Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Service System with Dependent Service and Patience Times

Chenguang (Allen) Wu, Achal Bassamboo, Ohad Perry

To cite this article:
Chenguang (Allen) Wu, Achal Bassamboo, Ohad Perry (2019) Service System with Dependent Service and Patience Times. Management Science 65(3):1151-1172. https://doi.org/10.1287/mnsc.2017.2983

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 65, No. 3, March 2019, pp. 1151–1172
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Service System with Dependent Service and Patience Times

Chenguang (Allen) Wu,[a,b] **Achal Bassamboo,**[c] **Ohad Perry**[a]

[a] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208; [b] Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong; [c] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208

Contact: allenwu@u.northwestern.edu, http://orcid.org/0000-0002-2528-0286 (C(A)W); a-bassamboo@kellogg.northwestern.edu, http://orcid.org/0000-0001-7758-4751 (AB); ohad.perry@northwestern.edu, http://orcid.org/0000-0002-4584-3015 (OP)

Received: January 7, 2017
Revised: August 9, 2017
Accepted: October 9, 2017
Published Online in Articles in Advance: May 8, 2018

https://doi.org/10.1287/mnsc.2017.2983

Abstract. Motivated by recent empirical evidence, we consider a large service system in which the patience time of each customer depends on his service requirement. Our goal is to study *the impact of such dependence on key performance measures*, such as expected waiting times and average queue length, as well as on optimal capacity decisions. Since the dependence structure renders exact analysis intractable, we employ a stationary fluid approximation that is based on the entire joint distribution of the service and patience times. Our results show that even moderate dependence has significant impacts on system performance, so considering the patience and service times to be independent when they are in fact dependent is futile. We further demonstrate that Pearson's correlation coefficient, which is commonly used to measure and rank dependence, is an insufficient statistic, and that the entire joint distribution is required for comparative statics. Thus, we propose a novel framework, incorporating the fluid model with bivariate dependence orders and copulas, to study the impacts of the aforementioned dependence. We then demonstrate how that framework can be applied to facilitate revenue optimization when staffing and abandonment costs are incurred. Finally, the effectiveness of the fluid-based approximations and optimal-staffing prescriptions is demonstrated via simulations.

History: Accepted by Noah Gans, stochastic models and simulation.
Funding: The first and third authors were partially supported by the National Science Foundation [Grant CMMI 1436518].

Keywords: service systems • dependent primitives • fluid approximation • bivariate dependence order • copulas • capacity sizing

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 De Vries et al. (2017). In this paper, we study the impact of such dependence on system performance and optimal staffing.

To motivate our analytical study, we 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, we 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 we 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 $\rho := \lambda/(s\mu) = 1.1$. We 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 our simulation study, we 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 1. Simulation Estimations of Stationary Performance Measures ($\lambda = 110, s = 100$)

Correlation	Queue length	Throughput	Wait of served customers	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	$99.0\% \pm 0.48\%$

Table 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 10 independent simulation runs, each of 3,000 time units, with the first 1,000 time units serving as a warm-up period.[1] The 95% confidence intervals, calculated using the t-distribution with nine degrees of freedom, are also given.

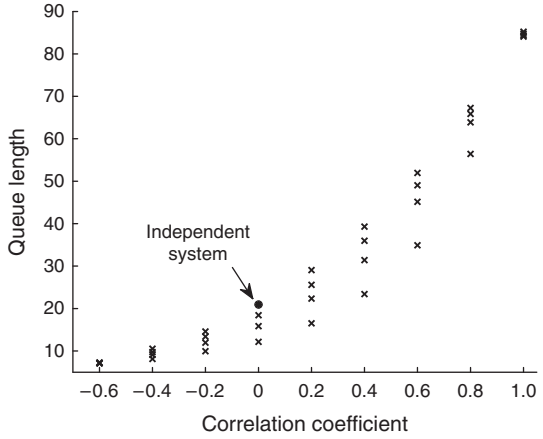
Observe that, under positive correlation, the expected queue length and expected offered wait (defined as the average time that an infinitely patient customer would wait before entering service) are approximately twice as large as those in the independent case and four times as large as those in the 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, for example, Garnett et al. (2002).

The reason for the substantial differences between the above performance metrics under different correlations 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. We conclude that *even moderate correlation can substantially affect the system performance*, and therefore staffing decisions, as a result of 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, the knowledge of the correlation coefficient between those latter two distributions is sufficient to determine the performance. To answer this question, we perform another simulation study in which we consider systems having the same correlation between the service time and patience but differing in the corresponding joint distributions. Specifically, we 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. We 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 for more details. The simulated steady-state expected queue length for each of the 36 systems is shown in Figure 1.

We make two important observations: 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.

Figure 1. Simulated Expected Queue Length for Different Joint Distributions



We conclude that *the correlation coefficient is not a sufficient statistic to determine system performance*.

Second, we find that we 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, we consider a many-server queueing system with a single pool of statistically homogeneous agents. We 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 we 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. (We 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, *we 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 were 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.

We note that the dependence may also have substantial impacts on critically loaded systems in some cases, because a nonnegligible proportion of the customers is delayed in queue. Since our analysis is motivated by asymptotic considerations and, in particular, by a weak law of large numbers (which we 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 1. For applications in which those stochastic fluctuations are nevertheless significant (because the systems is not sufficiently large), we propose a heuristic refinement in Section 6.3.

1.1. Main Goals and Contribution

Goals. In this work we 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 1 shows, quantifying the impact of the dependence on the queueing system requires more refined measures of dependency than simple correlation.

To this end, we 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 we 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 from 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), we 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 Section 4.1), so that the dependence structure and its impact on the fluid model can be studied.

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

I. We 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, the waiting time of served customers) can be easily computed. We demonstrate via simulation experiments that our fluid model is an effective and accurate approximation. See Section 4 for our fluid model.

II. We 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, we 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, we compare systems differing from each other only by the dependence structure. To this end, we rank the "strength" of the dependence by utilizing the positive quadrant dependence (PQD) stochastic order; see Section 3 for a background of PQD order and Section 5 for performance analysis.

III. We 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, we compute the fluid-optimal staffing, as well as provide structural results regarding how the dependence affects that optimal staffing. In addition, on the basis of our fluid analysis, we 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 Section 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. We therefore demonstrate that our 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, we focus on the class of Gaussian copulas (see Section 3.2 for details) whose relative tractability makes them attractive, and thus prevalent, in modeling.

2. Related Literature

Related Queueing Models. As was mentioned above, the fluid model we 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 et al. (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 it is used 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. 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 (2018) 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. (2017)

analyze an $M/M(f)/n$ queueing model (with no abandonment) in which service times are exponentially distributed with a mean that 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. We, on the other hand, consider large and overloaded systems.

Bivariate Stochastic Order and Copulas. Recall that one of the goals in this paper is to compare and rank systems having identical marginal distributions of service and patience times but different dependence structures. To this end, we employ the PQD order mentioned above and copulas. We refer to Scarsini and Shaked (1996) and Shaked and Shanthikumar (2007) for 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.

3. Measures of Dependence

In this section, we describe the measures of dependence that we will use in this paper. We provide more details in Appendix A. Let S and T be two random variables with a 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.

We consider the set of all bivariate distributions with the same marginal densities f_S and f_T, which we 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 this set is not empty It can be shown that there are many other joint distributions in this set; see Section 3.2.) We first employ a stochastic order, introduced in Section 3.1, to rank the strength of the dependence of the elements in $\mathcal{F}(f_S, f_T)$. We then discuss how to use copulas to represent joint distributions in Section 3.2.

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 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 $\mathbb{P}(X_1 > x_1, X_2 > x_2) \leq \mathbb{P}(X_1 > x_1) \cdot \mathbb{P}(X_2 > x_2)$ 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, for example, Shaked and Shanthikumar (2007, chap. 9).

Definition 1 (PQD Order)**.** For random vectors (X_1, X_2) with a joint cumulative distribution function (cdf) G and (Y_1, Y_2) with a joint cdf H, suppose that G and H have the same marginal cdfs F_1 and F_2. We say that (X_1, X_2) is smaller than (Y_1, Y_2) in the PQD order, denoted by $(X_1, X_2) \leq_{\text{PQD}} (Y_1, Y_2)$, if

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

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

We can analogously define NQD order by switching the inequalities between G and H—namely, $(X_1, X_2) \leq_{\text{NQD}} (Y_1, Y_2)$ 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 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.

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

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 cdfs $F_1, \ldots, F_d$, is a joint cdf on the unit cube $[0,1]^d$ with uniformly distributed marginals, such that

$$C(F_1(x_1), \ldots, F_d(x_d)) = F(x_1, \ldots, x_d) \quad \text{for all } x_1, \ldots, x_d,\ d \geq 2. \tag{1}$$

By Sklar's theorem (e.g., Clemen and Reilly 1999, section 2), a copula exists uniquely for any given joint cdf F if the marginals are continuous (as we 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 (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, we could choose any marginal distributions for S and T, and we 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 our 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 such as, for example, the Gaussian copula, t-copula, and various Archimedean copulas (e.g., Frank, Joe, AMH, and Gumbel copulas). Because of its tractability, the Gaussian copula plays a fundamental role in modeling dependent distributions. We will therefore focus on this class of copulas and demonstrate how our results translate to the corresponding joint distributions.

We 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 the 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 we 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.[2] Moreover, Lemma 3 in Appendix A.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.

4. Model

We consider a multiserver 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. We 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 our 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, we assume that $\{(S_i, T_i)\colon i \geq 1\}$ 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. We use S and T to denote generic random variables having joint density f, and marginals f_S and f_T. We 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. We 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*. We consider overloaded systems in which the arrival rate is larger than the total service capacity and thus a nonnegligible fraction of customers abandon the system. It will be shown in Proposition 2 that if $\lambda > s\mu$, or equivalently, $\rho > 1$, then the system is overloaded for any joint distribution f.

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. We therefore employ a deterministic fluid model, as in Whitt (2006a) and Bassamboo and Randhawa (2015), to approximate the stationary queueing system, and we demonstrate the effectiveness of that fluid

model via simulations. To construct the fluid model, we replace the stochastic arrival, service, and abandonment processes by corresponding deterministic flows. In particular, we 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 our setting, the workload in the system depends on the waiting time; to characterize it, we define the *work evolution function*,

$$\phi(w) := \int_w^\infty \int_0^\infty x f(x,y)\, dx\, dy, \tag{2}$$

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 $\phi(w) = F_T^c(w) \cdot \mathbb{E}[S \mid T > w]$, where $F_T^c := 1 - F_T$ is the proportion of fluid that remains in queue after waiting w time units, and $\mathbb{E}[S \mid T > w]$ 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:

$$\lambda \phi(\bar{w}) = s. \tag{3}$$

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

Proposition 1. *If $\rho > 1$, then there exists a unique $\bar{w} > 0$ that solves Equation (3).*

We refer to the unique solution $\bar{w}$ to (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, we 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; that is,

$$a(w) := \mathbb{E}[S \mid T > w]. \tag{4}$$

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}, we can characterize the *effective traffic intensity* to the system:

$$\rho_{\text{eff}} := \frac{\lambda}{s\mu_{\text{eff}}}. \tag{5}$$

Next, dividing both sides of the equality in (3) by $s\mu$ gives

$$\rho \phi(\bar{w}) = 1/\mu. \tag{6}$$

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 cdf F_T of patience time, we see that (6) can be represented via

$$\rho F_T^c(\bar{w}) = \frac{\mu_{\text{eff}}}{\mu}, \quad \text{or equivalently, } \frac{\lambda F_T^c(\bar{w})}{s} = \mu_{\text{eff}}. \tag{7}$$

The first equality in (7) is a generalization of equation (3.9) in Whitt (2006a), which states that $\rho F_T^c(w) = 1$ in the independent model. The second equality in (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}.

We note that when S and T are positively dependent, $a(w) = \mathbb{E}(S \mid T > w)$ 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, we compute the throughput and stationary fluid queue, which we denote by R and Q, respectively. Clearly, we have $R = s\mu_{\text{eff}}$, so that

$$R = s\mu_{\text{eff}} = s\mu\rho F_T^c(\bar{w}) = \lambda F_T^c(\bar{w}), \tag{8}$$

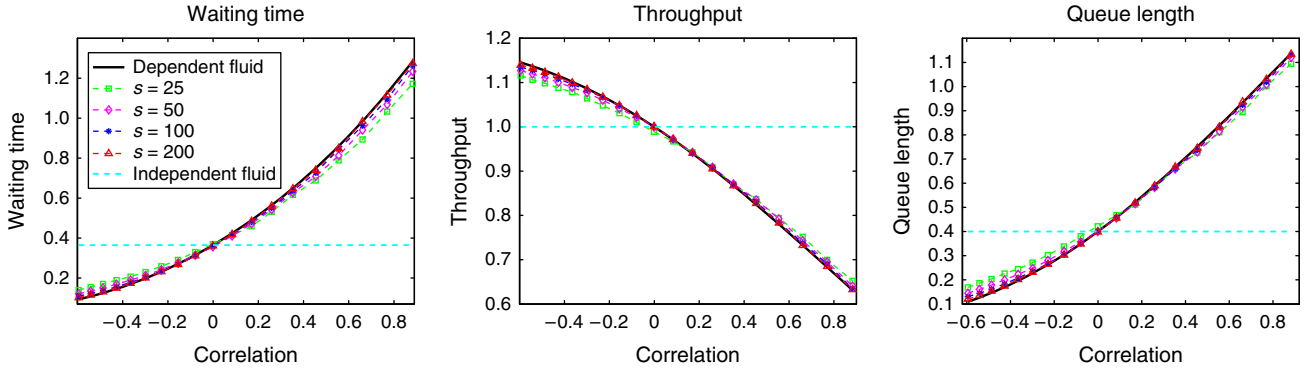
where the second equality follows from (7). The expression for the steady-state fluid queue length Q is derived as follows: the amount of fluid that enters the queue over an interval $[t, t + dx)$ is λdx, and the proportion of that fluid remaining in the queue t time units later after arrival is $F_T^c(t)$. Since all arriving fluid that is served waits exactly $\bar{w}$, it holds that

$$Q = \lambda \int_0^{\bar{w}} F_T^c(x)\, dx. \tag{9}$$

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 $\mathcal{D}$, we can compute the offered wait $\bar{w}$ via (3), from which a_{eff} and μ_{eff} can be easily recovered via (7). Given these latter two variables, we can compute the stationary throughput R in (8) and fluid queue Q in (9).

Note that in an overloaded system—that is, with $\bar{w} > 0$ (Proposition 1)—our fluid model captures "predictable" queueing effects, which are due to insufficient

Figure 2. (Color online) Simulation and Fluid Model Under Different System Sizes and Dependencies, $\rho = \lambda/s\mu = 1.2$, $s \in \{25, 50, 100, 200\}$



Notes. 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 a Gaussian copula.)

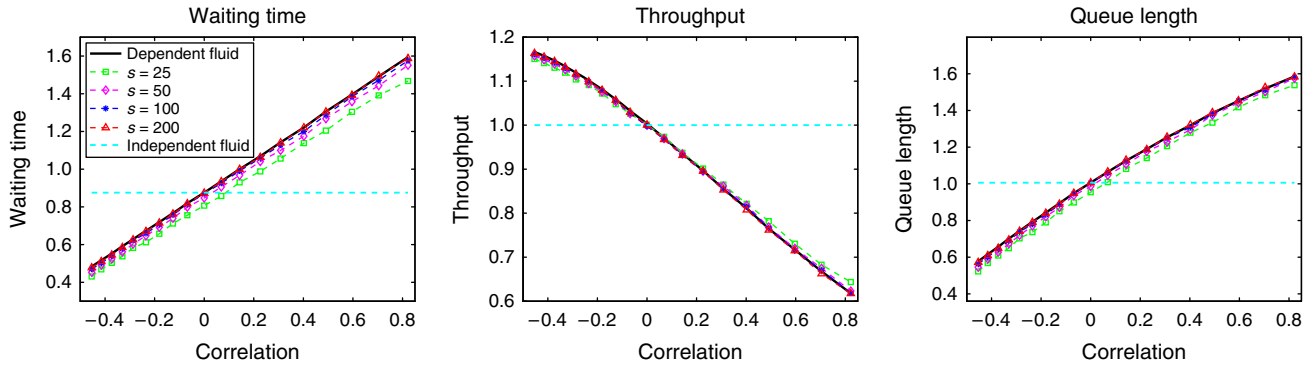
service capacity. This is different from non-overloaded systems, in which queueing is due to stochasticity associated with the arrival and service process. Specifically, the fluid model does not capture queueing effects that are due to random fluctuations. The following remark elaborates on this point from an asymptotic perspective.

Remark 1. Even though we 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 that is distributed as the stationary queue in the s system, we conjecture that $Q_s(\infty)/s$ converges in distribution to Q in (9) and that a similar result holds for the stationary distribution of the service process. In particular, we expect our fluid model to become more accurate as the size of the system increases, although our simulation experiments (depicted in Figures 2 and 3) 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 $\rho \le 1$). See also Proposition 2.

4.2. Numerical Examples

We now examine the accuracy of the fluid approximation for overloaded systems via simulation. To conduct the numerical experiments, we 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 we consider. In the first numerical study, depicted in Figure 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

Figure 3. (Color online) Simulation and Fluid Model Under Different System Sizes and Dependencies, $\rho = \lambda/s\mu = 1.2$, $s \in \{25, 50, 100, 200\}$



Notes. Interarrival time distribution *Erlang*$(2, 2\lambda)$, service time distribution LN(1, 2), and patience time distribution LN(2, 2). (The joint distribution of service time and patience time is generated via a Gaussian copula.)

from the exponential assumption, we perform another numerical study, depicted in Figure 3, in which we 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 we 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 we plot the simulated average waiting time of *served customers*, the average throughput, and average queue length in steady state, and we compare those simulation results (curves indicated by the number of agents s) to the corresponding fluid estimates (the "Dependent fluid" curves). The throughput and queue length are both plotted scaled by the number of agents s. To compare the result to the independent model, we 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. Furthermore, as was already demonstrated in Section 1, the independent model does not give useful approximations even for systems with moderate dependence.

We next demonstrate how the effective traffic intensity ρ_{eff} in (5) changes with the nominal traffic intensity ρ and the joint distribution f; the results are shown in Table 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 $r = -0.4$, 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 abandon), but only about 7.4% (a proportion $(1.08 - 1)/1.08$) end up abandoning in our example with negative correlation. By 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 our example with positive correlation. We study the economic aspect of our results in Section 6 in the context of optimal staffing.

Table 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

5. Performance Analysis

Recall that the fluid model is fully characterized by the primitive data set $\mathcal{D} = (\lambda, s, f)$. In this section, we 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 Section 5.1 we 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 Section 5.2 we quantify how the throughput is impacted by the dependence structure, employing the PQD order and Gaussian copula discussed in Sections 3.1 and 3.2. To this end, we fix λ and s and the two marginal densities f_S and f_T, and we vary the joint distribution f.

However, we 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 $\bar{w} > 0$.) We 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 $\rho \le 1$ 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 $\rho \le 1$.

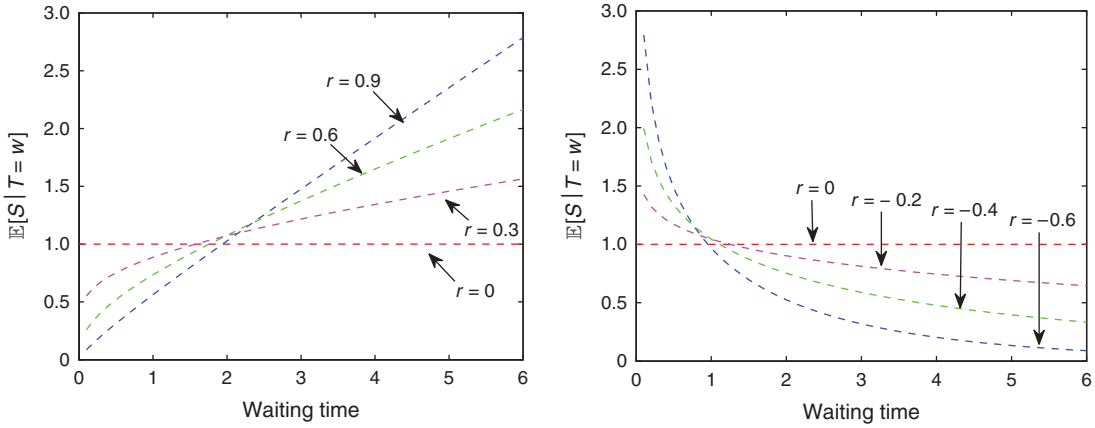
Proposition 2. *The following three statements are equivalent*:

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

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

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

Figure 4. (Color online) Conditional Service Time Under Different Distributions Generated by Gaussian Copula



Notes. Positive dependence (left), $r > 0$ and ICST. Negative dependence (right), $r < 0$ and DCST. Independent case, $r = 0$ and CCST.

Throughout this section we assume that $\rho > 1$. Let

$$g(w) := \mathbb{E}[S \mid T = w]. \tag{10}$$

We 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.3 we provide natural sufficient conditions for monotone conditional service time (MCST), and we link the monotonicity of the CST to PQD and Gaussian copula introduced in Section 3. In particular, Lemma 5 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 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.

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

We 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 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 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 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 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 we will show in Section 6.

The following corollary follows immediately from Lemma 5 and Proposition 3.

Corollary 1. *For $(S, T) \in \mathcal{G}$, $R(\lambda)$ is decreasing if $r > 0$, and $R(\lambda)$ is increasing if $r < 0$.*

We 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 $0 \le s < \lambda/\mu$—namely, the domain of $R(s)$ is $[0, \lambda/\mu)$.

Proposition 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 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,

Table 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	Absolute	Percentage	Simulation	Fluid	Absolute	Percentage
100	1	98.01	100.00	1.99	2.03	94.34	100.00	5.66	6.00
105	1.05	101.66	103.20	1.55	1.52	93.09	93.44	0.35	0.38
110	1.1	104.82	106.00	1.18	1.13	90.08	89.79	0.29	0.33
120	1.2	110.26	110.96	0.70	0.63	84.78	84.75	0.03	0.03
130	1.3	114.89	115.37	0.48	0.41	81.17	81.16	0.01	0.01
150	1.5	122.75	123.08	0.33	0.27	76.12	76.11	0.01	0.01

the relation between the throughput and the capacity is linear. The structural properties of $R(s)$, stated in Proposition 4, facilitate the analysis of the staffing problem for revenue maximization in Section 6.

The intuition behind the fact that the throughput grows at a rate faster or 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, we obtain the following corollary to Proposition 4.

Corollary 2. *For $(S,T) \in \mathcal{G}$, $R(s)$ is convex increasing if $r > 0$, and $R(s)$ is concave increasing if $r < 0$.*

5.2. Impact of Dependence Between Service and Patience on Performance

We now consider how the strength of the dependence, as ranked by PQD order, impacts system performance. To this end, we 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 Section 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 5. *If $(S_1,T_1) \leq_{\text{PQD}} (S_2,T_2)$, 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 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 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 $Q_1 < Q_2$.*

5.3. Numerical Examples

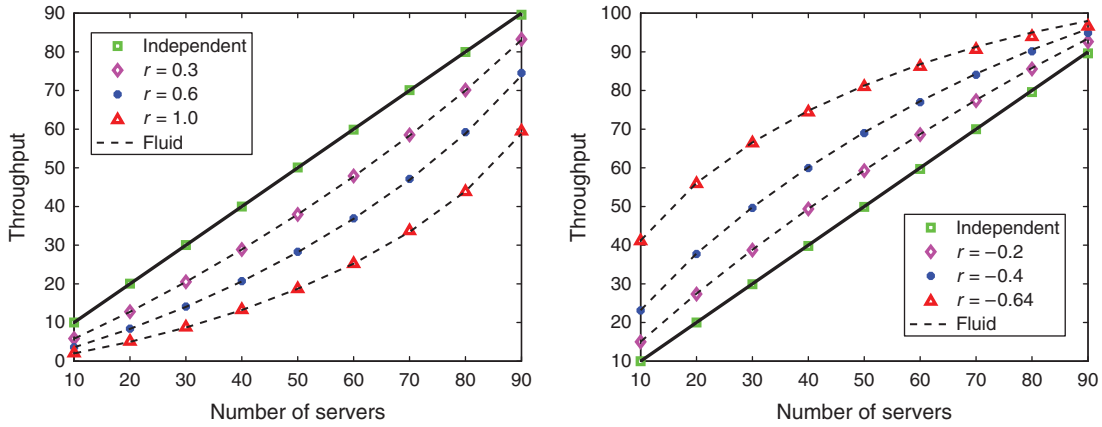
We first demonstrate the statement of Proposition 3 and Corollary 1. In Table 3 we compare the throughput of different systems where the capacity s is fixed at 100 and the nominal traffic intensity ρ increases from 1.0 to 1.5 as the arrival rate λ varies. We also compare the throughput calculated by our 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. We remind the reader that for any $\rho \geq 1$, the throughput is fixed at $s\mu = 100$ when S and T are independent.

We 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 1), play a dominant role when the fluid estimate is zero for the system. We consider the staffing problem in the next section and propose a heuristic refinement that is based on diffusion approximations for critically loaded systems.

We next validate the result in Proposition 4. Figure 5 compares the throughputs obtained from simulations (discrete marks) and from fluid models (dashed line) under different capacities and joint distributions. We 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 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).

Finally, we numerically validate the result in Proposition 5. We take S and T as in the former two examples,

Figure 5. (Color online) A Comparison of Throughputs Under Different Capacities ($\lambda = 100$, s Ranges from 10 to 90)



Notes. Positive dependence (left): throughput convex increasing with s. Negative dependence (right): throughput concave increasing with s. The independent case (solid lines with squares): throughput is linear increasing with s.

and again we employ a Gaussian copula to generate their joint distribution. We fix $\lambda = 120$ and $s = 100$, and we plot the throughput as a function of the correlation coefficient r. Figure 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 $r = 1$.

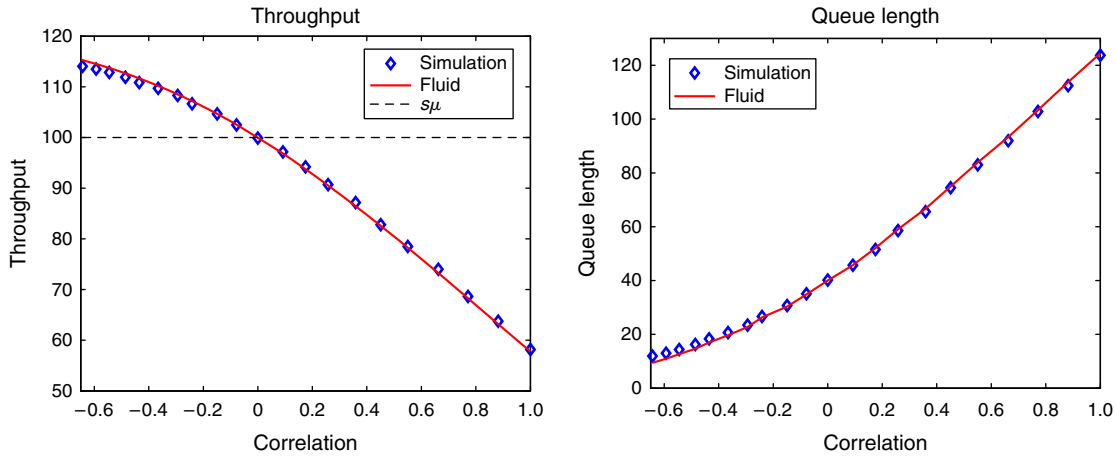
6. Economics of Capacity Sizing

In this section we apply the results derived for the stationary fluid model to develop fluid-optimal solutions to a capacity-sizing problem under a linear cost structure. We start in Section 6.1 by considering the optimal staffing under the first-in-first-out (FIFO) policy.[3] It is significant that the analysis we apply was performed for overloaded systems having $\rho > 1$ (recall Proposition 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 6. In the latter case, our fluid model is too crude an approximation for the stochastic system (since the queue and thus the proportion of abandonment are both null in the fluid model of critically loaded systems), and stochastic refinements must be considered. Thus, in Section 6.3 we 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.

6.1. Capacity Sizing Under FIFO Policy

We 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

Figure 6. (Color online) A Comparison of Throughputs and Queue Lengths for Different Systems with Service and Patience Times Generated by Gaussian Copulas



arrival rate λ, we consider the following cost optimization problem for the fluid system:

$$\min_{s\geq 0} C_\lambda(s) := cs + p\alpha_\lambda(s), \tag{11}$$

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.

Equivalent to (11), we 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 (11) is equivalent to

$$\min_{0\leq s\leq \lambda/\mu} C_\lambda(s) = cs + p\alpha_\lambda(s). \tag{12}$$

Note that $\alpha_\lambda(s) = \lambda F_T(w)$, where w solves (3), so that

$$\begin{aligned} C_\lambda(s) &= cs + p\alpha_\lambda(s) = \lambda[c\phi(w) + pF_T(w)] \\ &= \lambda[cF^c_T(w)a(w) + pF_T(w)]. \end{aligned}$$

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

$$\min_{w\geq 0} \bar{C}_\lambda(w) := cF^c_T(w)a(w) + pF_T(w). \tag{13}$$

Differentiating $\bar{C}_\lambda(w)$ with respect to w gives $\bar{C}'_\lambda(w) = f_T(w)(p - cg(w))$, and setting the derivative to 0 gives us the following first-order condition: $g(w) = p/c$. To interpret the latter equality, note that $p/g(w)$ represents the marginal revenue of adding capacity; in optimality, this marginal revenue must equal the marginal cost c of added capacity. The above derivation gives rise to the following proposition. Let $g(\infty)$ denote the limit of $g(w)$ as $w \to \infty$, whenever the limit exists.

Proposition 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.*

We remark that the conditions in the second part of the proposition are always satisfied when (S, T) is generated by a Gaussian copula 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 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 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 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 4. *If $(S_1, T_1) \leq_{\text{PQD}} (S_2, T_2)$, then $C^*_1 \leq C^*_2$.*

It follows from Corollary 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 4. 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, we need the following definition. Let h be a real-valued function. We 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.

Table 4. Optimal Staffing of Systems with Dependencies ($\lambda = 100$)

Correlation	Fluid optimal		Simulation optimal		Optimality gap	
r	Capacity	Cost	Capacity	Cost	Absolute	Percentage
			$p/c = 0.8$			
−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
			$p/c = 1.25$			
−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
			$p/c = 3.5$			
−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

Consider the setting of Corollary 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 (10), $i = 1, 2$.

Proposition 7. *Assume that* (i) $(S_1, T_1) \leq_{\text{PQD}} (S_2, T_2)$; (ii) f_i *has a DCST,* $i = 1, 2$; *and* (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* $p/c \in (M, g_2(0))$.

We note that condition (iii) in Proposition 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 $h'(0) \neq 0$; or (3) h admits a Taylor series expansion at 0.

If the bivariates are generated via a Gaussian copula, the monotonicity of the optimal staffing in Proposition 7 is strict, as stated in the following corollary.

Corollary 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* $p/c > M$.

6.2. Numerical Study

We now present numerical and simulation examples to demonstrate the accuracy and the limitations of the optimal fluid solution to problem (11) described in Proposition 6. The system we 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, we fix $c = 1$ and vary the penalty p; in particular, we 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 4 we 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, we 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. Second, we 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 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 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 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 paper. We mention that extensive simulation experiments suggest that the safety capacity needed 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, Garnett et al. 2002). In the next section we propose an algorithm to compute effective staffing recommendations for critically loaded systems with dependencies that are based on our characterization of the effective service rate combined with existing results for the independent model.

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, we 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 Garnett et al.: For a service system with s agents, define $\beta = (s - \lambda/\mu)/\sqrt{\lambda/\mu}$. Then the abandonment ratio can be approximated by

$$\mathbb{P}(\mathrm{Ab}) \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\mu/\theta})}\right]\left[1 + \frac{h(\beta\sqrt{\mu/\theta})}{\sqrt{\mu/\theta}\,h(-\beta)}\right]^{-1}, \tag{14}$$

where h is the hazard function of the standard normal random variable. Approximate the abandonment rate $\alpha_\lambda(s) = \lambda \cdot \mathbb{P}(\mathrm{Ab})$ and compute the optimal staffing level s_0 that solves (11) (without dependence). Let $\mathbb{P}^*(\mathrm{Ab})$ 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 $\mathbb{P}^*(\mathrm{Ab})$ computed in (i)—namely, for which $F_T(w^*) = \mathbb{P}^*(\mathrm{Ab})$. Compute the effective service rate $\mu^*_{\mathrm{eff}} = 1/a(w^*)$.

(iii) Employ the approximation in (14) once again, this time with service rate μ^*_{eff}, and compute the capacity s^* for which the proportion of abandonment is equal to $\mathbb{P}^*(\mathrm{Ab})$.

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 5 demonstrates the substantial improvements obtained by employing the above procedure. In this table, the capacity and resulting cost obtained using our 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 our heuristic when $p = 3.5$ and $r = 1.0$, compared with 26% under the fluid prescription.

Table 5. Optimal Staffing of Systems with Dependencies: Simulations, Fluid Prescriptions, and Heuristic ($\lambda = 100, c = 1, p = 3.5$)

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

7. Summary

We 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 analysis intractable even if the marginal service time and patience are exponentially distributed, we 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, we applied the fundamental PQD stochastic order and the CST to obtain structural results regarding the impact of the dependence on the fluid model. Our qualitative results were shown to hold for the important family of Gaussian copulas, which is often employed in practice to analyze joint distributions because of its analytical tractability.

We then implemented the framework we 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, we 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 our 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; for example, 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.

Acknowledgments

This research was conducted while the first author was a Ph.D. student at Northwestern University. The authors thank the department editor, associate editor, and three anonymous reviewers for their careful reading of the paper, and for providing constructive feedback. The authors also thank Barry Nelson for helpful discussions.

Appendix A. More on Copulas and Conditional Service Time

A.1. Generating Gaussian Copulas and t-Copulas

We now provide details on how to generate dependent bivariates (S, T) via a Gaussian copula and t-copulas. The procedures we describe below are used to generate different joint distributions that correspond to Figure 1.

Let r_G be a number in $[-1, 1]$, and let $\Phi(\cdot)$ denote the cdf of the 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; we 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, we build on the following lemma; see theorem 2 of Cario and

Nelson (1997) for its proof and for further details. Let $\underline{r}$ and $\bar{r}$ be the minimal and maximal attainable correlation coefficients of S and T, respectively. (Note that $\underline{r}$ may be larger than -1 and $\bar{r}$ may be smaller than 1; for example, if S and T are both exponential random variables, then $\underline{r} \approx -0.64$.)

Lemma 2. *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 $r\colon x \mapsto [\underline{r}, \bar{r}]$ is strictly increasing, with $r(-1) = \underline{r}$ and $r(1) = \bar{r}$.*

In particular, the minimal and maximal attainable correlation between two marginal distributions can be generated via NORTA. Moreover, because of the monotonicity of $r(x)$ and its inverse, it is easy to find the value of r_G that gives any prespecified 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 the Gaussian copula are independent if and only if they are uncorrelated.

We next describe the procedure proposed in Demarta and McNeil (2005) 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.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. We 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—that is, 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 nonparametric 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. We therefore have the following lemma.

Lemma 3. *If for $(S_1,T_1),(S_2,T_2) \in \mathcal{G}$ it holds that $r_1 < r_2$, then $(S_1,T_1) \leq_{\mathrm{PQD}} (S_2,T_2)$.*

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.3. Relating the CST, PQD Order, and Gaussian Copula Together

The following two lemmas provide natural sufficient conditions for MCST, and they link the monotonicity of the CST to PQD and Gaussian copula.

Lemma 4. *If $\mathbb{P}(S > u \mid T = w)$ is increasing in w, then (S,T) is PQD and has an ICST. If $\mathbb{P}(S > u \mid T = w)$ is decreasing in w, then (S,T) is NQD and has a DCST.*

Lemma 4 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 5. *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*; *and* (iii) *if $r = 0$, then (S,T) has a CCST.*

Appendix B. Time to Stationarity

In general, many-server queueing systems in heavy traffic tend to converge to stationarity much faster than single-server systems; see, for example, the discussion in E.C.1 in Perry and Whitt (2009). We now demonstrate via simulations that our system with dependence indeed converges quickly to its stationary behavior. We simulate systems with and without dependence, starting the systems at two extreme initial conditions; the systems in Figure B.1(a) are initialized

Figure B.1. (Color online) Convergence of Queue Length of Stochastic System to Steady State

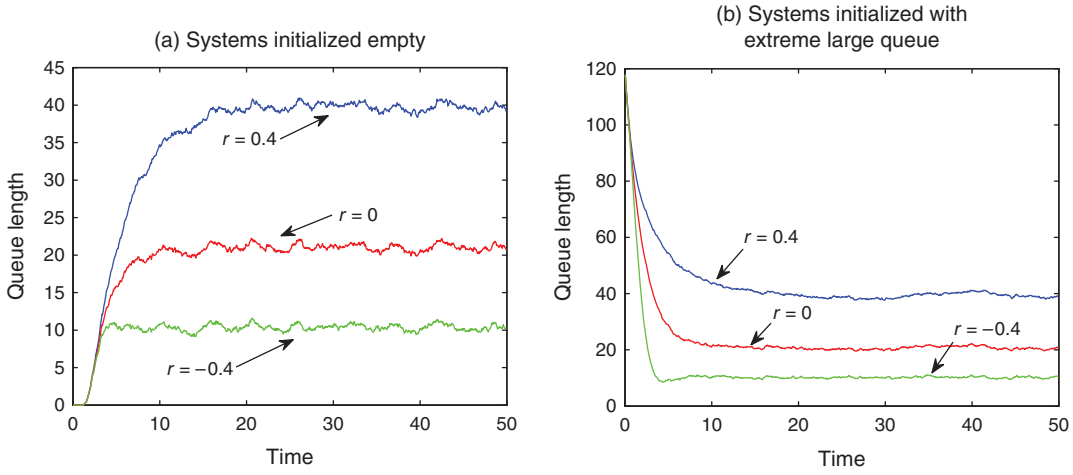


Table C.1. Optimal Staffing Under LIFO and Positive Dependence ($\lambda = 100$)

	$p/c = 1.25$			$p/c = 3.5$		
	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

empty, and the initial queue length of the systems depicted in Figure B.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 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, we 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, we conjecture that the same is true for the dependent model. We further remark that we consider extreme initial conditions 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 B.1.

Appendix C. Capacity Sizing Under a Throughput-Maximizing Policy

In this appendix, we 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 8. *The throughput-maximizing policy is FIFO if f has a DCST, and is last-in-first-out (LIFO) if f has an ICST. If f has a CCST, any nonidling policy yields the same throughput.*

In particular, since congestion is beneficial when the dependence is negative, we 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 we cannot identify those customers upon arrival, the best we 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 Gaussian copulas, the conditions on MCST in Proposition 8 reduce to a condition on the sign of the correlation coefficient.

Corollary 6. *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 Section 6.1 regarding the optimal capacity under a negative dependence still applies, because FIFO is the optimal policy in this case. Hence we 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. We conclude that when the throughput-maximizing control policy is adopted, the capacity prescribed in Proposition C.1 remains optimal. Numerical studies for the system considered in Section 6.3, presented in Table C.1, show that the fluid-optimal capacity is fairly accurate under the throughput-maximizing policy.

In ending, we remark that in overloaded systems, customers will be left to wait with no chance of ever entering service if LIFO is employed, and so it is infeasible to employ in observable service systems. Nevertheless, from the fluid perspective, we can achieve the same throughput by employing an admission control policy that 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.

Appendix D. Proofs

D.1. Auxiliary Results

Before presenting the proofs of the results in the paper, we state two auxiliary results that will be employed in our proofs below.

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

Lemma 6. *If $(S_1, T_1) \leq_{\text{PQD}} (S_2, T_2)$, then $\mathbb{E}(S_1 \mid T_1 \leq z) \geq \mathbb{E}(S_2 \mid T_2 \leq z)$ and $\mathbb{E}(S_1 \mid T_1 > z) \leq \mathbb{E}(S_2 \mid T_2 > z)$ for all $z \geq 0$.*

For the next auxiliary result, whose statement follows easily from (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 $\phi(\cdot)$, we obtain the following lemma.

Lemma 7. *We have that $w(s)$ is strictly decreasing on $(0, \lambda/\mu)$.*

D.2. Proofs of the Main Results in the Paper

We now prove the main results (propositions and corollaries) in the paper in the order in which they appear.

Proof of Proposition 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 (3). □

Proof of Proposition 2. We start by showing that $\bar{w} > 0$ if and only if $\rho > 0$. First, it follows immediately from the fact that $\phi(w) \le \phi(0) = \mathbb{E}[S] = 1/\mu$ for all $w \ge 0$, so that (6) is not well defined when $\rho \le 1$. In particular, there exists no overload equilibrium for the fluid model in this case.

To prove the other direction, we assume that $\rho > 1$ and make the contradictory assumption that $\bar{w} = 0$. It then follows from (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 (7). Thus, it must hold that $\bar{w} > 0$.

We next prove that $\rho > 1$ if and only if $\rho_{\text{eff}} > 1$. To this end, observe that, by (5), $\rho_{\text{eff}} \le 1$ is equivalent to $s\mu_{\text{eff}} \ge \lambda$, which, together with the second equality in (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 (5), $\rho_{\text{eff}} > 1$ implies that $\mu_{\text{eff}} < \lambda/s = \rho\mu$. It then follows from the first equality in (7) that $\bar{w} > 0$, and thus $\rho > 1$. □

Proof of Proposition 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 (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). □

Proof of Corollary 1. Corollary 1 follows from Proposition 3 and Lemma 5. □

Proof of Proposition 4. The offered wait w solving (3) is a function of s, which we 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 (3) with respect to s gives $-\lambda \int_0^\infty x f(x, w(s))\,dx \cdot w'(s) = 1$, so that

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

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 (D.1) gives

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

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. □

Proof of Corollary 2. Corollary 2 follows from Proposition 4 and Lemma 5. □

Proof of Proposition 5. It suffices to prove that $w_1 \le w_2$, because the stated inequalities for R_i and Q_i, $i = 1, 2$, will follow immediately from (8) and (9) and the fact that T_1 and T_2 have the same marginal cdf F_T. To this end, we will prove that the following inequality holds for ϕ in (2):

$$\phi_1(z) \le \phi_2(z), \quad \text{for all } z \ge 0. \tag{D.2}$$

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

It remains to show that (D.2) holds. Note that $\psi_z(s,t) = (s\mathbf{1}\{t > z\})$ 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 of Shaked and Shanthikumar 2007, p. 395), $(S_1, T_1) \le_{\text{PQD}} (S_2, T_2)$ implies $\mathbb{E}[\psi_z(S_1, T_1)] \le \mathbb{E}[\psi_z(S_2, T_2)]$; that is, $\phi_1(z) \le \phi_2(z)$. □

Proof of Corollary 3. The statement of the corollary follows from the fact that the inequality in (D.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

$$\begin{aligned}\mathbb{E}(S_i \mid T_i > z) &= \int_0^\infty \mathbb{P}(S_i > u \mid T_i > z)\,du \\ &= \frac{\int_0^\infty \mathbb{P}(S_i > u, T_i > z)\,du}{F_T^c(z)}.\end{aligned} \tag{D.3}$$

In the proof of Lemma 3 we 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 (D.3) that $\mathbb{E}(S_1 \mid T_1 > z) < \mathbb{E}(S_2 \mid T_2 > z)$ for all $z > 0$. Hence, $w_1 < w_2$, implying that $R_1 > R_2$ and $Q_1 < Q_2$. □

Proof of Proposition 6. If g is increasing, then by Proposition 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, we optimize the cost function $\bar{C}_\lambda(s)$ in (12). Or equivalently, we minimize $\bar{C}_\lambda(w)$ in (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 (7): $s_\lambda^* = \lambda F_T^c(w^*)a(w^*)$. If $p/c \ge g(0)$, then $\bar{C}'_\lambda(w) \ge 0$ for all w, so that $w^* = 0$, and $s_\lambda^* = \lambda/\mu$ is fluid optimal. If $p/c \le g(\infty)$, then $\bar{C}'_\lambda(w) \le 0$ for all w; hence $w^* = \infty$ and $s_\lambda^* = 0$. □

Proof of Corollary 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 system with service time and patience time (S_i, T_i). Then,

$$\begin{aligned}C_2^* &= p\lambda + cs_2^* - pR_2(s_2^*) \ge p\lambda + cs_2^* - pR_1(s_2^*) \\ &\ge p\lambda + cs_1^* - pR_1(s_1^*) = C_1^*,\end{aligned}$$

where the first inequality follows from Proposition 5 and the second inequality follows from the optimality of s_1^* for a system with service and patience time (S_1, T_1). □

Proof of Proposition 7. The first-order condition (13) of the capacity optimization problem gives

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

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

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

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

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

Since T_1 and T_2 have the same marginal cdf F_T, $\mathbb{E}(S_1 \mid T_1 \leq \delta) < \mathbb{E}(S_2 \mid T_2 \leq \delta)$, contradicting Lemma 6. Hence, (D.4) must hold.

We 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

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

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 7 and the second inequality follows from Proposition 5, contradicting (D.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, we consider the cases $h(0) > 0$ 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 because of 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.)

We 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, we 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 $t_0 = \inf\{z > 0\colon h(z) = 0\}$, where $\inf(\varnothing) := \infty$. We first claim that $t_0 > 0$. Indeed, if $t_0 = 0$, then there must exist a positive sequence $\{z_n\colon n \geq 1\}$ 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 $z = 0$. We therefore have $t_0 > 0$. We next show that 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 $\mathbb{E}(S_1 \mid T_1 \leq \delta) < \mathbb{E}(S_2 \mid T_2 \leq \delta)$ for all $\delta > 0$, a contradiction to Lemma 6, we 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 $z \in [0, t_0]$. □

Proof of Corollary 5. If $(S_1, T_1), (S_2, T_2) \in \mathcal{G}(f_S, f_T)$ satisfy $r_1 < r_2 < 0$, 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 7 gives $t_0 < \infty$. For all $z \in (0, t_0)$, we 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 7 can be used to show that $s_1^* < s_2^*$. □

D.3. Proofs of the Lemmas in the Paper

Proof of Lemma 3. As demonstrated in Appendix A.1,

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

where $\stackrel{d}{=}$ 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 2 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 $(S_1, T_1) \leq_{\text{PQD}} (S_2, T_2)$. By theorem 9.A.1 of Shaked and Shanthikumar (2007, p. 390), PQD ordering is preserved under component-wise 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 $(\Gamma_1, \Xi_1) \leq_{\text{PQD}} (\Gamma_2, \Xi_2)$. 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), and (2) association ordering implies PQD ordering (Shaked and Shanthikumar 2007, p. 417, theorem 9.E.2).

We now prove a stronger version of the lemma, which we employ in the proof of Corollary 3, requiring a strict form of the PQD order; in particular, we 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, we 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 2 that the cdfs $\{F \in \mathcal{G}\}$ can be indexed by the correlation coefficient r of (S, T). Moreover, again by Lemma 2, there exists a bijection mapping from the cdfs in $\mathcal{G}$ to the family of bivariate normal random variables with a zero mean vector indexed by their correlation coefficient r_G. Thus, we can equivalently parameterize the elements $\{F \in \mathcal{G}\}$ by the correlation coefficient r_G of the underlying bivariate normal random variables, and we show that $\{F_{r_G}(x_1, y_1)\colon -1 \leq r_G \leq 1\} \equiv \{F_r(x_1, y_1)\colon \underline{r} \leq r \leq \bar{r}\}$ is increasing in r_G, and thus in r, for all $x_1, y_1 \geq 0$. (Recall that $\underline{r}$ and $\bar{r}$ denote the minimal and maximal attainable correlation coefficients of S and T, respectively; see Appendix A.1.)

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

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

For Φ and ϕ denoting the cdf and probability density function (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, $\Gamma \stackrel{d}{=} \gamma(S)$ and $\Xi \stackrel{d}{=} \xi(T)$ so that the joint density of (S,T) is

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

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

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

so that

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

Now,

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

Plugging (D.7) into (D.6),

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

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

Proof of Lemma 4. Note that $g(w) = \mathbb{E}[S \mid T = w] = \int_0^\infty \mathbb{P}(S > u \mid T = w)\,du$. Since $\mathbb{P}(S > u \mid T = w)$ is increasing in w and f_S is fixed, g is necessarily increasing. It remains to show that $\mathbb{P}(S > u \mid T = w)$ increasing in w implies PQD. Following Block et al. (1985), we say that (S,T) is positively dependent through stochastic ordering (PDS) if $\mathbb{P}(S > u \mid T = w)$ is increasing in w. That PDS implies PQD is proved in Block et al. (1985, p. 82). □

Proof of Lemma 5. By Lemma 4, we need to show that if $(S,T) \in \mathcal{G}$, then $\mathbb{P}(S > u \mid T = w)$ is strictly increasing in w (PDS) (see the proof of Lemma 4) 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 variable (Γ,Ξ) with correlation coefficient r_G. By Block et al. (1985, theorem 2.1), PDS is preserved under component-wise increasing transformation of random vectors. Since $r > 0$ implies $r_G > 0$ by Lemma 2, 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. □

Endnotes

[1] In fact, the convergence of the stochastic systems to the steady state is fairly fast; see Appendix B for more details.

[2] 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.

[3] The analysis in Section 6.1 is extended in Appendix C to consider the optimal control policies.

References

Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57(1):66–81.

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* 58(5):1398–1413.

Bassamboo A, Randhawa RS (2015) Scheduling homogeneous impatient customers. *Management Sci.* 62(7):2129–2147.

Bassamboo A, Randhawa RS, Zeevi A (2010) Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Sci.* 56(10):1668–1686.

Block HW, Savits TH, Shaked M (1985) A concept of negative dependence using stochastic ordering. *Statistics Probab. Lett.* 3(2):81–86.

Boxma OJ, 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, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.

Cario MC, Nelson BL (1997) Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Delphi Packard Electric Systems, Warren, OH.

Chan CW, Farias VF, Escobar GJ (2017) The impact of delays on service times in the intensive care unit. *Management Sci.* 63(7): 2049–2072.

Clemen RT, Reilly T (1999) Correlations and copulas for decision and risk analysis. *Management Sci.* 45(2):208–224.

Colangelo A, Scarsini M, Shaked M (2006) Some positive dependence stochastic orders. *J. Multivariate Anal.* 97(1):46–78.

Corbett CJ, Rajaram K (2006) A generalization of the inventory pooling effect to nonnormal dependent demand. *Manufacturing Service Oper. Management* 8(4):351–358.

Demarta S, McNeil AJ (2005) The t copula and related copulas. *Internat. Statist. Rev.* 73(1):111–129.

De Vries J, Roy D, De Koster R (2017) Worth the wait? How waiting influences customer behavior and their inclination to return. Working paper, VU University, Amsterdam.

Embrechts P, McNeil A, Straumann D (2002) Correlation and dependence in risk management: Properties and pitfalls. *Risk Management: Value at Risk and Beyond*, 1st ed. (Cambridge University Press, Cambridge, UK), 176–223.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

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

Joe H (1997) *Multivariate Models and Multivariate Dependence Concepts* (CRC Press, Boca Raton, FL).

Kang W, Ramanan K, et al. (2010) Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* 20(6):2204–2260.

Li AA, Whitt W (2014) Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. *Performance Evaluation* 80(October):82–101.

Liu Y, 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, Whitt W (2011b) A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.* 59(4):835–846.

Mak H-Y, Shen Z-JM (2014) Pooling and dependence of demand and yield in multiple-location inventory systems. *Manufacturing Service Oper. Management* 16(2):263–269.

Müller A (2000) On the waiting times in queues with dependency between interarrival and service times. *Oper. Res. Lett.* 26(1): 43–47.

Müller A, Scarsini M (2001) Stochastic comparison of random vectors with a common copula. *Math. Oper. Res.* 26(4):723–740.

Nelsen RB (2013) *An Introduction to Copulas*, Vol. 139 (Springer Science & Business Media, New York).

Pang G, Whitt W (2012) The impact of dependent service times on large-scale service systems. *Manufacturing Service Oper. Management* 14(2):262–278.

Pang G, Whitt W (2013) Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems* 73(2):119–146.

Perry O, Whitt W (2009) Responding to unexpected overloads in large-scale service systems. *Management Sci.* 55(8):1353–1367.

Reich M (2012) The offered-load process: Modeling, inference and applications. Unpublished Ph.D. thesis, Technion—Israel Institute of Technology, Haifa.

Ren ZJ, Zhou Y-P (2008) Call center outsourcing: Coordinating staffing level and service quality. *Management Sci.* 54(2): 369–383.

Scarsini M, Shaked M (1996) Positive dependence orders: A survey. Heyde CC, Prohorov YV, Pyke R, Rachev ST, eds. *Athens Conf. Appl. Probab. Time Ser. Anal.* (Springer, Berlin), 70–91.

Shaked M, Shanthikumar JG (2007) *Stochastic Orders* (Springer Science & Business Media, New York).

Sharakhmetov S, Ibragimov R (2002) A characterization of joint distribution of two-valued random variables and its applications. *J. Multivariate Anal.* 83(2):389–408.

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. *Oper. Res.* 54(1):37–54.

Whitt W (2006b) Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* 15(1):88–102.

Whitt W, You W (2018) Using robust queueing to expose the impact of dependence in single-server queues. *Oper. Res.* 66(1):184–199.

Zhang J (2013) Fluid models of many-server queues with abandonment. *Queueing Systems* 73(2):147–193.