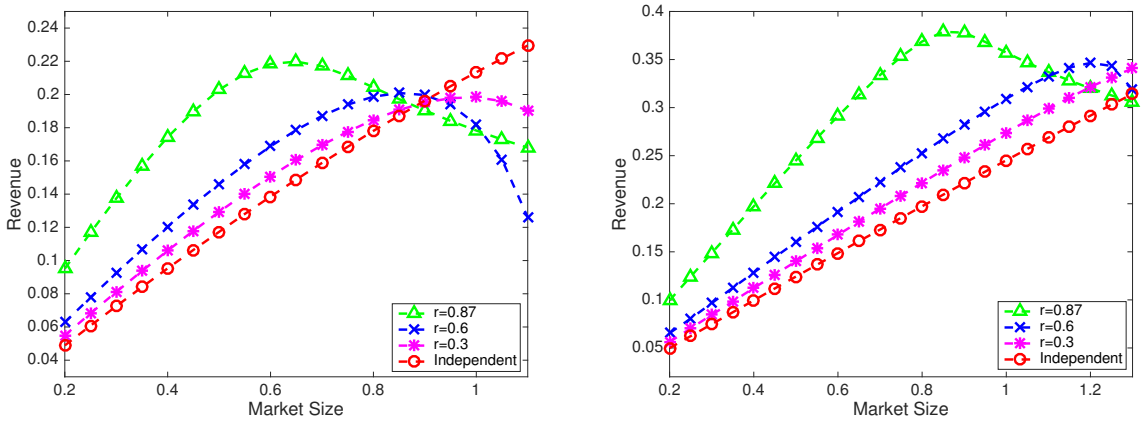
(1) When the market size Λ is sufficiently small;

(2) When the delay sensitivity c is sufficiently small and the provider's capacity can cover the entire market, i.e., $\Lambda < \mu$.

Proposition 4.17 shows that the intuitive construction of price function (4.20) works well when the queueing effect is not significant. The positive dependence can be exploited under the heuristic to generate more revenue compared to the independent model. However, when queueing effect is significant, the comparison result could reverse, as we numerically demonstrate below.

I use a numerical example to test the performance of my proposed heuristic. I let the service value V be uniformly distributed in $[0, 1]$ and the service requirement S be exponentially distributed with rate μ. For systems with dependencies, we generate the joint distributions of V and S using Gaussian copulas. In the system where V and S are co-monotonic, the correlation coefficient is 0.87, which is the largest among all attainable correlation coefficients for the bivariate random variable with the same marginals. To capture a moderate positive dependence, we consider correlation coefficient of V and S to be 0.6 and 0.3 in the other two systems. I compare the revenue for these three systems under the heuristic pricing function (4.20) with that for an independent system under the optimal price. Results with different delay sensitivities are presented in Figure 4.5.

I observe that when the market size is not very large, the heuristic pricing function (4.20) can be used to generate more revenue for a system under a positive dependence than that under no dependence. The numerical results provide strong evidence that a positive dependence can be useful to improve revenue under SB pricing. The results also show that the benefit of SB pricing is the highest in the presence of co-monotonicity between



$c = 0.1$ $c = 0.01$

Figure 4.5. Optimal Revenue under One SB Policy

service value and service requirement, which allows the provider to use SB pricing to fully extract the surplus of joining customers. By contrast, the provider can only extract a partial surplus from customers under a general dependence in which the service value and service requirement are not co-monotonic.

APPENDIX A

Appendix for Chapter 2

A.1. More on Copulas and Conditional Service Time

A.1.1. Generating Gaussian Copulas and t-Copulas

I now provide details on how to generate dependent bivariates (S, T) via Gaussian copula and t-copulas. The procedures I describe below are used to generate different joint distributions which correspond to Figure 2.1.

Let r_G be a number in $[-1, 1]$, and let $\Phi(\cdot)$ denote the cdf of standard normal random variable. The following NORTA procedure, which was proposed in Cario and Nelson (1997), produces a bivariate (S, T) with some correlation coefficient r that is a bijective function of r_G; I elaborate below.

Generating (S, T) Using Gaussian Copula (NORTA)

1. Generate two independent standard normal random variables Z_1 and Z_2.

2. Let $V_1 = Z_1$ and $V_2 = r_G Z_1 + \sqrt{1 - r_G^2} Z_2$. Then V_1 and V_2 are two standard normal random variables with correlation coefficient r_G.

3. Let $S = F_S^{-1}(\Phi(V_1))$ and $T = F_T^{-1}(\Phi(V_2))$. The correlation coefficient r between the random variables S and T generated via the algorithm above is a continuous function of r_G. To generate a bivariate (S, T) with a specific correlation r, I build on the following lemma; see Cario and Nelson (1997) Proposition 2 for its proof and for further details. Let $\underline{r}$ and $\overline{r}$ be the minimal and maximal attainable correlation coefficients of S and

T, respectively. (Note that $\underline{r}$ may be larger than -1 and $\overline{r}$ may be smaller than 1; for example, if S and T are both exponential random variables, then $\underline{r} \approx -0.64$.)

Lemma A.1. *For two densities f_S and f_T and a fixed number $x \in [-1, 1]$, let S_x and T_x be the random variables generated via NORTA by taking $r_G = x$, and let $r(x)$ denote the correlation between S_x and T_x. Then $r : x \mapsto [\underline{r}, \overline{r}]$ is strictly increasing, with $r(-1) = \underline{r}$ and $r(1) = \overline{r}$.*

In particular, the minimal and maximal attainable correlation between two marginal distributions can be generated via NORTA. Moreover, due to the monotonicity of $r(x)$ and its inverse, it is easy to find the value of r_G that gives any pre-specified attainable correlation coefficient. Finally, it can be easily verified that $r(r_G) = 0$ if and only if $r_G = 0$, so that two random variables generated by Gaussian copula are independent if and only if they are uncorrelated.

I next describe the procedure proposed in **?** for generating a bivariate (S, T) using t-copula.

Generating (S, T) Using t-Copula with Degree n

1. Generate two independent standard normal random variables Z_1 and Z_2.

2. Let $V_1 = Z_1$ and $V_2 = r_t Z_1 + \sqrt{1 - r_t^2} Z_2$. Then V_1 and V_2 are two standard normal random variables with correlation coefficient r_t.

3. Generate a random variable Y having the chi-square distribution with n degrees of freedom, and let $U = n/Y$.

4. Let $X_1 = \sqrt{U} V_1$ and $X_2 = \sqrt{U} V_2$.

5. Let $S = F_S^{-1}(t_n(X_1))$ and $T = F_T^{-1}(t_n(X_2))$, where $t_n(\cdot)$ is the cdf of the t-distribution with n degrees of freedom.

A.1.2. Ranking Gaussian Copulas with Given Marginals

Recall that $\mathcal{G} := \mathcal{G}(f_S, f_T)$ denotes the set of joint distributions generated by the Gaussian copula with fixed marginals f_S and f_T. I state a few properties of $\mathcal{G}$. First, a bivariate with any attainable correlation coefficient can be generated by a Gaussian copula, and is characterized by its correlation. In particular, if (S_1, T_1) and (S_2, T_2) are two *distinct* elements in $\mathcal{G}$, then their respective correlation coefficients are necessarily different, i.e., either $r_1 < r_2$, or $r_2 < r_1$, where r_i is the correlation coefficient of (S_i, T_i), $i = 1, 2$. Thus, an important advantage of focusing on the set of bivariates with fixed marginals that are generated by Gaussian copulas, is that the corresponding joint distributions are fully characterized by the correlation coefficient r, so that one parameter can be used as a measure of dependence (as opposed to PQD order, which is a non-parametric measure of dependence). Second, bivariates in the set $\mathcal{G}$ are independent if and only if they are uncorrelated. Third, the class of bivariates generated by the Gaussian copula can be ranked by PQD order, as was mentioned above. I therefore have the following lemma.

Lemma A.2. *If for* $(S_1, T_1), (S_2, T_2) \in \mathcal{G}$ *it holds that* $r_1 < r_2$*, then* $(S_1, T_1) \leq_{PQD} (S_2, T_2)$.

Note that the condition $r_1 < r_2$ is assumed without loss of generality, since the correlation coefficients of any two distinct elements in $\mathcal{G}$ must be strictly ordered.

A.1.3. Relating the CST, PQD order and Gaussian Copula together

The following two lemmas provide natural sufficient conditions for Monotone Conditional Service Time (MCST), and link the monotonicity of the CST to PQD and Gaussian copula.

Lemma A.3. *If $\mathbb{P}(S > u|T = w)$ is increasing in w, then (S,T) is PQD and has an ICST. If $\mathbb{P}(S > u|T = w)$ is decreasing in w, then (S,T) is NQD and has a DCST.*

Lemma A.3 provides a natural sufficient condition for (S,T) to be PQD (NQD) and have an MCST. When $(S,T) \in \mathcal{G}$, both PQD and monotonicity of the CST are determined by the sign of the correlation coefficient r, as the next lemma shows.

Lemma A.4. *Let $(S,T) \in \mathcal{G}$ with correlation coefficient r. Then (i) if $r > 0$, then (S,T) is PQD and has an ICST; (ii) if $r < 0$, then (S,T) is NQD and has a DCST; (iii) if $r = 0$, then (S,T) has a CCST.*

A.2. Time to Stationarity

In general, many-server queueing systems in heavy traffic tend to converge to stationarity much faster than single-server systems; see, e.g., the discussion in E.C.1 in Perry and Whitt (2009). I now demonstrate via simulations that my system with dependence indeed converges quickly to its stationary behavior. I simulate systems with and without dependence, starting the systems at two extreme initial conditions; the systems in Figure A.1(a) are initialized empty, and the initial queue length of the systems depicted in Figure A.1(b) is much larger than the stationary queue. The system parameters and

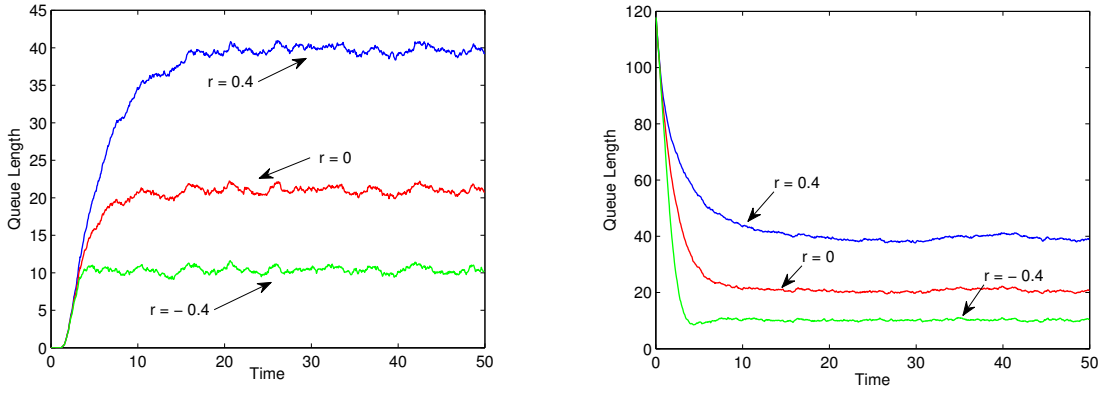
distributions are the same as in the numerical experiment presented in Table 2.1. Specifically, the system has an arrival rate $\lambda = 110$ and $s = 100$ agents, and the service and patience times are exponentially distributed with rate $\mu = 1$ and $\theta = 1/2$, respectively. For each simulated system, I take averages of 500 independent runs, and use the queue length metric to demonstrate the convergence.

Observe that the shape of the trajectories of queues in the dependent models is similar to that of the independent model. Since it is known that both the stochastic system and its fluid limit converge exponentially fast to stationarity in the independent case, I conjecture that the same is true for the dependent model. I further remark that I consider extreme initial conditions in order to make the shape of the trajectories apparent. However, in practice, a stationary analysis is performed over time blocks, with the initial condition of the fluid model being much closer to its stationary point. (Similarly, the initial distribution is much closer to the stationary one, where the distance is measured via an appropriate metric.) Therefore, the actual time it takes to be sufficiently close to stationarity is much shorter than that in the examples shown in Figure A.1.

A.3. Capacity Sizing under a Throughput-Maximizing Policy

In this appendix, I consider the capacity sizing problem when applying the optimal control policy to maximize throughput. The following proposition follows directly from Proposition 3 in Bassamboo and Randhawa (2015).

Proposition A.1. *The throughput-maximizing policy is FIFO if f has a DCST and LIFO if f has an ICST. If f has a CCST, any nonidling policy yields the same throughput.*



(a) Systems initialized empty (b) Systems initialized with extreme large queue

Figure A.1. Convergence of queue length of stochastic system to steady state

In particular, since congestion is beneficial when the dependence is negative, I would like to serve customers in the order at which they arrive, so that customers having short patience, but long service requirements, voluntarily abandon the system. However, under positive dependence, less patient customers are also those who tend to require short services, and since I cannot identify those customers upon arrival, the best I can do is to have $\mu_{\text{eff}} = \mu$. This effective service rate can be achieved (in the fluid model) by employing LIFO, since the waiting of customers who enter service is negligible, and so no screening of customers occurs.

For bivariates generated by Guassian copulas, the conditions on MCST in Proposition A.1 reduce to a condition on the sign of the correlation coefficient.

Corollary A.1. *Let $(S,T) \in \mathcal{G}$. Then the throughput-maximizing policy is FIFO if $r < 0$ and LIFO if $r > 0$. Any nonidling policy yields the same throughput if $r = 0$.*

The discussion in §2.6.1 regarding the optimal capacity under a negative dependence still applies, because FIFO is the optimal policy in this case. Hence I only need to

consider the case with a positive dependence, for which LIFO is optimal. Under LIFO, the throughput is equal to $s\mu$, so that the profit $\Pi_\lambda(s)$ is simply equal to $(p\mu - c)s$, as in the independent model. I conclude that when the throughput-maximizing control policy is adopted, the capacity prescribed in Proposition A.1 remains optimal. Numerical studies for the system considered in §2.6.3, presented in Table A.1, show that the fluid-optimal capacity is fairly accurate under the throughput-maximizing policy.

Table A.1. Optimal staffing under LIFO and positive dependence ($\lambda = 100$)

	$p/c = 1.25$			$p/c = 3.5$		
	Capacity		Cost Gap	Capacity		Cost Gap
r	Optimal	Fluid	Percentage	Optimal	Fluid	Percentage
0	94	100	1.4%	104	100	1.0%
0.2	95	100	0.7%	105	100	1.9%
0.4	95	100	0.5%	105	100	2.8%
0.6	98	100	0.3%	107	100	3.6%
0.8	99	100	0.3%	108	100	4.2%
1	100	100	0.0%	108	100	4.7%

In ending I remark that in overloaded systems, customers will be left to wait with no chance of ever entering service if LIFO is employed, and so is infeasible to employ in observable service systems. Nevertheless, from the fluid perspective, I can achieve the same throughput by employing an admission control policy which rejects arrivals if the number of customers waiting in queue is larger than a certain threshold, and this threshold is negligible for the fluid model.

A.4. Proofs

A.4.1. Auxiliary Results

Before presenting the proofs of the results in this chapter I state two auxiliary results which will be employed in my proofs below.

The proof of the following lemma can be found in Shaked and Shanthikumar (2007, p. 389).

Lemma A.5. *If* $(S_1, T_1) \leq_{PQD} (S_2, T_2)$*, then* $\mathbb{E}(S_1|T_1 \leq z) \geq \mathbb{E}(S_2|T_2 \leq z)$ *and* $\mathbb{E}(S_1|T_1 > z) \leq \mathbb{E}(S_2|T_2 > z)$ *for all* $z \geq 0$*.*

For the next auxiliary result, whose statement follows easily from (2.3), let $w(s)$ denote the steady state offered wait as a function of the capacity s when λ and f are kept fixed. Using the monotonicity of $\phi(\cdot)$, I obtain the following lemma.

Lemma A.6. *I have that* $w(s)$ *is strictly decreasing on* $(0, \lambda/\mu)$*.*

A.4.2. Proofs of the Main Results in Chapter 2

I now prove the main results (Propositions and Corollaries) in the chapter in the order in which they appear.

Proof of Proposition 2.1. Since f_S and f_T are strictly positive over $[0, \infty)$, $\phi(w)$ is strictly decreasing. Thus, there exists a unique solution to (2.3). Q.E.D.

Proof of Proposition 2.2. I start by showing that $\bar{w} > 0$ if and only if $\rho > 0$. First, it follows immediately from the fact that $\phi(w) \leq \phi(0) = \mathbb{E}[S] = \frac{1}{\mu}$ for all $w \geq 0$, that (2.6) is not well defined when $\rho \leq 1$. In particular, there exists no overload equilibrium for the fluid model in this case.

To prove the other direction, I assume that $\rho > 1$ and make the contradictory assumption that $\bar{w} = 0$. It then follows from (2.4) that $a_{\text{eff}} = a(\bar{w}) = \mathbb{E}[S]$, so that $\mu_{\text{eff}} = 1/a_{\text{eff}} = \mu$, contradicting the first equality in (2.7). Thus, it must hold that $\bar{w} > 0$.

I next prove that $\rho > 1$ if and only if $\rho_{\text{eff}} > 1$. To this end, observe that, by (2.5), $\rho_{\text{eff}} \leq 1$ is equivalent to $s\mu_{\text{eff}} \geq \lambda$ which, together with the second equality in (2.7), implies that $\bar{w} = 0$. Hence, by the preceding argument, $\rho > 1$ implies that $\rho_{\text{eff}} > 1$ as well. For the other direction, note that, by (2.5), $\rho_{\text{eff}} > 1$ implies that $\mu_{\text{eff}} < \lambda/s = \rho\mu$. It then follows from the first equality in (2.7) that $\bar{w} > 0$ and thus $\rho > 1$. Q.E.D.

Proof of Proposition 2.3. By Corollary 4.1 in Reich (2012), if $g(w)$ is increasing (decreasing), then $a(w)$ is also increasing (decreasing). The throughput $R(\lambda) = s/a(w(\lambda))$ is increasing (decreasing) in $w(\lambda)$ if $a(w(\lambda))$ is decreasing (increasing) in $w(\lambda)$. By (2.3), given s and f, the offered wait $w(\lambda)$, as a function of λ, is increasing in λ. Thus, $R(\lambda)$ is increasing (decreasing) in λ if a is decreasing (increasing), which is implied by having g decreasing (increasing). Q.E.D.

Proof of Corollary 2.1. Corollary 2.1 follows from Proposition 2.3 and Lemma A.4. Q.E.D.

Proof of Proposition 2.4. The offered wait w solving (2.3) is a function of s, which I denote by $w(s)$. It can be easily verify that $w(s)$ is continuously differentiable in s. Note that $w(s)$ is strictly decreasing in s, so that $w'(s) < 0$. Differentiating both sides of (2.3) with respect to s gives $-\lambda \int_0^\infty x f(x, w(s))dx \cdot w'(s) = 1$, so that

$$-\lambda w'(s) = \left(\int_0^\infty x f(x, w(s))dx \right)^{-1} \tag{A.1}$$

The throughput $R = \lambda F_T^c(w(s))$ is decreasing in $w(s)$, and hence increasing in s. Taking the derivative of $R(s)$, $R'(s) = -\lambda f_T(w(s))w'(s)$, and plugging the value of $-\lambda w'(s)$ in (A.1), gives

$$R'(s) = \frac{f_T(w(s))}{\int_0^\infty x f(x, w(s))dx} = \frac{1}{\mathbb{E}(S|T = w(s))} = \frac{1}{g(w(s))}.$$

Therefore $R'(s) > 0$ for all $s \in (0, \lambda/\mu)$. If g is increasing, then $R'(s)$ is increasing in s, hence $R(s)$ is convex in s. Analogously, $R(s)$ is concave in s if g is decreasing. Q.E.D.

Proof of Cororllary 2.2. Corollary 2.2 follows from Proposition 2.4 and Lemma A.4. Q.E.D.

Proof of Proposition 2.5. It suffices to prove that $w_1 \leq w_2$, because the stated inequalities for R_i and Q_i, $i = 1, 2$, will follow immediately from (2.8) and (2.9) and the fact that T_1 and T_2 have the same marginal cdf F_T. To this end, I will prove that the following inequality holds for ϕ in (2.2).

$$\phi_1(z) \leq \phi_2(z), \quad \text{for all } z \geq 0. \tag{A.2}$$

Indeed, if (A.2) holds, then $\rho\phi_1(w_2) \leq \rho\phi_2(w_2) = 1/\mu$. Since $\rho\phi_i(w_i) = 1/\mu$ for $i = 1, 2$, and since ϕ_1 is strictly decreasing and $\rho\phi_1(w_1) = 1/\mu$, (A.2) implies that $w_1 \leq w_2$.

It remains to show (A.2) holds. Note that $\psi_z(s, t) = (s \cdot \mathbf{1}\{t > z\})$ is a supermodular function in (s, t). It also holds that $\phi_i(z) = \mathbb{E}[\psi_z(S_i, T_i)]$. Since PQD ordering and supermodular ordering are equivalent in the bivariate case (see 9.A.18, Shaked and Shanthikumar (2007, p. 395)), $(S_1, T_1) \leq_{PQD} (S_2, T_2)$ implies $\mathbb{E}[\psi_z(S_1, T_1)] \leq \mathbb{E}[\psi_z(S_2, T_2)]$,

i.e., $\phi_1(z) \leq \phi_2(z)$. Q.E.D.

Proof of Corollary 2.3. The statement of the corollary follows from the fact that the inequality in (A.2) is strict for $(S_1, T_1), (S_2, T_2) \in \mathcal{G}(f_S, f_T)$ with $r_1 < r_2$. To show this, note that

$$\text{(A.3)} \qquad \mathbb{E}(S_i|T_i > z) = \int_0^\infty \mathbb{P}(S_i > u|T_i > z)du = \frac{\int_0^\infty \mathbb{P}(S_i > u, T_i > z)du}{F_T^c(z)}.$$

In the proof of Lemma A.2 below I show that $\mathbb{P}(S_1 > u, T_1 > z) < \mathbb{P}(S_2 > u, T_2 > z)$ for all $u, z > 0$. It then follows from (A.3) that $\mathbb{E}(S_1|T_1 > z) < \mathbb{E}(S_2|T_2 > z)$ for all $z > 0$. Hence, $w_1 < w_2$, implying that $R_1 > R_2$ and $Q_1 < Q_2$. Q.E.D.

Proof of Proposition 2.6. If g is increasing, then by Proposition 2.4, $R(s)$ is convex increasing in s. Hence, the profit function $\Pi_\lambda(s)$ is convex in s and maximizing $\Pi_\lambda(s)$ gives a corner solution. Note that $\Pi_\lambda(0) = 0$ and $\Pi_\lambda(\lambda/\mu) = (p\mu - c)\lambda/\mu$. Hence $\Pi_\lambda(\lambda/\mu) > 0$ if and only if $p\mu > c$. In other words, $s^*_\lambda = \lambda/\mu$ is optimal if and only if $p\mu > c$.

Next, if g is decreasing, I optimize the cost function $\bar{C}_\lambda(s)$ in (2.12). Or equivalently, I minimize $\bar{C}_\lambda(w)$ in (2.13). The derivative of $\bar{C}_\lambda(w)$ is $\bar{C}'_\lambda(w) = f_T(w)\,(p - cg(w))$, since g is decreasing, $\bar{C}_\lambda(w)$ is quasiconvex in w. Hence, any local minimizer is globally optimal. If $g(\infty) < p/c < g(0)$, since g is continuous, there is a unique w^* that solves $g(w^*) = p/c$ and w^* is the optimizer. The optimal capacity is given by (2.7): $s^*_\lambda = \lambda F_T^c(w^*)a(w^*)$. If $p/c \geq g(0)$, then $\bar{C}'_\lambda(w) \geq 0$ for all w, so that $w^* = 0$ and $s^*_\lambda = \lambda/\mu$ is fluid optimal. If $p/c \leq g(\infty)$, then $\bar{C}'_\lambda(w) \leq 0$ for all w, hence $w^* = \infty$ and $s^*_\lambda = 0$. Q.E.D.

Proof of Corollary 2.4. The cost function for a system whose service time and patience time are distributed as S_i and T_i is

$$C_i(s) = cs + p\alpha_i(s) = cs + p(\lambda - R_i(s)) = p\lambda + cs - pR_i(s).$$

Let s_i^* be the optimal capacity for a syetem with service time and patience time (S_i, T_i). Then

$$C_2^* = p\lambda + cs_2^* - pR_2(s_2^*) \geq p\lambda + cs_2^* - pR_1(s_2^*) \geq p\lambda + cs_1^* - pR_1(s_1^*) = C_1^*,$$

where the first inequality follows from Proposition 2.5 and the second inequality follows from the optimality of s_1^* for a system with service and patience time (S_1, T_1). Q.E.D.

Proof of Proposition 2.7. The first-order condition (2.13) of the capacity optimization problem gives

$$\mathbb{E}(S_1|T_1 = w_1(s_1^*)) = \mathbb{E}(S_2|T_2 = w_2(s_2^*)) = p/c,$$

where $w_i(s)$ is the offered wait for a system with capacity s and service and patience time (S_i, T_i). I will next show that

$$g_1(0) := \mathbb{E}(S_1|T_1 = 0) \geq \mathbb{E}(S_2|T_2 = 0) =: g_2(0). \tag{A.4}$$

To prove (A.4), I take the contradictory assumption that $\mathbb{E}(S_1|T_1 = 0) < \mathbb{E}(S_2|T_2 = 0)$. As I assume continuity of g_i for $i = 1, 2$, I can therefore find a $\delta > 0$, such that $\mathbb{E}(S_1|T_1 =$

$z) < \mathbb{E}(S_2|T_2 = z)$ for all $z < \delta$. Note that

$$\mathbb{E}(S_i|T_i \leq z) = \frac{\int_0^z \mathbb{E}(S_i|T_i = t) f_T(t)dt}{F_T(z)}.$$

Since T_1 and T_2 have the same marginal cdf F_T, $\mathbb{E}(S_1|T_1 \leq \delta) < \mathbb{E}(S_2|T_2 \leq \delta)$, contradicting Lemma A.5. Hence, (A.4) must hold.

I will show below that there exists a t_0, $0 < t_0 < \infty$, such that for all $z \in [0, t_0]$, it holds that $g_1(z) \geq g_2(z)$. For that t_0, let $M := g_2(t_0)$. Since g_2 is strictly decreasing, $M < g_2(0) \leq g_1(0)$. If $M < p/c < g_2(0) \leq g_1(0)$, then the equality $g_1(w_1(s_1^*)) = g_2(w_2(s_2^*)) = p/c$ and the fact that g_1 and g_2 are both strictly decreasing functions imply that

$$w_1(s_1^*) \geq w_2(s_2^*). \tag{A.5}$$

Observe that the inequality $s_1^* > s_2^*$ implies that $w_1(s_1^*) < w_1(s_2^*) \leq w_2(s_2^*)$, where the first inequality follows from Lemma A.6 and the second inequality follows from Proposition 2.5, contradicting (A.5). Hence, it must hold that $s_1^* \leq s_2^*$ as stated.

It remains to show the existence of a finite $t_0 > 0$, such that $g_1(z) \geq g_2(z)$ for all $z \in [0, t_0]$. To this end, I consider the the case $h(0) > 0$ and $h(0) = 0$ separately. Assume first that $h(0) > 0$. In this case, $g_1(0) > g_2(0)$ so that $g_1(z) > g_2(z)$ in a right neighborhood of 0 due to the right continuity of g_1 and g_2 at 0. Define $t_0 := \inf_{z \geq 0}\{g_1(z) \leq g_2(z)\}$. Note that $t_0 < \infty$ because $\int_0^\infty f_T(y)g_1(y)dy = \int_0^\infty f_T(y)g_2(y)dy = \mathbb{E}[S]$. (If $g_1(z) > g_2(z)$ for all $z \geq 0$, then this latter equality cannot hold.)

I next consider the case $h(0) = 0$. If $h(t) = 0$ for all t in some right neighborhood of 0, namely, if there exists $t_0 > 0$ such that $h(z) = 0$ for all $z \in [0, t_0]$, then it trivially

holds that $g_1(z) \geq g_2(z)$ for all $z \in [0, t_0]$. Hence, I need only consider the case in which $h(0) = 0$ and h is not identically equal to 0 in any right neighborhood of 0. That is, for any $\epsilon > 0$, there exists $t \in (0, \epsilon)$ such that $h(t) \neq 0$. Define $t_0 = \inf\{z > 0 : h(z) = 0\}$, where $\inf(\emptyset) := \infty$. I first claim that $t_0 > 0$. Indeed, if $t_0 = 0$, then there must exists a positive sequence $\{z_n : n \geq 1\}$ such that $h(z_n) = 0$ and $z_n \to 0$ as $n \to \infty$, contradicting the assumption that the principle of permanence holds for h at $z = 0$. I therefore have $t_0 > 0$. I next show t_0 is finite and that $h(z) \geq 0$ for all $z \in [0, t_0]$, so that $g_1(z) \geq g_2(z)$ for all $z \in [0, t_0]$. If $t_0 = \infty$, note that by the definition of t_0, it holds that $h(z) > 0$ or $h(z) < 0$ for all $z > 0$. (Otherwise, if the value of h changes sign in $(0, t_0)$, then the continuity of h implies that there exists a $\hat{z} \in (0, t_0)$ such that $h(\hat{z}) = 0$, contradicting the definition of t_0.) Since $h(z) < 0$ for all $z > 0$ implies that $\mathbb{E}(S_1|T_1 \leq \delta) < \mathbb{E}(S_2|T_2 \leq \delta)$ for all $\delta > 0$, a contradiction to Lemma A.5, I necessarily have $h(z) > 0$ for all $z \in (0, t_0)$. But then $g_1(z) > g_2(z)$ for all $z > 0$ which, as was shown above for the case $h(0) > 0$, contradicts the fact that $\mathbb{E}[S_1] = \mathbb{E}[S_2]$. Hence, it must hold that $t_0 < \infty$. Repeating the same argument above shows that $h(z) > 0$ for all $z \in (0, t_0)$, so that $h(z) \geq 0$ for all $z \in [0, t_0]$.Q.E.D.

Proof of Corollary 2.5. If $(S_1, T_1), (S_2, T_2) \in \mathcal{G}(f_S, f_T)$ satisfying $r_1 < r_2 < 0$, then $g_1(z) > g_2(z)$ for sufficiently small $z > 0$. Define $t_0 := \inf_z\{g_1(z) \leq g_2(z)\}$, then $t_0 > 0$. A similar argument to the one in the proof of Proposition 2.7 gives $t_0 < \infty$. For all $z \in (0, t_0)$, I have $g_1(z) > g_2(z)$. Define $M := g_2(t_0)$. If $p/c > M$, then the first order condition $g_1(w_1(s_1^*)) = g_2(w_2(s_2^*)) = p/c$ implies $w_1(s_1^*) > w(s_2^*)$. A similar argument to the one in the proof of Proposition 2.7 can be used to show that $s_1^* < s_2^*$. Q.E.D.

A.4.3. Proofs of the Lemmas in the Paper

Proof of Lemma A.2. As demonstrated in Appendix A.1.1,

$$(S_i, T_i) \stackrel{\mathrm{d}}{=} (F_S^{-1}(\Phi(\Gamma_i)), F_T^{-1}(\Phi(\Xi_i))),\ i = 1, 2,$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution and (Γ_i, Ξ_i) is a bivariate normal random variable with correlation coefficient r_G^i. It follows from Lemma A.1 that $r_1 < r_2$ if and only if $r_G^1 < r_G^2$. Therefore, it suffices to show that if $r_G^1 \leq r_G^2$, then $(S_1, T_1) \leq_{PQD} (S_2, T_2)$. By Proposition 9.A.1 of Shaked and Shanthikumar (2007, p. 390), PQD ordering is preserved under componentwise increasing transformation of random vectors. Since $F_S^{-1}(\Phi(\cdot))$ and $F_T^{-1}(\Phi(\cdot))$ are both increasing, it suffices to show that if $r_G^1 \leq r_G^2$, then $(\Gamma_1, \Xi_1) \leq_{PQD} (\Gamma_2, \Xi_2)$. This latter result follows from the facts that (1) bivariate normal distributions with the same marginals are monotone in the association ordering with respect to their correlation coefficient (Shaked and Shanthikumar (2007, p. 419, Example 9.E.6)); (2) association ordering implies PQD ordering (Shaked and Shanthikumar (2007, p. 417, Proposition 9.E.2)).

I now prove a stronger version of the lemma, which I employ in the proof of Corollary 2.3, requiring a strict form of the PQD order; in particular, I prove that, if $r_1 < r_2$, then $\mathbb{P}(S_1 \leq x, T_1 \leq y) < \mathbb{P}(S_2 \leq x, T_2 \leq y)$, for all $x, y > 0$. With an abuse of notation, I write $F \in \mathcal{G} := \mathcal{G}(f_S, f_T)$ if F is the joint cdf of a bivariate $(S, T) \in \mathcal{G}$. Since a bivariate normal random variable is completely characterized by its mean and correlation coefficient r_G, it follows from Lemma A.1 that the cdf's $\{F \in \mathcal{G}\}$ can be indexed by the correlation coefficient r of (S, T). Moreover, again by Lemma A.1, there exists a bijection mapping

from the cdf's in $\mathcal{G}$ to the family of bivariate normal random variables with a zero mean vector indexed by their correlation coefficient r_G. Thus, I can equivalently parameterize the elements $\{F \in \mathcal{G}\}$ by the correlation coefficient r_G of the underlying bivariate normal random variables, and show that $\{F_{r_G}(x_1, y_1) : -1 \leq r_G \leq 1\} \equiv \{F_r(x_1, y_1) : \underline{r} \leq r \leq \overline{r}\}$ is increasing in r_G, and thus in r, for all $x_1, y_1 \geq 0$. (Recall that $\underline{r}$ and $\overline{r}$ denote the minimal and maximal attainable correlation coefficients of S and T, respectively; see Appendix A.1.1.)

Let φ_{r_G} denote the density function of (Γ, Ξ) with correlation coefficient r_G:

$$\varphi_{r_G}(u_1, u_2) = \frac{1}{2\pi\sqrt{1 - r_G^2}} \exp\left(-\frac{u_1^2 - 2r_G u_1 u_2 + u_2^2}{2(1 - r_G^2)}\right).$$

For Φ and ϕ denoting the cdf and pdf of the standard normal random variable, respectively, let $\gamma(x) := \Phi^{-1}(F_S(x))$ and $\xi(y) := \Phi^{-1}(F_T(y))$. Then $\Gamma \overset{\mathrm{d}}{=} \gamma(S)$ and $\Xi \overset{\mathrm{d}}{=} \xi(T)$ so that the joint density of (S, T) is

$$f_{r_G}(x, y) := \frac{1}{2\pi\sqrt{1 - r_G^2}} \exp\left(-\frac{\gamma(x)^2 - 2r_G\gamma(x)\xi(y) + \xi(y)^2}{2(1 - r_G^2)}\right) \gamma'(x)\xi'(y),$$

where $\gamma'(x) = f_S(x)/\phi(\gamma(x))$ and $\xi'(y) = f_T(y)/\phi(\xi(y))$. Then

$$\begin{aligned} F_{r_G}(x_1, y_1) &= \int_0^{y_1} \int_0^{x_1} f_{r_G}(x, y) dx dy \\ &= \int_0^{y_1} \int_0^{x_1} \frac{1}{2\pi\sqrt{1 - r_G^2}} \exp\left(-\frac{\gamma(x)^2 - 2r_G\gamma(x)\xi(y) + \xi(y)^2}{2(1 - r_G^2)}\right) \gamma'(x)\xi'(y) dx dy, \end{aligned}$$

so that

$$
\begin{aligned}
\frac{\partial F_{r_G}(x_1,y_1)}{\partial r_G} =& \int_0^{y_1}\int_0^{x_1} \frac{\partial\left(\frac{1}{2\pi\sqrt{1-r_G^2}}\exp\left(-\frac{\gamma(x)^2-2r_G\gamma(x)\xi(y)+\xi(y)^2}{2(1-r_G^2)}\right)\gamma'(x)\xi'(y)\right)}{\partial r_G}dxdy \\
=& \frac{1}{2\pi}\int_0^{y_1}\int_0^{x_1} e^{-\frac{\xi(y)^2}{2}}\frac{\partial\left(\exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)}\right)\gamma'(x)\xi'(y)\right)}{\partial r_G}dxdy \\
=& \frac{1}{2\pi}\int_0^{y_1}\int_0^{x_1} e^{-\frac{\xi(y)^2}{2}}\Bigg\{\frac{\frac{\partial\left(\exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)}\right)\gamma'(x)\xi'(y)\right)}{\partial r_G}\sqrt{1-r_G^2}}{1-r_G^2} \\
&+\frac{\left(\exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)}\right)\gamma'(x)\xi'(y)\right)\frac{2r_G}{2\sqrt{1-r_G^2}}}{1-r_G^2}\Bigg\}dxdy \\
=& \frac{1}{2\pi\sqrt{1-r_G^2}}\int_0^{y_1}\int_0^{x_1} e^{-\frac{\xi(y)^2}{2}}\xi'(y)\Big\{\gamma'(x)\frac{\partial\left(\exp(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)})\right)}{\partial r_G} \\
&+\gamma'(x)\frac{\left(\exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)}\right)\right)r_G}{1-r_G^2}\Big\}dxdy. \qquad \text{(A.6)}
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\int_0^{x_1} \gamma'(x) \frac{\partial \left(\exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)} \right) \right)}{\partial r_G} dx \\
=& -\int_0^{x_1} \gamma'(x) \exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)} \right) \frac{r_G\left[\gamma(x)-r_G\xi(y)\right]\left[\gamma(x)-\frac{\xi(y)}{r_G}\right]}{(1-r_G^2)^2} dx \\
=& -\frac{1}{(1-r_G^2)^2} \int_0^{x_1} \exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)} \right) r_G \left[\gamma(x)-\frac{\xi(y)}{r_G}\right] d\left[\gamma(x)-r_G\xi(y)\right]^2 \\
=& \frac{1}{1-r_G^2} \Big\{ r_G \left[\gamma(x)-\frac{\xi(y)}{r_G}\right] \exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)} \right) \Big|_{x=0}^{x_1} \\
&- \int_0^{x_1} r_G\gamma'(x) \exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)} \right) dx \Big\} \\
=& \frac{1}{1-r_G^2} \Big\{ r_G \left[\gamma(x_1)-\frac{\xi(y)}{r_G}\right] \exp\left(-\frac{(\gamma(x_1)-r_G\xi(y))^2}{2(1-r_G^2)} \right) \\
&- \int_0^{x_1} r_G\gamma'(x) \exp\left(-\frac{(\gamma(x)-r_G\xi(y))^2}{2(1-r_G^2)} \right) dx \Big\}. \qquad \text{(A.7)}
\end{aligned}
$$

Plug (A.7) into (A.6),

$$
\begin{aligned}
&\frac{\partial F_{r_G}(x_1,y_1)}{\partial r_G} \\
=& \frac{1}{2\pi(1-r_G^2)^{3/2}} \int_0^{y_1} r_G e^{-\frac{\xi(y)^2}{2}} \xi'(y) \left[\gamma(x_1)-\frac{\xi(y)}{r_G}\right] \exp\left(-\frac{(\gamma(x_1)-r_G\xi(y))^2}{2(1-r_G^2)} \right) dy \\
=& -\frac{1}{2\pi(1-r_G^2)^{3/2}} \int_0^{y_1} e^{-\frac{\gamma(x_1)^2}{2}} \xi'(y)(\xi(y)-r_G\gamma(x_1)) \exp\left(-\frac{(\xi(y)-r_G\gamma(x_1))^2}{2(1-r_G^2)} \right) dy \\
=& \frac{1}{2\pi(1-r_G^2)^{1/2}} \Big[e^{-\frac{\gamma(x_1)^2}{2}} \exp\left(-\frac{(\xi(y)-r_G\gamma(x_1))^2}{2(1-r_G^2)} \right) \Big] \Big|_{y=0}^{y_1} \\
=& \frac{1}{2\pi(1-r_G^2)^{1/2}} \Big[e^{-\frac{\gamma(x_1)^2}{2}} \exp\left(-\frac{(\xi(y_1)-r_G\gamma(x_1))^2}{2(1-r_G^2)} \right) \Big].
\end{aligned}
$$

It follows that $\frac{\partial F_{r_G}(x_1,y_1)}{\partial r_G} > 0$ for all $x_1, y_1 > 0$, implying the statement of the lemma. Q.E.D.

Proof of Lemma A.3. Note that $g(w) = \mathbb{E}[S|T = w] = \int_0^\infty \mathbb{P}(S > u|T = w)du$. Since $\mathbb{P}(S > u|T = w)$ is increasing in w and f_S is fixed, g is necessarily increasing. It remains to show that $\mathbb{P}(S > u|T = w)$ increasing in w implies PQD. Following Block et al. (1985), I say that (S, T) is Positively Dependent through Stochastic Ordering (PDS) if $\mathbb{P}(S > u|T = w)$ is increasing in w. That PDS implies PQD is proved in Block et al. (1985, p.82). Q.E.D.

Proof of Lemma A.4. By Lemma A.3, I need to show that, if $(S, T) \in \mathcal{G}$, then $\mathbb{P}(S > u|T = w)$ is strictly increasing in w (PDS) (see the proof of Lemma A.3) if $r > 0$, and strictly decreasing in w if $r < 0$. Note that if $(S, T) \in \mathcal{G}(f_S, f_T)$, then $(S, T) \stackrel{d}{=} (F_S^{-1}(\Phi(\Gamma)), F_T^{-1}(\Phi(\Xi)))$ for a bivariate normal random variables (Γ, Ξ) with correlation coefficient r_G. By Block et al. (1985, Proposition 2.1), PDS is preserved under componentwise increasing transformation of random vectors. Since $r > 0$ implies $r_G > 0$ by Lemma A.1, and since $F_S^{-1}(\Phi(\cdot))$ and $F_T^{-1}(\Phi(\cdot))$ are both increasing, it suffices to show that (Γ, Ξ) is PDS if $r_G > 0$. This latter result is established in Block et al. (1985, Example 4.1). The proof for $r < 0$ is similar. Q.E.D.

APPENDIX B

Appendix for Chapter 3

B.1. An Algorithm to Solve the Fluid Model

In this section I provider an algorithm to numerically solve the continuous-time fluid model developed in §3.3.2 by computing the various performance functions in discrete time scales. Building on the discretization method proposed in Whitt (2006a), I adapt the method to incorporate the dependence in my model, and also complement and correct some implementation details in Whitt (2006a).

I assume that all events take place in the discrete time scale $\{i\delta : i \geq 0\}$ for small $\delta > 0$. I specify the order of events if more than one event takes place simultaneously: First, customers who complete service depart, second, waiting customers in queue enter service, third, impatient customers in queue elect to abandon, finally, new arrivals are added to the system. When counting the queue length $Q(n)$, I don't include $q(n, 0)$, which is the new arrival at time n. Define the probability mass functions (pmf) for customer l:

$$g(i|j) = \mathbb{P}(S_l = i\delta | T_l > Z_l, Z_l = j\delta) \approx \psi(j\delta, i\delta)\delta,$$

$$G(i|j) = \sum_{k=1}^{i} g(k|j),\ G^c(i|j) = 1 - G(i|j),$$

$$g_T(i) = \mathbb{P}(T_l = i\delta) \approx f_T(i\delta)\delta,\ G_T(i) = \sum_{k=1}^{i} g_T(k),\ G_T^c(i) = 1 - G_T(i),$$

where ψ and f_T are the pdf of the conditional service time and patience time in the continuous model, respectively. I discuss how to discretize the fluid model in two cases depending on whether all servers are currently busy.

Case 1: $B(n-1) = 1$.

If $B(n-1) = 1$, then all servers are busy after time $n-1$. In this case, the total service rate

$$\begin{aligned}\sigma(n) =& \sum_{i=1}^{n-1} \Bigg\{ I(n-i-1) \Bigg[\sum_{j=c_{n-i}+1}^{n-i-1} q(n-i-1,j) g(i|j) \\ &+ \Bigg(b(n-i,0) - \sum_{j=c_{n-i}+1}^{n-i-1} q(n-i-1,j) \Bigg) g(i|c_{n-i}) \Bigg] \\ &+ (1 - I(n-i-1)) \left[\sum_{j=0}^{n-i-1} q(n-i-1,j) g(i|j) \right] \Bigg\} + \sum_{i=n}^{\infty} b(0, i-n) g(n|w_B(i-n)),\end{aligned} \tag{B.1}$$

where $I(n)$ and c_n are defined in (B.5) and (B.6) below. I use c_n to denote the smallest waiting time experienced by the fluid that enters service at time n, which approximates $w(n\delta)$ in the continuous model. $I(n)$ is an indicator function which tracks whether all the queue content at time $n-1$ and new arrival $q(n-1,0)$ are cleared by service input at time n. If the queue and new arrival is not cleared at time n, then $I(n) = 1$ and c_n defined in (B.6) below is strictly positive.

To explain the expression of the total service rate in (B.1), first pick an arbitrary $i < n$ and consider the fluid content that enters service at time $n-i$ and remains in service at time $n-1$. This fluid may consist of separate fluid densities with different waiting times. Hence I compute the service rate of this fluid by summing up the conditional

service rates of its individual fluid, conditioned on the corresponding waiting times. The quantity of the individual fluid with waiting time j that remains in service at time $n-1$ is $q(n-i-1,j)G^c(i-1|j)$ and the conditional hazard rate of this individual fluid is $g(i|j)/G^c(i-1/j)$. It then follows that the instantaneous service rate of this individual fluid is $q(n-i-1,j)g(i|j)$, which is the product of the quantity of the individual fluid at time $n-i$ and the conditional pmf conditioned on its waiting time. Now I consider the waiting times of the fluid entering service at time $n-i$. If $I(n-i)=1$, then the smallest waiting time of individual fluids is c_{n-i}, and the quantity of fluid with the smallest waiting time c_{n-i} is $b(n-i,0)-\sum_{j=c_{n-i}+1}^{n-i} q(n-i-1,j)$. If $I(n-i)=0$, then the smallest waiting time of individual fluids is zero. The second term in (B.1) is the service rate of the initial fluid that remains in service at time n. A similar argument gives the fluid density $b(n,i)$ in service:

$$
\text{(B.2)} \qquad b(n,i) = \begin{cases} I(n-i-1)\Big[\sum_{j=c_{n-i}+1}^{n-i-1} q(n-i-1,j)G^c(i|j) \\ \quad + \Big(b(n-i,0) - \sum_{j=c_{n-i}+1}^{n-i-1} q(n-i-1,j)\Big) G^c(i|c_{n-i})\Big] \\ \quad + (1-I(n-i-1))\Big[\sum_{j=0}^{n-i-1} q(n-i-1,j)G^c(i|j)\Big] & \text{for } i<n \\ b(0,i-n)G^c(n|w_B(i-n)) & \text{for } i \geq n. \end{cases}
$$

I give the boundary conditions when there is a positive queue. (I point out that Equation (6.17) in Whitt (2006a) is incorrect.)

$$
\text{(B.3)} \qquad \begin{aligned} b(n,0) &= \min\{\sigma(n)\delta, Q(n-1)+q(n-1,0)\}, \\ q(n,0) &= \lambda\delta. \end{aligned}
$$

Update fluid content in service B in the next time epoch:

$$B(n) = \sum_{i=0}^{n} b(n,i). \tag{B.4}$$

Note in (B.3), the service input $b(n,0)$ depends on the relative value of $\sigma(n)\delta$ and $Q(n-1)+q(n-1,0)$. I define an indicator function to track which is smaller. (This detail complements Whitt (2006a).)

$$I(n) = \mathbf{1}_{[\sigma(n)<Q(n-1)+q(n-1,0)]} \quad \text{for } n \geq 0. \tag{B.5}$$

Using c_n to denote the smallest waiting time of fluid that enters service at time n, I compute c_n in two cases depending on $I(n)$.

(1) $\sigma(n) < Q(n-1) + q(n-1,0)$.

In this case $I(n) = 1$ and $b(n,0) = \sigma(n)$. Further,

$$\begin{aligned}
q(n,i) &= 0, \quad i \geq c_n + 2, \\
q(n,c_n+1) &= (1-P_n)q(n-1,c_n)\frac{G_T^c(c_n+1)}{G_T^c(c_n)}, \\
q(n,i) &= q(n-1,i-1)\frac{G_T^c(i)}{G_T^c(i-1)}, \text{ for } i \leq c_n,
\end{aligned}$$

where the integer c_n and the probability p_n are determined by

$$\sum_{i=c_n+1}^{\infty} q(n-1,i) \leq \sigma(n) < \sum_{i=c_n}^{\infty} q(n-1,i), \tag{B.6}$$

$$P_n = \frac{\sigma(n) - \sum_{i=c_n+1}^{\infty} q(n-1,i)}{q(n-1,c_n)}. \tag{B.7}$$

Other performance functions can be computed as follows:

$$\alpha_n = \sum_{i=0}^{c_n-1} q(n-1,i)\frac{g_T(i+1)}{G_T^c(i)} + (1-P_n)q(n-1,c_n)\frac{g_T(c_n+1)}{G_T^c(c_n)}, \tag{B.8}$$

$$Q(n) = \sum_{i=1}^{c_n} q(n,i). \tag{B.9}$$

(2) $\sigma(n) \geq Q(n-1) + q(n-1,0)$.

In this case $I(n) = 0$ and $b(n) = Q(n-1) + q(n-1,0)$. Define $c_n = 0$ and $P_n = 1$. Further,

$$\alpha(n) = 0,\ Q(n) = 0,\ \text{and } q(n,i) = 0 \text{ for all } i > 0.$$

Case 2: $B(n-1) < 1$.

If $B(n-1) < 1$, then there is idle server after time $n-1$. In this case, (B.1) and (B.2) still hold. However, instead of (B.3), the boundary conditions are characterized by

$$b(n,0) = \min\left\{\sigma(n)\delta + 1 - B(n-1), q(n-1,0)\right\},$$

$$q(n,0) = \lambda\delta.$$

Queue contents are updated:

$$q(n,i) = 0,\ \text{for } i \geq 2,$$

$$q(n,1) = \left(q(n-1,0) - b(n,0)\right)^+ G_T^c(1). \tag{B.10}$$

Other performance functions are computed as follows:

$$I(n) = \mathbf{1}_{[\sigma(n)\delta+1-B(n-1)<q(n-1,0)]},$$

$$c_n = 0,$$

$$\alpha(n) = (q(n-1,0) - b(n,0))^+ f_T(1).$$

B.2. Proofs

Proof of Proposition 3.1. It suffices to show the nested σ and w in (3.12) and (3.14) have a unique solution. I discuss two cases depending on whether the system is UL or OL at time 0.

Case 1: If the system is OL at time 0, then $w(t) > 0$ for all $t < T_{OL}$ where T_{OL} is the termination time of the OL interval. I will show there exists a unique solution (w, σ) to (3.12) and (3.14).

Define an operator $\Gamma : \mathbb{C}^2 \mapsto \mathbb{C}^2$ via

$$\Gamma(l,m) = \left(\int_0^t l(t-x)\psi(m(t-x),x)dx + c(t), \int_0^t \left[1 - \frac{l(x)}{\tilde{q}(x,m(x))}\right] dx + w(0) \right).$$

Then it is evident that (w, σ) is a fixed point of the operator Γ. I use the complete normed space $\mathbb{C}^2$ with the following norm over the interval $[0, t]$,

$$\|(l,m)\|_t = \sup_{0\le u\le t} \{|l(u)|\} + \sup_{0\le u\le t} \{|m(u)|\}.$$

I will apply the Banach fixed point theorem by showing Γ is a contraction mapping. Note that the expression of $\tilde{q}(t, w(t))$ in (3.15) depends on the relative value of t and $w(t)$.

Therefore, I will analyze Γ separately for $t < v(0)$ and $t \geq v(0)$, where $v(0)$ is the time for the initial queue to be cleared. For $t < v(0)$, the initial queue is not cleared yet. Hence $t < w(t)$ for all $t < v(0)$. Restricting Γ on time $[0, v(0)]$, then

$$\begin{aligned}
&\Gamma(l, m) \\
&= \left(\int_0^u l(u-x)\psi(m(u-x), x)du + c(t), \int_0^t \left[1 - \frac{l(x)}{q(0, m(x)-x)\frac{F_T^c(m(x))}{F_T^c(m(x)-x)}} \right] dx + w(0) \right).
\end{aligned}$$

The total service rate of the initial fluid content in service is bounded, $\sup_{0 \leq u \leq t} |c(u)| < \infty$. Hence, Γ maps $\mathbb{C}^2$ to $\mathbb{C}^2$.

$$\begin{aligned}
&\|\Gamma(l_1, m_1) - \Gamma(l_2, m_2)\|_t \\
&= \sup_{0 \leq u \leq t} \left\{ \left| \int_0^u l_1(u-x)\psi(m_1(u-x), x) - l_2(u-x)\psi(m_2(u-x), x)dx \right| \right\} \\
&\quad + \sup_{0 \leq u \leq t} \left\{ \left| \int_0^t \frac{l_1(x)}{q(0, m_1(x)-x)\frac{F_T^c(m_1(x))}{F_T^c(m_1(x)-x)}} - \frac{l_2(x)}{q(0, m_2(x)-x)\frac{F_T^c(m_1(x))}{F_T^c(m_2(x)-x)}} dx \right| \right\}.
\end{aligned}$$

I analyze the first term:

$$
\begin{aligned}
&\left| \int_0^u l_1(u-x)\psi(m_1(u-x),x) - l_2(u-x)\psi(m_2(u-x),x)dx \right| \\
\leq &\int_0^u \left| l_1(u-x)\psi(m_1(u-x),x) - l_2(u-x)\psi(m_2(u-x),x) \right| dx \\
= &\int_0^u \Big| \left[l_1(u-x) - l_2(u-x) \right] \psi(m_1(u-x),x) \\
&+ l_2(u-x)[\psi(m_1(u-x),x) - \psi(m_2(u-x),x)] \Big| dx \\
\leq &\int_0^u \left| \left[l_1(u-x) - l_2(u-x) \right] \psi(m_1(u-x),x) \right| \\
&+ \left| l_2(u-x)[\psi(m_1(u-x),x) - \psi(m_2(u-x),x)] \right| dx \\
\leq &\sup_{0\leq x\leq u} |l_1(x) - l_2(x)| \int_0^u \psi(m_1(u-x),x)dx \\
&+ \sup_{0\leq x\leq u} |l_2(x)| \cdot \int_0^u \left| \psi(m_1(u-x),x) - \psi(m_2(u-x),x) \right| dx \\
\overset{(a)}{\leq} &\sup_{0\leq x\leq u} |l_1(x) - l_2(x)| \cdot M_1 u + \sup_{0\leq x\leq u} |m_1(x) - m_2(x)| \cdot M_2 u \quad \text{for some } M_1 \text{ and } M_2 \\
\leq &(M_1 + M_2)u \cdot \left[\sup_{0\leq x\leq u} |l_1(x) - l_2(x)| + \sup_{0\leq x\leq u} |m_1(x) - m_2(x)| \right],
\end{aligned}
$$

where inequality (a) follows from the continuity of ψ. In particular, since m_1 is continuous, hence $m_1(x)$ is bounded for $0 \leq x \leq t$. Then the continuity of ψ implies there exists some $M_1 > 0$ such that $\psi(m_1(u-x),x) < M_1$ for all $0 \leq x \leq t$. On the other hand,the continuity of ψ also implies there exists $\hat{M}_2 > 0$ such that $\left|\psi(m_1(u-x),x) - \psi(m_2(u-x),x)\right| < \hat{M}_2 |m_1(t-x) - m_2(t-x)|$ for $0 \leq x \leq t$. Let $M_2 = \hat{M}_2 \cdot \sup_{0\leq x\leq t} |l_2(x)|$. I next analyze

the second term:

$$
\begin{aligned}
&\int_0^u \left| \frac{l_1(x)}{q(0, m_1(x)-x)\frac{F_T^c(m_1(x))}{F_T^c(m_1(x)-x)}} - \frac{l_2(x)}{q(0, m_2(x)-x)\frac{F_T^c(m_1(x))}{F_T^c(m_2(x)-x)}} \right| dx \\
\leq &\int_0^u \left| \frac{l_1(x)-l_2(x)}{q(0, m_1(x)-x)\frac{F_T^c(m_1(x))}{F_T^c(m_1(x)-x)}} \right| \\
&+ \left| \frac{l_2(x)}{q(0, m_1(x)-x)\frac{F_T^c(m_1(x))}{F_T^c(m_1(x)-x)}} - \frac{l_2(x)}{q(0, m_2(x)-x)\frac{F_T^c(m_1(x))}{F_T^c(m_2(x)-x)}} \right| dx \\
\overset{(b)}{\leq} &\sup_{0\leq x\leq u} |l_1(x)-l_2(x)| \cdot M_3 u + \sup_{0\leq x\leq u} |m_1(x)-m_2(x)| \cdot M_4 u \quad \text{for some } M_3 \text{ and } M_4 \\
\leq &(M_3+M_4)u \cdot \left[\sup_{0\leq x\leq u} |l_1(x)-l_2(x)| + \sup_{0\leq x\leq u} |m_1(x)-m_2(x)| \right],
\end{aligned}
$$

where inequality (b) follows from the continuity of initial conditions and the distribution of the conditional service time. Combining the results together, I have

$$
\begin{aligned}
&\|\Gamma(l_1, m_1) - \Gamma(l_2, m_2)\|_t \\
\leq &\sup_{0\leq u\leq t} (M_1+M_2+M_3+M_4)u \cdot \left[\sup_{0\leq x\leq u} |l_1(x)-l_2(x)| + \sup_{0\leq x\leq u} |m_1(x)-m_2(x)| \right] \\
\leq &(M_1+M_2+M_3+M_4)t \cdot \left[\sup_{0\leq x\leq t} |l_1(x)-l_2(x)| + \sup_{0\leq x\leq t} |m_1(x)-m_2(x)| \right].
\end{aligned}
$$

Let $\delta = 1/2(M_1+M_2+M_3+M_4)$, then Γ is a contraction mapping on $[0, \delta]$. Therefore, (w, σ) uniquely solves (3.12) and (3.14) on $[0, \delta]$. Then I can recursively consider successive intervals of length δ to show (w, σ) uniquely solves (3.12) and (3.14) on $[0, t]$ for $t < v(0)$.

Next I consider $t > v(0)$. Now I have $\tilde{q}(t, w(t)) = \lambda(t) F_T^c(w(t))$. I consider the original system at time $v(0)$ to be a new system at time 0 by adjusting $b(0, \cdot)$ to $\hat{b}(0, \cdot)$ and $q(0, \cdot)$ to $\hat{q}(0, \cdot)$ accordingly. In particular, I perform a backward time shift to replace t with

$t - v(0)$ wherever t appears. The new system is OL at time 0 and I will show the new tail $\hat{b}(0,\cdot)$ is relatively bounded relative to new initial conditional service time.

$$\begin{aligned}
\hat{C}_1(t) &:= \sup_{0\le u\le t}\Big[\int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x), v(0)+u+x)}{\bar{\Psi}(w_B(x),x)}dx \\
&\quad + \int_0^{v(0)} q(x, w(x))\psi(w(x), v(0)-x+u)dx\Big] \\
&= \sup_{0\le u\le t}\big[\int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x), v(0)+u+x)}{\bar{\Psi}(w_B(x),x)}dx \\
&\quad + \int_0^{v(0)} q(0, w(x)-x)\frac{F_T^c(w(x))}{F_T^c(w(x)-x)}\psi(w(x), v(0)-x+u)dx\big] \\
&\overset{(c)}{<} \sup_{0\le u\le t}\int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x), v(0)+u+x)}{\bar{\Psi}(w_B(x),x)}dx + M_5\int_0^{v(0)} q(0, w(x)-x)dx \\
&= \sup_{0\le u\le t}\int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x), v(0)+u+x)}{\bar{\Psi}(w_B(x),x)}dx + M_5\cdot Q(0) \\
&< \infty,
\end{aligned}$$

where inequality (c) follows from the continuity of ψ. I can use a similar argument to show Γ is a contraction mapping on $[v(0), v(0)+\delta]$ for small δ. By considering recursively, I can show (w,σ) uniquely solves (3.12) and (3.14) on $[0, T_{OL}]$. If $T_{OL} = \infty$, then the system stays OL all the time and I have shown the existence and uniqueness of (w,σ) on $[0,\infty)$. If $T_{OL} < \infty$, then I consider the original system at time T_{OL} to be a new system at time 0 by adjusting $b(0,\cdot)$ to $\hat{b}(0,\cdot)$ and $q(0,\cdot)$ to $\hat{q}(0,\cdot)$ accordingly. I analyze the UL interval in Case 2. Here I only need to establish that the new tail $\hat{b}(0,\cdot)$ is relatively

bounded relative to new initial conditional service time.

$$
\begin{aligned}
&\hat{C}_2(t) \\
&= \sup_{0\le u\le t}\Bigg[\int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x),T_{OL}+u+x)}{\bar{\Psi}(w_B(x),x)}dx + \int_0^{v(0)} q(x,w(x))\psi(w(x),T_{OL}-x+u)dx \\
&\quad + \int_{v(0)}^{T_{OL}} q(x,w(x))\psi(w(x),t+T_{OL}-x)dx\Bigg] \\
&\overset{(d)}{\le} \sup_{0\le u\le t}\hat{C}_1(T_{OL}-v(0)+t) + M_6(T_{OL}-v(0)) \sup_{0\le x\le T_{OL}}\lambda(x) \\
&< \infty,
\end{aligned}
$$

where inequality (d) follows from the definition of $\hat{C}_1$ and $q(x,w(x)) \le \lambda(x-w(x))$.

Case 2: If the system is UL at time 0, then $w(t) = 0$ for all $t < T_{UL}$ where T_{UL} is the termination time of the UL interval. No queue is accumulated in this interval and b evolves according to (3.4) with $w(t) = 0$ for all $t < T_{UL}$.

$$
\begin{aligned}
\sigma(t) &= \int_0^t b(t-x,0)\bar{\Psi}(w(t-x),x)h(w(t-x),x)dx \\
&\quad + \int_t^\infty b(0,x-t)\bar{\Psi}(w_B(t-x),x)h(w_B(t-x),x)du \\
&= \int_0^t \lambda(x)\psi(0,x)dx + c(t).
\end{aligned}
$$

If $T_{UL} = \infty$, then the system is UL all the time and all performance functions are explicitly determined. If $T_{UL} < \infty$, then I consider the original system at time T_{UL} to be a new system at time 0 by adjusting $b(0,\cdot)$ to $\hat{b}(0,\cdot)$ accordingly. The new system is OL at time 0 and I will show the new tail $\hat{b}(0,\cdot)$ is relatively bounded relative to new initial conditional service time. Then I can employ the argument in Case 1 to show the existence of a

solution (w, σ). Note that

$$\begin{aligned}\hat{C}(t) &= \sup_{0\le u\le t}\left[\int_{\mathcal{S}_B(0)} \frac{b(0,x)\psi(w_B(x), T_{UL}+u+x)}{\bar{\Psi}(w_B(x),x)}dx + \int_0^{T_{UL}} \lambda(x)\psi(0, T_{UL}+u-x)dx\right] \\ &< \sup_{0\le u\le t}\int_{\mathcal{S}_B(0)} \frac{b(0,x)f_z(T_{UL}+u+x|w_B(x))}{\bar{F}_S(x|w_B(x))}dx + \sup_{0\le x\le T_{UL}}\lambda(x) \\ &< \infty.\end{aligned}$$

Finally, I can show the existence and uniqueness of (w, σ) on the entire timeline by recursively considering alternating OL and UL intervals. Q.E.D.

Proof of Proposition 3.2. I invoke the proof of a similar comparison result in Liu and Whitt (2012). I first claim $\sigma_1(t) \le \sigma_2(t)$ when $B_1(t) = B_2(t)$. This is straightforward because

$$\begin{aligned}\sigma_1(t) =& \int_0^\infty b(t,x)h_1(w_1(t-x),x)dx \\ \le& \sup_{z\ge 0, x\in\mathcal{S}_1(z)} h_1(z,x)\int_0^\infty b(t,x)dx \\ =& B_1(t)\sup_{z\ge 0, x\in\mathcal{S}_1(z)} h_1(z,x).\end{aligned}$$

Similarly,

$$\sigma_2(t) \ge B_2(t)\inf_{z\ge 0, x\in\mathcal{S}_2(z)} h_2(z,x).$$

By (3.16), it follows that $\sigma_1(t) \le \sigma_2(t)$ when $B_1(t) = B_2(t)$.

I discuss three cases: (i) when both systems are UL, (ii) when the upper system is OL and the lower system is UL, and (iii) when both systems are OL. I apply mathematical

induction over the successive alternating intervals of the three cases by assuming the initial conditions for each succeeding interval satisfy the ordering assumed in the proposition.

I first consider when both systems are UL. Then $w_1(t) = w_2(t) = 0$ until one of the system becomes OL. Suppose the ordering $B_1(t_1) \geq B_2(t_1)$ holds for some t_1 in a common UL interval of both systems. If there exists $t_2 > t_1$ such that $B_1(t_2) < B_2(t_2)$. Then by the continuity of B_i, there exists t_0 satisfying $t_1 \leq t_0 < t_2$ such that $B_1(t_0) = B_2(t_0)$. Note that in an UL interval, $B_i'(t) = \lambda(t) - \sigma_i(t)$. I have showed that $\sigma_1(t_0) \leq \sigma_2(t_0)$ since $B_1(t_0) = B_2(t_0)$. Then it follows that $B_1'(t_0) \geq B_2'(t_0)$. This contradicts the assumption that there exists $t_2 > t_0$ such that $B_1(t_2) < B_2(t_2)$. I also find that the UL termination times are ordered as well. This is because $B_1 \geq B_2$ in the UL interval of B_1, which I will show later. Whenever system 2 becomes OL, i.e., $B_2(t) = 1$ and $\lambda > \sigma_2(t)$, I must have $B_1(t) \geq B_2(t) = 1$ and $\lambda(t) > \sigma_2(t) \geq \sigma_1(t)$. This implies system 1 is OL at the UL termination time of system 2. Therefore, the UL termination time of system 1 must come before that of system 2.

When the upper system is OL while the lower system is UL, it is obvious that the ordering $B_1(t) \geq B_2(t)$ and $w_1(t) \geq w_2(t)$ remain valid.

I consider when both systems are OL. Suppose the ordering $w_1(t_1) \geq w_2(t_1)$ holds for some t_1 in a common OL interval of both systems. If there exists $t_2 > t_1$ such that $w_1(t_2) < w_2(t_2)$. Then by the continuity of w_i, there exists t_0 satisfying $t_1 \leq t_0 < t_2$ such that $w_1(t_0) = w_2(t_0)$. By (3.14), $w_i'(t) = 1 - \sigma_i(t)/\tilde{q}_i(t, w(t))$. On one hand, $\tilde{q}_1(t_0, w(t_0)) = \tilde{q}_1(t_0, w(t_0))$. On the other, $\sigma_1(t_0) \leq \sigma_2(t_0)$ since $B_1(t_0) = B_2(t_0)$. Then it follows that $w_1'(t_0) \geq w_2'(t_0)$. This contradicts the assumption that there exists $t_2 > t_0$ such that $B_1(t_2) < B_2(t_2)$.

By Corollary 3 of Liu and Whitt (2012), it holds that

$$q_i(t,x) = q_i(t-x,0)F_T^c(x)\mathbf{1}_{[x\le w_i(t)\wedge t]} + q(0,x-t)\frac{F_T^c(x)}{F_T^c(x-t)}\mathbf{1}_{[t\le x\le w_i(t)]}.$$

Since $w_1(t) \ge w_2(t)$, hence $q_1(t,\cdot) \ge q_2(t,\cdot)$. Q.E.D.

Proof of Proposition 3.3. In stationarity, the service input (rate that fluid enters service) must equal service output (rate that fluid completes service). This implies $b(t,0) = \sigma(t) = \sigma$. Then the basic evolution equation (3.4) implies $b(t,x) = \sigma\Psi(w(t-x),x)$. In stationarity, all performance functions are time-independent. In particular, $w(t-x)$ does not evolve with t. So there exists a function $l(\cdot)$ such that $w(t-x) = l(x)$ for all t. I claim that $l(\cdot)$ is a constant. Pick an arbitrary pair $(x,y) \in \mathbb{R}_+^2$ satisfying $y > x$, I show $l(x) = l(y)$. Pick any $t_1 > x + l(x)$ and define $t_2 = t_1 - x + y > t_1$. Then $l(x) = w(t_1 - x) = w(t_2 - y) = l(y)$. This implies $l(\cdot)$ is a constant, which I denote by w^*. Then $b(x) = \sigma\bar{\Psi}(w^*,x)$ for all $x \ge 0$, which, by integrating implies $B = \sigma/\mu(w^*)$.

Because the total fluid input must equal the total fluid output, I must have $\lambda = \sigma + \alpha$. Note that when $Q > 0$, I must have $B = 1$. On the other hand, $Q > 0$ is equivalent to $\alpha > 0$ and $w^* = 0$. If $\lambda > \mu(0)$, the system has an overload and I claim that $Q > 0$. Take the contradictory assumption that $Q = 0$, then $w = 0$ so that $\sigma = \mu(0) < \lambda$. This implies $\alpha = \lambda - \sigma > 0$, contradicting $Q = 0$. Hence $Q > 0$ and $B = 1$. If $\lambda \le \mu(0)$, I discuss two possible scenarios: (i) $Q > 0$ so that $B = 1$, $\alpha > 0$ and $w^* > 0$. (ii) $Q = 0$ so that $\alpha = 0$ and $\sigma = \lambda = B\mu(0)$ hence $B = \lambda/\mu(0)$. The following discussion will rule out the first scenario.

Note that when $w^* > 0$, $B = 1$ must hold, hence $\sigma = \mu(w^*)$. Then $\alpha = \lambda - \sigma = \lambda - \mu(w^*)$. The abandonment rate α, on the other hand, can be computed by

$$\alpha = \int_0^{w^*} q(x)h_T(x)dx = \int_0^{w^*} q(0)F_T^c(x)\frac{f_T(x)}{F_T^c(x)}dx = \int_0^{w^*} \lambda f_T(x)dx = \lambda F_T(w^*).$$

Plug in $\alpha = \lambda - \mu(w^*)$ and I obtain $\lambda F_T^c(w^*) = \mu(w^*)$, which is exactly (3.19). Under Condition 3.1, there is a unique solution $w^* > 0$ that solves (3.19) if and only if $\lambda > \mu(0)$. Hence the first scenario discussed above when $\lambda \leq \mu(0)$ is ruled out. Therefore, when $\lambda \leq \mu(0)$, I must have $Q = \alpha = w^* = 0$, $\sigma = \lambda$ and $B = \lambda/\mu(0)$.

Now it remains to show the vector $(b, q, \sigma, \alpha, Q, B, w, v)$ given above is indeed a stationary point. In other words, if the system starts with it, the performance functions do no change over time.

When $\lambda \leq \mu(0)$, the system is underloaded or critically loaded. The maximal service capacity can handle all incoming fluid, hence $w(t) = 0$ for all $t > 0$ if the system is initialized with no queue. The initial fluid in service satisfies $b(0, x) = \lambda F_z^c(x|0)$ and hence the initial total service rate

$$\sigma(0) = \int_0^{\infty} b(0, x)(0, x)dx = \int_0^{\infty} \lambda\bar{\Psi}(0, x)\frac{\psi(0, x)}{\bar{\Psi}(0, x)}dx = \lambda.$$

Then it follows for small $t > 0$,

$$\begin{aligned} b(t, x) &= b(t - x, 0)\bar{\Psi}(0, x)\mathbf{1}_{[0\leq x\leq t]} + b(0, x - t)\frac{\bar{\Psi}(0, x)}{\bar{\Psi}(0, x - t)}\mathbf{1}_{[x>t]} \\ &= \lambda\bar{\Psi}(0, x)\mathbf{1}_{[0\leq x\leq t]} + \lambda\bar{\Psi}(0, x - t)\frac{\bar{\Psi}(0, x)}{\bar{\Psi}(0, x - t)}\mathbf{1}_{[x>t]} = \lambda\bar{\Psi}(0, x) = b(0, x). \end{aligned}$$

When $\lambda > \mu(0)$, I have $\sigma(0) = \mu(w^*)$ and $b(0,x) = F_z^c(x|w^*)\mu(w^*)$. In this case,

$$c(t) = \int_0^\infty \frac{b(0,x)\psi(w^*,t+x)}{\bar{\Psi}(w^*,x)}dx = \int_0^\infty \psi(w^*,t+x)\mu(w^*)dx = \bar{\Psi}(w^*,t)\mu(w^*).$$

If $w(0) = w^*$, then by (3.13) I have $w'(0+) = 1 - \sigma(0)/(\lambda F_T^c(w^*)) = 0$. In an OL interval, I have $\sigma(t) = b(t,0)$. Hence for small $t > 0$, $b(t,0)$ solves

$$b(t,0) = \int_0^t b(t-x,0)\psi(w^*,x)dx + c(t) = \int_0^t b(t-x,0)\psi(w^*,x)dx + \bar{\Psi}(w^*,t)\mu(w^*).$$

Noting that $\sup_{0\le u\le t} c(u) < \infty$, I can use a similar approach in the proof of Proposition 3.1 to show the above equation has a unique fixed point $b(s) = \mu(w^*)$ for all $s \in [0,t]$. The fluid density in queue satisfies

$$\begin{aligned} q(t,x) &= \lambda F_T^c(x)\mathbf{1}_{[x\le t]} + q(0,x-t)\frac{F_T^c(x)}{F_T^c(x-t)}\mathbf{1}_{[t<t\le w(t)]} \\ &= \lambda F_T^c(x)\mathbf{1}_{[0\le x\le w(t)]}. \end{aligned}$$

Since $\sigma(t) = b(t,0)$ are constants, I have $w'(t) = 1 - \sigma(t)/(\lambda F_T^c(w(t))) = 0$ so that $w(t) = w^*$ and $q(t,x) = q(x)$. Therefore, all performance functions are constants with $0 \le t \le \delta$ for small δ and thus for all $t \ge 0$. Q.E.D.

Proof of Proposition 3.4. Following the argument in the proof of Proposition 3.3, the stationary point of the fluid model for an *overloaded* system can be found by solving (3.19). In this case, the uniqueness of the stationary point is equivalent to the uniqueness of the solution to (3.19). Note that $\lambda\phi(0) = \lambda/\mu(0) > 1$ and $\lambda\phi(\infty) = 0$, by the continuity of $\phi(z)$, (3.19) must have at least one solution. I show that either of two conditions is

sufficient to guarantee the uniqueness of the solution. If $\phi(z)$ is unimodal, then the curve $y = \lambda\phi(z)$ can only cross the horizontal line $y = 1$ once, otherwise there will be more than one mode, contradicting the unimodality of $\phi(z)$. If $1/\phi(z)$ is convex in z, it is sufficient to show the solution to $\xi(z) = 1/\phi(z) - \lambda = 0$ is unique. Define $z_0 = \inf_{z=0}\{\xi(z) = 0\}$, then it must hold that $\xi'(z_0) > 0$. Otherwise if $\xi'(z_0) \leq 0$, then by the convexity of $\xi(z)$, $\xi'(z) \leq 0$ for all $z \leq z_0$. It then follows that $\xi(z_0) \leq \xi(0) = \mu(0) - \lambda < 0$, contradicting the definition of z_0. Thus, I have $\xi'(z_0) > 0$ and the convexity of $\xi(z)$ implies $\xi'(z) > 0$ for all $z > z_0$, which further implies z_0 is the unique solution.

If the system is underloaded with $\lambda < \mu(0)$, Proposition 3.3 prescribes one equilibrium with no queue. This equilibrium is unique if and only if an equilibrium with a strictly positive queue does not exist, i.e., (3.19) does not have a solution. For this latter case to hold, it follows immediately that $\phi(z) < 1/\lambda$ for all $z \geq 0$. Q.E.D.

Proof of Proposition 3.5. By definition of the hazard rate of conditional service time,

$$\begin{aligned}
h(z,x) =& \frac{\psi(z,x)}{\bar{\Psi}(z,x)} \\
=& \frac{f_{S|T}(x|T>z)}{F^c_{S|T}(x|T>z)} \\
=& \frac{\int_z^\infty f(x,y)dy/\mathbb{P}(T>z)}{\mathbb{P}(S>x,T>z)/\mathbb{P}(T>z)} \\
=& \frac{\int_z^\infty f(x,y)dy}{\mathbb{P}(S>x,T>z)} \\
=& \frac{\mathbb{P}(T>z|S=x)f_S(S=x)}{\mathbb{P}(T>z|S>x)F^c_S(x)} \\
=& \frac{\mathbb{P}(T>z|S=x)}{\mathbb{P}(T>z|S>x))} \cdot \mu,
\end{aligned}$$

where I use f_S and F_S to denote the unconditional marginal distributions of S. It then suffices to show $\mathbb{P}(T > z|S = x)/\mathbb{P}(T > z|S > x) \leq 1$ if (S, T) are PRD and $\mathbb{P}(T > z|S = x)/\mathbb{P}(T > z|S > x) \geq 1$ if (S, T) are NRD. I show the first half and the second half follows analogously.

$$\begin{aligned}
\mathbb{P}(T > z|S > x) &= \frac{\mathbb{P}(T > z, S > x)}{\mathbb{P}(S > x)} \\
&= \frac{\int_x^\infty \mathbb{P}(T > z|S = x) f_S(y) dy}{\mathbb{P}(S > x)} \\
&\leq \frac{\mathbb{P}(T > z|S = x) \int_x^\infty f_S(y) dy}{\mathbb{P}(S > x)} \\
&= \mathbb{P}(T > z|S = x),
\end{aligned}$$

where in the inequality I have used the fact that $\mathbb{P}(T > z|S = x)$ is increasing in x.Q.E.D.

Proof of Proposition 3.6. I show the result under condition (i). The proof under other two conditions follows analogously. I complete the proof in four steps. Proposition 3.1 implies there is a unique (w, σ) that solves the fluid model.

Step 1: Let $t_0 := v(0)$ be the time that the initial queue content is cleared. I first claim that for any $t_1 > t_0$, if $w'(t) \geq 0$ for $t \in [t_0, t_1]$, then $\sigma'(t) \leq 0$ for $t \in [t_0, t_1]$. To show this, first note that the differentiability of σ follows from the smoothness of μ. Lemma 3.2 implies that

$$\sigma(t) = \int_0^\infty b(t, x)\mu(w(t - x))dx.$$

Thus,

$$\begin{aligned}\sigma(t+\Delta t) &= \int_0^{\infty} b(t+\Delta t, x)\mu(w(t+\Delta t-x))dx \\ &= \int_0^{\Delta t} b(t+\Delta t, x)\mu(w(t+\Delta t-x))dx + \int_{\Delta t}^{\infty} b(t+\Delta t, x)\mu(w(t+\Delta t-x))dx \\ &= \int_0^{\Delta t} b(t+\Delta t, x)\mu(w(t+\Delta t-x))dx + \int_0^{\infty} b(t+\Delta t, x+\Delta t)\mu(w(t-x))dx.\end{aligned}$$

Subtracting $\sigma(t+\Delta t)$ by $\sigma(t)$ gives

$$\begin{aligned}&\sigma(t+\Delta t)-\sigma(t) \\ =&\int_0^{\Delta t} b(t+\Delta t, x)\mu(w(t+\Delta t-x))dx - \int_0^{\infty} \left[b(t,x)-b(t+\Delta t, x+\Delta t)\right]\mu(w(t-x))dx \\ \overset{(e)}{\leq}&\mu(w(t))\int_0^{\Delta t} b(t+\Delta t, x)dx - \mu(w(t))\int_0^{\infty}\left[b(t,x)-b(t+\Delta t, x+\Delta t)\right]\mu(w(t-x))dx \\ =&\mu(w(t))\left[\int_0^{\infty} b(t+\Delta t, x)dx - \int_0^{\infty} b(t,x)dx\right] \\ =&0,\end{aligned}$$

where inequality (e) follows because w is decreasing in $[t_0, t_1]$ and $w(t) > w(t_0) = w(v(0)) \geq w(x)$ for $x < 0$. Let Δt be sufficiently small, then it follows that $\sigma'(t) \leq 0$.

Step 2: I next show $w'(t) \geq 0$ for all $t \geq t_0$. Suppose this is not true, i.e., there exists $t_3 > t_0$ such that $w'(t_3) < 0$. Since $w'(t_0) \geq 0$ as assumed in condition (i) and $w'(t)$ is continuously differentiable as I assume the smoothness of the initial conditions, define $t_2 := \sup\{t > t_0 : w'(t) \geq 0\}$. Since $w'(t_3) < 0$, it holds that $t_2 < \infty$. By the definition of t_2, I have $w'(t) \geq 0$ for all $t \in [t_0, t_2)$. Differentiating $w'(t) = 1 - \frac{\sigma(t)}{\lambda F_T^c(w(t))}$ on both sides

gives

$$w''(t_2) = -\frac{\sigma'(t_2)F_T^c(w(t_2)) + \sigma(t_2)f_T(w(t_2))w'(t_2)}{(\lambda F_T^c(w(t_2))^2} = -\frac{\sigma'(t_2)F_T^c(w(t_2))}{(\lambda F_T^c(w(t_2))^2} \geq 0.$$

In the last inequality I have used $\sigma'(t_2) \leq 0$, as proved in Step 1. The above equation contradicts the assumption that $w'(t_3) < 0$.

Step 3: I show that $w(t)$ can never hit $w^*(\lambda)$. If $w(t) = w^*(\lambda)$ for some $t > t_0$, then

$$w'(t) = 1 - \frac{\sigma(t)}{\lambda F_T^c(w(t))} = 1 - \frac{\sigma(t)}{\lambda F_T^c(w^*(\lambda))} < 1 - \frac{\mu(w^*(\lambda))}{\lambda F_T^c(w^*(\lambda))} = 0.$$

The inequality follows because

$$\sigma(t) = \int_0^\infty b(t,x)\mu(w(t-x)) < \mu(w(t)) \int_0^\infty b(t,x)dx = \mu(w(t)) = \mu(w^*(\lambda)).$$

Therefore, $w(t) < w^*(\lambda)$ for all $t > t_0$.

Step 4: Since $w(t)$ is increasing for $t > t_0$ (Step 2) and is bounded above by $w^*(\lambda)$ (Step 3), $w(t)$ has a limit $\tilde{w}$ as $t \to \infty$. I show the limit $\tilde{w} = w^*(\lambda)$. If $\tilde{w} < w^*(\lambda)$, then by (3.12),

$$\sigma(t) = \int_0^t \sigma(t-u)f_z(u|w(t-u))du + c(t).$$

First note that

$$\begin{aligned} c(t) &= \int_{\mathcal{S}_B(0)} \frac{b(0,x) f_z(t+x|w_B(x))}{\bar{F}_S(x|w_B(x))} dx \\ &= \int_{\mathcal{S}_B(0)} b(0,x) \exp(-\mu(w_B(x))t) dx \\ &\leq \exp(-\mu(w^*(\lambda))t) B(0) \to 0 \quad \text{as } t \to \infty. \end{aligned}$$

Next,

$$\begin{aligned} &\int_0^t \sigma(t-u) f_z(u|w(t-u)) du \\ =&\int_0^t \sigma(u) f_z(t-u|w(u)) du \\ =&\int_0^{t/2} \sigma(u) f_z(t-u|w(u)) du + \int_{t/2}^t \sigma(u) f_z(t-u|w(u)) du \\ =&\int_0^{t/2} \sigma(u)\mu(w(u)) \exp(-\mu(w(u))(t-u)) du + \int_{t/2}^t \sigma(u)\mu(w(u)) \exp(-\mu(w(u)(t-u)) du. \end{aligned}$$

I analyze the first term,

$$\begin{aligned} &\int_0^{t/2} \sigma(u)\mu(w(u)) \exp(-\mu(w(u))(t-u)) du \\ \leq&\mu(0)^2 \int_0^{t/2} \exp(-\mu(w(u))(t-u)) du \\ \leq&\mu(0)^2 \int_0^{t/2} \exp(-\mu(w^*(\lambda))(t-u)) du \\ \leq&\frac{\mu(0)^2}{\mu(w^*(\lambda))} \big(\exp(-\mu(w^*(\lambda))(t/2)) - \exp(-\mu(w^*(\lambda))t \big) \\ \to& 0 \quad \text{as} \quad t \to \infty. \end{aligned}$$

The second term,

$$\begin{aligned}
&\lim_{t\to\infty}\int_{t/2}^{t}\sigma(u)\mu(w(u))\exp(-\mu(w(u)(t-u))du \\
=&\lim_{t\to\infty}\int_{t/2}^{t}b(t,t-u)\mu(w(u))du \\
=&\lim_{t\to\infty}\int_{0}^{t/2}b(t,u)\mu(w(t-u))du \\
\overset{(f)}{=}&\lim_{t\to\infty}\int_{0}^{t/2}b(t,u)\mu(\tilde{w})du \\
\overset{(g)}{=}&\mu(\tilde{w}).
\end{aligned}$$

In equality (f), I use dominated convergence theorem to take limit of $\mu(w(t-u))$ in the integration. In equality (g), I use the fact that $\int_{t/2}^{\infty}b(t,u)\to 0$ as $t\to\infty$. The proof of this is similar to the analysis of the first term and is omitted for brevity. Hence it holds that

$$\lim_{t\to\infty}\sigma(t)=\mu(\tilde{w}).$$

This implies

$$\lim_{t\to\infty}w'(t)=\lim_{t\to\infty}1-\frac{\sigma(t)}{\lambda F_T^c(w(t))}=1-\frac{\mu(\tilde{w})}{\lambda F_T^c(\tilde{w})}>0.$$

The last inequality follows from Assumption 3.1 and $\tilde{w}<w^*(\lambda)$. Hence one can find δ such that $0<\delta<1-\mu(\tilde{w})/\lambda F_T^c(\tilde{w})$. It then follows that $w'(t)>\delta$ for sufficiently large t. This contradicts the assumption that $\lim_{t\to\infty}w(t)=\tilde{w}$.

Combining the results in Step 3 and Step 4, I conclude that $\lim_{t\to\infty}w(t)=w^*(\lambda)$. Q.E.D.

References

Afèche, P. (2013). Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management*, 15(3):423–443.

Afèche, P. and Pavlin, J. M. (2016). Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science*, 62(8):2412–2436.

Alizamir, S., de Véricourt, F., and Sun, P. (2013). Diagnostic accuracy under congestion. *Management Science*, 59(1):157–171.

Anand, K. S., Pac, M. F., and Veeraraghavan, S. (2011). Quality-speed conundrum: trade-offs in customer-intensive services. *Management Science*, 57(1):40–56.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., Yom-Tov, G. B., et al. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194.

Armony, M., Shimkin, N., and Whitt, W. (2009). The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81.

Baccelli, F., Boyer, P., and Hebuterne, G. (1984). Single-server queues with impatient customers. *Advances in Applied Probability*, 16(4):887–905.

Balachandran, K. R. and Radhakrishnan, S. (1994). Extensions to class dominance characteristics. *Management Science*, 40(10):1353–1360.

Bassamboo, A. and Randhawa, R. S. (2010). On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations research*, 58(5):1398–1413.

Bassamboo, A. and Randhawa, R. S. (2015). Scheduling homogeneous impatient customers. *Management Science*, 62(7):2129–2147.

Bassamboo, A., Randhawa, R. S., and Zeevi, A. (2010). Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56(10):1668–1686.

Batt, R. J. and Terwiesch, C. (2016). Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science*.

Block, H. W., Savits, T. H., and Shaked, M. (1985). A concept of negative dependence using stochastic ordering. *Statistics & probability letters*, 3(2):81–86.

Bocklund, L. and Hinton, B. (2008). Cost structure and distribution in todays contact centers. *Strategiccontact. com*.

Borst, S., Mandelbaum, A., and Reiman, M. I. (2004). Dimensioning large call centers. *Operations research*, 52(1):17–34.

Boxma, O. J. and Vlasiou, M. (2007). On queues with service and interarrival times depending on waiting times. *Queueing Systems*, 56(3-4):121–132.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50.

Cachon, G. P. and Feldman, P. (2011). Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing & Service Operations Management*, 13(2):244–260.

Cario, M. C. and Nelson, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Citeseer.

Chan, C. W., Farias, V. F., and Escobar, G. J. (2016). The impact of delays on service times in the intensive care unit. *Management Science*.

Chan, C. W., Yom-Tov, G., and Escobar, G. (2014). When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482.

Clemen, R. T. and Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224.

Colangelo, A., Scarsini, M., and Shaked, M. (2006). Some positive dependence stochastic orders. *Journal of Multivariate Analysis*, 97(1):46–78.

Corbett, C. J. and Rajaram, K. (2006). A generalization of the inventory pooling effect to non-normal dependent demand. *Manufacturing & Service Operations Management*, 8(4):351–358.

Delasay, M., Ingolfsson, A., and Kolfal, B. (2016). Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research*, 64(4):867–885.

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R., and Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31(1):3–33.

Dietz, D. C. (2011). Practical scheduling for call center operations. *Omega*, 39(5):550–557.

Dong, J., Feldman, P., and Yom-Tov, G. B. (2015). Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research*, 63(2):305–324.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, pages 176–223.

Factbook, C. (2017). The world factbook. central intelligence agency.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.

Garnett, O., Mandelbaum, A., and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227.

Gurvich, I., Lariviere, M., and Ozkan, C. (2016). Coverage, coarseness and classification: Determinants of social efficiency in priority queues.

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

Hassin, R. and Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media.

Hopp, W. J., Iravani, S. M., and Yuen, G. Y. (2007). Operations systems with discretionary task completion. *Management Science*, 53(1):61–77.

Hu, K., Allon, G., and Bassamboo, A. (2016). Understanding customers retrial in call centers: Preferences for service quality and service speed. Working paper.

Huang, T., Allon, G., and Bassamboo, A. (2013). Bounded rationality in service systems. *Manufacturing & Service Operations Management*, 15(2):263–279.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kang, W. and Ramanan, K. (2010). Fluid limits of many-server queues with reneging. *The Annals of Applied Probability*, 20(6):2204–2260.

Kc, D. S. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498.

Kc, D. S. and Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, 14(1):50–65.

Kim, Y. J. and Mannino, M. V. (2003). Optimal incentive-compatible pricing for m/g/1 queues. *Operations Research Letters*, 31(6):459–461.

Kuntz, L., Mennicken, R., and Scholtes, S. (2014). Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science*, 61(4):754–771.

Lederer, P. J. and Li, L. (1997). Pricing, production, scheduling, and delivery-time competition. *Operations Research*, 45(3):407–420.

Li, A. A. and Whitt, W. (2014). Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. *Performance Evaluation*, 80:82–101.

Liu, Y. and Whitt, W. (2011a). Large-time asymptotics for the g t/m t/s t+ gi t many-server fluid queue with abandonment. *Queueing systems*, 67(2):145–182.

Liu, Y. and Whitt, W. (2011b). A network of time-varying many-server fluid queues with customer abandonment. *Operations research*, 59(4):835–846.

Liu, Y. and Whitt, W. (2012). The gt/gi/st+ gi many-server fluid queue. *Queueing Systems*, 71(4):405–444.

Lovett, P., Kahn, J., and Greene, S. (2014). Early quick acuity score provides more complete data on emergency department walkouts. *PLoS One*, 9(1).

Maglaras, C., Yao, J., and Zeevi, A. (2016). Optimal price and delay differentiation in queueing systems. *Management Science*.

Mak, H.-Y. and Shen, Z.-J. M. (2014). Pooling and dependence of demand and yield in multiple-location inventory systems. *Manufacturing & Service Operations Management*, 16(2):263–269.

Mendelson, H. and Whang, S. (1990). Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations research*, 38(5):870–883.

Moyal, P. (2017). Coupling in the queue with impatience: case of several servers. *Working paper*.

Müller, A. (2000). On the waiting times in queues with dependency between interarrival and service times. *Operations Research Letters*, 26(1):43–47.

Müller, A. and Scarsini, M. (2001). Stochastic comparison of random vectors with a common copula. *Mathematics of operations research*, 26(4):723–740.

Nazerzadeh, H. and Randhawa, R. S. (2015). Near-optimality of coarse service grades for customer differentiation in queueing systems. *Available at SSRN 2438300*.

Nelsen, R. B. (2013). *An introduction to copulas*, volume 139. Springer Science & Business Media.

Pang, G. and Whitt, W. (2012). The impact of dependent service times on large-scale service systems. *Manufacturing & Service Operations Management*, 14(2):262–278.

Pang, G. and Whitt, W. (2013). Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems*, 73(2):119–146.

Perry, O. and Whitt, W. (2009). Responding to unexpected overloads in large-scale service systems. *Management Science*, 55(8):1353–1367.

Perry, O. and Whitt, W. (2015). Achieving rapid recovery in an overload control for large-scale service systems. *INFORMS journal on computing*, 27(3):491–506.

Perry, O. and Whitt, W. (2016). Chattering and congestion collapse in an overload switching control. *Stochastic Systems*, 6(1):132–210.

Reich, M. (2012). *The offered-load process: Modeling, inference and applications*. PhD thesis, Technion-Israel Institute of Technology.

Ren, Z. J. and Zhou, Y.-P. (2008). Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383.

Scarsini, M. and Shaked, M. (1996). Positive dependence orders: A survey. In *Athens conference on applied probability and time series analysis*, pages 70–91. Springer.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic orders*. Springer Science & Business Media.

Sharakhmetov, S. and Ibragimov, R. (2002). A characterization of joint distribution of two-valued random variables and its applications. *Journal of Multivariate Analysis*, 83(2):389–408.

Staats, B. R. and Gino, F. (2012). Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management science*, 58(6):1141–1159.

Tan, T. F. and Netessine, S. (2014). When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593.

Van Mieghem, J. A. (2000). Price and service discrimination in queuing systems: Incentive compatibility of gc μ scheduling. *Management Science*, 46(9):1249–1267.

Vries, J. d., Roy, D., and Koster, R. d. (2017). Worth the wait? how waiting influences customer behavior and their inclination to return. *Working paper*.

Whitt, W. (1990). Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems*, 6(1):335–351.

Whitt, W. (2006a). Fluid models for multiserver queues with abandonments. *Operations research*, 54(1):37–54.

Whitt, W. (2006b). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88.

Whitt, W. and You, W. (2016). Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research*.

Wu, C. A., Bassamboo, A., and Perry, O. (2017). Service systems with dependent service and patience times. *Management Science, forthcoming*.

Zhang, J. (2013). Fluid models of many-server queues with abandonment. *Queueing Systems*, 73(2):147–193.