OPERATIONS RESEARCH
Vol. 60, No. 4, July–August 2012, pp. 996–1009
ISSN 0030-364X (print) | ISSN 1526-5463 (online)
http://dx.doi.org/10.1287/opre.1120.1070

Overflow Networks: Approximations and Implications to Call Center Outsourcing

Itai Gurvich
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
i-gurvich@kellogg.northwestern.edu

Ohad Perry
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208,
ohad.perry@northwestern.edu

Motivated by call center cosourcing problems, we consider a service network operated under an overflow mechanism. Calls are first routed to an in-house (or dedicated) service station that has a finite waiting room. If the waiting room is full, the call is overflowed to an outside provider (an overflow station) that might also be serving overflows from other stations. We establish approximations for overflow networks with many servers under a resource-pooling assumption that stipulates, in our context, that the fraction of overflowed calls is nonnegligible. Our two main results are (i) an approximation for the overflow processes via limit theorems and (ii) asymptotic independence between each of the in-house stations and the overflow station. In particular, we show that, as the system becomes large, the dependency between each in-house station and the overflow station becomes negligible. Independence between stations in overflow networks is assumed in the literature on call centers, and we provide a rigorous support for those useful heuristics.

Subject classifications: overflow networks; cosourcing; heavy-traffic approximations; separation of time scales.
Area of review: Stochastic Models.
History: Received March 2011; revisions received July 2011, October 2011; accepted November 2011.

1. Introduction

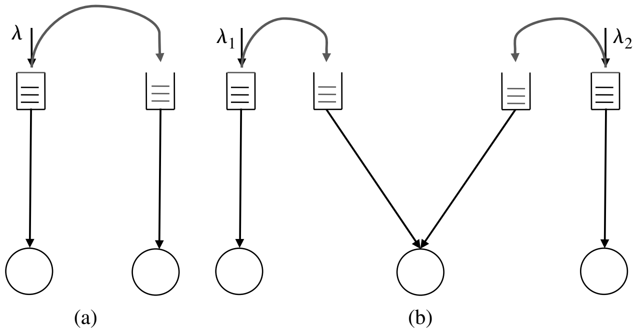
This work is motivated by call center applications and, in particular, call center outsourcing. Even though call centers often serve as a primary channel of interaction of firms with their customers, not all firms manage their call center operations in house. Some firms outsource their call center operations entirely, whereas others choose to serve a significant share of the customers in house, and route only some of the calls to an outside provider/outsourcer. The latter policy is sometimes referred to as cosourcing; see the detailed discussion in Zhou and Ren (2011). A network with cosourcing can be modeled by a queueing system with multiple queues overflowing some of the calls to an outsourcer. Figure 1 is a schematic depiction of two such networks. The outsourcer may provide a dedicated pool to each input stream, as in Figure 1(a)—which is prevalent in practice—or use a multiclass (and possibly multipool) configuration with skill-based routing (SBR) as in Figure 1(b).

Call overflow is a simple mechanism by which to divide the calls in real time between the in-house call center and the outsourcer. An arriving call is overflowed to the outsourcer when the queue length (found by this arrival) exceeds a prespecified threshold. Hence, the in-house call center operates as a queue with a finite waiting room. In this work we are primarily interested in the performance analysis of such overflow networks.

Our performance analysis should be placed in the context of, and is motivated by, optimization problems that emerge in the management of such distributed systems, with call center outsourcing being a primary example. In some settings (as studied, e.g., in Chevalier et al. 2004; see §2) there may be a central planner that makes the capacity planning and real-time control decisions for the entire network with the objective of minimizing total network costs subject to some quality-of-service (QoS) targets. Such a central planner/controller will be informed about the parameters across the network (exogenous parameters as well as decision variables) and may also have access to the real-time information about the state of each of the queues. Given the complexity of the network, the central planner faces a difficult optimization problem, and it is desirable to have simple prescriptions that utilize the information that is available to the planner.

When the network is managed in a decentralized manner (as is often the case in outsourcing), such information may not be readily available to "local" planners and controllers. In addition to practical prescriptions, one is interested in means to compare the performance of various coordination schemes for the decentralized network. Such comparisons are conducted in Gans and Zhou (2007); see §2. Given a QoS constraint that is placed on all customers—served in house or overflowed, one can then ask what is the

Figure 1. A network with in-house call centers and an outsourcer: (a) overflow is served by a dedicated pool; (b) overflows are served in a multiclass multipool system with SBR.



best outsourcing coordination mechanism that will guarantee that the constraint is met at a minimal capacity cost.

Coordination mechanisms may differ in the way in which information is shared and the way in which queues are pooled. Different coordination mechanisms will result in different queueing systems, as depicted in Figure 2. Figures 2(a) and 2(b) are the noncoordinated and coordinated versions of Figure 1(a), whereas Figures 2(c) and 2(d) correspond to the setting in Figure 1(b) in which the outsourcer uses a common system to serve multiple (in this case, two) input streams.

Figures 2(a) and 2(c) depict cases in which there is no pooling. The in-house call centers use some policy to overflow calls to the outsourcer who guarantees to meet a service-level target. No queues are pooled and no real-time information is shared between the parties. Partial coordination can be achieved by sharing real-time information. In the multiple-streams case, depicted in Figure 2(c), real-time information about the state of the queues in the in-house call centers may allow the outsourcer to intelligently choose his prioritization rule and, in turn, decrease his capacity costs. The level of coordination can be increased further by having joint virtual queues so that calls are pulled from a common queue (by either the in-house or the outsourcer agents). The resulting pooled systems are as depicted in Figures 2(b) and 2(d), and are referred to in the literature as the inverted-V (or $\bigwedge$) and M models, respectively.

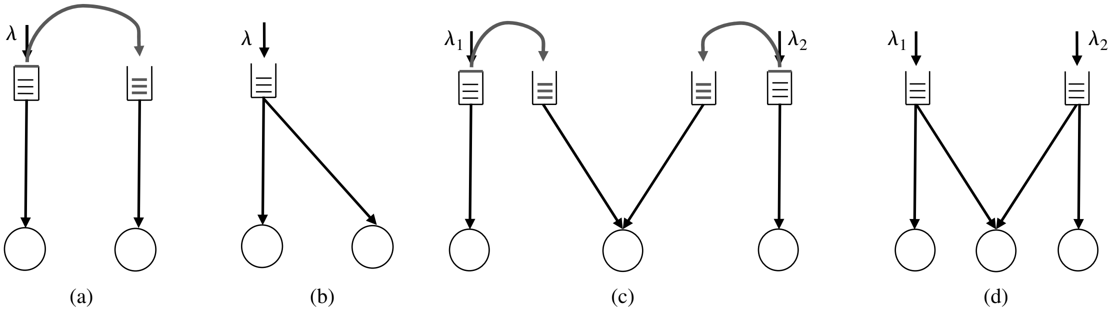
To compare the various schemes, one needs to evaluate the performance of the overflow network with respect to various QoS metrics. Whereas some metrics (such as the average speed of answer (ASA)) are separable via Little's law, most QoS metrics require knowledge of the joint distribution of the queues. Furthermore, for practical purposes, it is desirable to have accurate (but simple) approximations for the overflow processes and the queueing-system dynamics. Such approximations may facilitate solutions for the respective optimization problems of both the in-house call centers and the outsourcer. Our performance analysis, and the simplifications it introduces, has implications for decision making in both the centralized and decentralized settings. We conduct the performance analysis in a many-server heavy-traffic regime with *resource pooling*.

In the context of outsourcing, the resource-pooling condition can be interpreted as corresponding to nonnegligible cosourcing, namely, to settings in which the capacities of the in-house centers require that a nonnegligible fraction (but not all) of the calls be overflowed. The survey (ICMI Press 2006) indicates that the percentage of call centers that fall into this category is significant, and that a relatively small percentage of firms rely on an outsourcer to handle most or all of their call volume. There may be multiple reasons for this prevalence. Vendors, for example, may charge a minimal fixed cost (say, for hiring and training costs) that renders it profitable for the client to outsource more than a negligible fraction of his calls. Also, clients may face physical constraints on their in-house capacity that limit the volume of calls that can be handled in house in busy days. The explicit economic modeling of this choice is beyond the scope of this paper. Rather, we impose this "nonnegligible overflow" as an assumption; see §3.2 for the formal definition of this requirement.

Our main results are summarized below:

1. *The overflow process*: In the Markovian setting, the overflow process is a renewal process having an interarrival time distribution that can be identified explicitly by means of Laplace transforms; see §2. We improve the

Figure 2. Different coordination schemes for outsourcing: (a) overflow with dedicated outsourcer, (b) pooled network with dedicated outsourcer, (c) overflow with pooled outsourcer, (d) pooling with pooled outsourcer.



understanding of this process by using many-server approximations instead. Building on regenerative-process arguments, we prove limit theorems for the overflow process. We show that the overflow process can be approximated by a drifted Brownian motion whose mean and variance terms we identify as explicit functions of the capacity of the in-house call center. Interestingly, the instantaneous drift of this Brownian motion does not depend on the actual state of the in-house station or on the time point t. Rather, for each time t, the instantaneous drift depends only on the *long-run probability of blocking* in the in-house pool. That phenomenon is caused by an *averaging principle* (AP), which is a consequence of a separation of time scales, further discussed below.

Our results in this context have both theoretical and practical implications: first, the characterization of the limit provides insight into properties of the overflow process and, specifically, into the way in which its variability depends on the capacity of the in-house call center. Second, a simplified (closed-form) characterization of the overflow process is useful for purposes of capacity and prioritization optimization in overflow networks.

2. *Characterization of the joint distribution*: In terms of detailed analysis of the network, one would ideally be able to characterize for each t (or at least in steady state) the joint distribution of the number-in-system processes. For example, we are interested in the joint distribution of $(X_I(t), X_O(t))$, where $X_I(t)$ and $X_O(t)$ are the head-counts in the "in-house" station and the "outsourcer" station, respectively, at time t. Because a network with overflow does not have, in general, a product-form distribution, this joint distribution can be identified only via brute-force computation.

Simplifications in heavy traffic are often achieved via a reduction of dimensionality, typically referred to as state-space collapse (SSC). In our setting, the corresponding SSC result implies that, under appropriate diffusion scaling, $X_I(t)$ is approximated by a deterministic constant, whereas $X_O(t)$ is stochastic. It thus may seem that, through SSC, we achieve a great simplification for computing the joint distribution above.

However, this SSC result is somewhat crude and deceiving because, if both $X_I(t)$ and $X_O(t)$ are scaled in the same manner, meaningful information regarding $X_I(t)$ is lost. To gain a better understanding of the system, we must conduct a more refined analysis and consider the in-house station *without any scaling*, so that no part of the network degenerates in the limit. We then prove that, under the resource-pooling condition, the in-house station is asymptotically independent of the outsourcer station, i.e., that the dependency between the stations diminishes as the size of the system increases. In particular, we show that the joint distribution above "approaches" a product-form structure as the system size grows, for each time t, and not only in steady state. Such independence, as we formally prove here, is assumed heuristically in multiple papers that consider optimization problems for call centers with overflow; see §2.

The asymptotic independence is *not implied directly by the SSC mentioned above*. Rather, it requires a different analysis that builds on the fast oscillations of the in-house queue about the threshold determining the buffer size. These oscillations are not only relatively small—as reflected by the SSC result—but they are also sufficiently fast so that a separation of time scales occurs in the limit. In particular, the in-house queue operates in a faster time scale than that of the outsourcer, so that, relative to the outsourcer, the in-house queue approaches its steady state instantaneously at each time point t, a phenomenon typically referred to as "pointwise stationarity."

3. *Implications to outsourcing*: Our performance analysis has the following implications for the comparison of the different coordination schemes in Figure 2: (i) The complexity of the overflow network in Figure 2(c) is no greater than that of three independent queueing systems: two simple ones (with a single customer class and a single agent group), and one that corresponds to a multiclass single-pool system (often referred to as the V model). (ii) The overflow networks with and without real-time information sharing are, in a sense, equivalent. This strong statement follows from our result that the absence of real-time information on the state of the stations has at most negligible effect on the optimal prioritization chosen by the outsourcer and on his corresponding capacity costs. As a result, comparing the overflow network in Figure 2(c) (without real-time information sharing) to the pooled network in Figure 2(d) is equivalent to comparing a centrally controlled V model to the centrally controlled pooled system.

Finally, although results in the spirit of our separation of time scales (and the resulting pointwise stationarity and AP) are often complicated to prove, the specific structure of the network that we study allows us to provide relatively simple proofs, so that mathematical complexity does not obscure the underlying intuition.

2. Literature Review

Four streams of literature are directly related to our work: (i) queueing models of call centers, (ii) queueing systems with blocking, (iii) call center outsourcing, and (iv) pointwise stationarity and averaging principles in queueing systems.

The literature on queueing aspects of call centers is now vast, and we refer the reader to the three survey papers Akşin et al. (2007), Gans et al. (2003), and Koole and Pot (2006). The last reference focuses specifically on multiclass multiskill call centers. Below we discuss only the papers that are most relevant to our work.

For a survey on service outsourcing and, more specifically, on call center outsourcing, we refer the reader to Zhou and Ren (2011). Most of that literature focuses on settings in which the firm outsources all of its calls. Cosourcing is studied in Gans and Zhou (2007), taking into account the queueing effects. The focus of that paper is on the comparison of different outsourcing schemes with respect to the way in which capacity and control are coordinated in the network. A combination of dynamic programming and simulation is used to draw conclusions about the performance of the different schemes. In essence, the paper is concerned with the trade-off between the level of coordination (information sharing or pooling) and the cost of capacity in the absence of such coordination. The model studied in Gans and Zhou (2007) is different than the one we study here: in their model, the in-house call center serves two classes of customers of which only the lower-priority calls may be overflowed and the overflow is based on the number of idle servers in the in-house call center rather than on the buffer space. The analysis in Gans and Zhou (2007) underscores the difficulty in evaluating coordinating schemes given the relative intractability of the underlying overflow network. Our results facilitate such analysis for the family of models discussed in §1. It is likely, but yet to be proved, that results of similar spirit hold for the model Gans and Zhou (2007).

Approximations and bounds for overflow queues have been proposed both in the queueing literature and, more specifically, in the context of call centers, two notable examples being the papers Koole and Talim (2000) and Frankx et al. (2006). These papers also contain an account of earlier heuristic approximations and bounds. In Koole and Talim (2000) the overflow process is approximated by a Poisson process. In Frankx et al. (2006) the approximation is improved by using, instead, a renewal process with hyperexponential interarrival times. The authors study the loss probability in a call center with sequential overflows. In addition to the overflow approximation, the different stations are treated heuristically as being independent. In fact, in most papers that study call centers with overflow, independence between the stations is employed (explicitly or implicitly) in the construction of approximations for the overflow processes; see, e.g., Frankx et al. (2006), Avramidis et al. (2009), Chevalier and Van den Schrieck (2008), Chevalier and Tabordon (2003), Bhulai and Roubos (2010), and references therein.

Two other papers that are closely related to ours are Chevalier et al. (2004) and Chevalier and Tabordon (2003). These papers consider a pure-loss multistation system (there is no queueing in any station) having a set of dedicated stations and one pool of generalists (fully flexible servers). The focus in those papers is on using heuristics, based on the Hayward's approximation, to compute the probability of blocking and to provide staffing recommendations.

Loss queues, and more generally loss networks, have received significant attention in the queueing systems literature outside of a specific application context. Most papers focus on identifying blocking probabilities in such networks. For all but the simplest Markovian networks, the analysis of the blocking probabilities is complicated so that many papers resort to heavy-traffic approximations. Examples in the single-queue context are Massey and Whitt (1996), Whitt (1984), and, in the network context, Heyman (1987) and Hunt and Kurtz (1994); see also references therein.

One of our main results is concerned with approximations for the overflow process. Exact analysis of the interarrival time of the overflow renewal process via Laplace transforms are given, for example, in van Doorn (1984). In addition, various heuristic approximations have been proposed; see, e.g., Pourbabai (1987) and references therein. We take neither of these approaches. Instead, we achieve simplification by considering heavy-traffic limits. In the Halfin-Whitt regime, also referred to as the quality and efficiency driven (QED) regime, limits for the $M/M/N/K$ queue (having a Poisson arrival process, exponential service times, N agents, and a finite buffer of size K) have been studied in several papers; see Pang et al. (2007) and references therein. There are also various papers considering limits for the $M/M/N+M$ queue (with abandonment but without blocking; the $+M$ stands for exponential abandonments) in various heavy-traffic regimes; see e.g., Whitt (2004) and references therein.

The optimality of a threshold-based overflow in an outsourcing setting has been established, for example, in Koçağa and Ward (2010). There, the authors consider the in-house call center in isolation and prove that a threshold-based overflow policy is asymptotically optimal for a call center in the Halfin-Whitt (QED) many-server regime that seeks to minimize the combined costs of overflow, waiting time, and customer abandonment. It is important that in the QED regime the fraction of overflowed calls is negligible. We, in contrast, study the network comprising both the in-house call center and the outsourcer, and analyze the interaction between the two under the assumption that the fraction of calls overflowed is nonnegligible.

Finally, in terms of the relevant technical literature, results in which a process is approximated at each time point by a long-run average behavior of a related ("fast") process are often said to exhibit an AP. An AP appears in the limit whenever (at least) one of the processes evolves in a faster time scale than the other processes considered, so that the prelimit "fast" process is replaced by a simpler process whose parameters reflect long-run average quantities. In our paper, the AP is useful in simplifying the system performance analysis. There are several papers that deal with AP results in queueing systems, and we review the most relevant among these.

In Hunt and Kurtz (1994), functional law-of-large-numbers (FLLN) (or "fluid limits") are considered for large loss networks (with overflows between the various stations), and an AP-type result is established for the idle-capacity process. In the context of multiclass multipool systems with skill-based routing (SBR), our work is closely related to the sequence of papers from Perry and Whitt (2009; 2011a, b; 2012a, b). The latter reference considers a network of two customer classes with two server pools, and proposes a threshold-based routing policy to minimize convex holding costs. The proposed policy induces an AP. The sequence of papers (Perry and Whitt 2011a, b; 2012a, b) provides the technical support for Perry and Whitt (2009) by establishing corresponding functional limit theorems (FLLN as well as FCLT). Our model is different than that of Perry and Whitt (2009) in that we have overflow (customers are routed upon arrival) rather than routing (customers being "pulled" from queues), but there are some important similarities. In both models it is the fast oscillation around the thresholds that creates the AP.

The AP is related to pointwise stationary approximations (see, e.g., Bassamboo et al. 2009, Whitt 1991, Perry and Whitt 2012a, and references therein) because both phenomena are driven by a separation of time scales. Whereas the former, however, is concerned with process approximations, the latter is concerned with fixed times t. In the diffusion limit, the AP "replaces" a fast time-scale process whose instantaneous drift and variance are state dependent with a process whose instantaneous parameters are constants and in which the instantaneous drift is, at each time point, equal to the original process's long-run average. The pointwise stationarity result focuses on a given time point t and is concerned with the fast process achieving its steady state instantaneously, again at each time point.

For most of the paper we will focus on the simpler setting in Figure 1(a), but in §EC.2 we will show how our results extend to the setting with SBR in Figure 1(b). When we discuss the SBR setting we will highlight how, with our results, analysis of the outsourcer station can draw on established results provided the SBR protocol has certain properties. The queue-and-idleness ratio (QIR) controls, studied in Gurvich and Whitt (2009a, b; 2010) is one family of routing rules that has the desired properties, but many other controls are possible; see §EC.2.

Contribution to Existing Literature. Our contribution is fourfold: First, we provide a simple, yet rigorously justified, approximation for the overflow process in large systems, when the proportion of overflowed customers is non-negligible. Second, we establish an (atypical) asymptotic independence result showing that the complex overflow network exhibits an "asymptotic" product-form distribution. This result justifies some of the heuristics used in the existing literature, as reviewed above. Third, we provide tools that can be used to explicitly take into account the queueing effects when optimizing overflow networks or analyzing, for example, contracts and outsourcing schemes. Finally, in our setting, the separation of time scales phenomenon carries useful implications to the management of the underlying service-system. To the best of our knowledge, ours is the first instance where such separation of time scales leads to a pointwise stationary product-form distribution. Moreover, the mathematical analysis in this paper is simpler than in some of the papers reviewed above, especially in terms of the AP. This relative simplicity makes the instantaneous stationarity and fast averaging phenomenons more accessible and revealing.

Organization of the Remainder of the Paper. We introduce the model in §3, starting with the simple setting in Figure 1(a). This allows us to discuss outsourcing problems more formally. Those problems are used to motivate the main results, which are stated in §4. Some concluding remarks appear in §5. All the results are proved in the e-companion, where we also analyze the extension to the multiclass setting, as the one in Figure 1(b). An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1070.

3. The Model

We initially consider a system consisting of a single in-house station, which we refer to as station I, and an outsourcer station which we refer to as station O. Station I has N_I servers and a finite waiting room of size $K_I \geqslant 0$. In turn, there can be at most K_I customers in queue and at most a total of $N_I + K_I$ customers in the station at any given time. Exogenous arrivals follow a Poisson process $A = \{A(t),\ t \geqslant 0\}$ with rate λ. A customer that arrives to find less than $N_I + K_I$ customers in station I (waiting or being served) enters this station. The service discipline is first come first served (FCFS), and the service time is exponential with rate μ_I. Customers that find exactly $N_I + K_I$ customers in the station upon their arrival are overflowed to station O. We denote by $A_O(t)$ the number of calls that arrived by time t (inclusive) and were overflowed. The process $A_O = (A_O(t),\ t \geqslant 0)$ is the overflow process.

For most of the paper, the overflow station is itself a single-class single-pool system to which the sole input stream consists of the overflows from station I; see Figure 1(a). The overflow station has N_O servers and an infinite waiting space, the service discipline is FCFS, and service times are exponential with rate μ_O. This setup thus corresponds to an outsourcer serving each input stream through a dedicated facility; see Figure 1(a). In §EC.2 we will show how the analysis extends to the setting in Figure 1(b) where multiple overflow streams are served in one facility with SBR.

Finally, customers (callers) may abandon at any point during their wait in station I or station O. We assume that customers have exponential patience with rate $\theta > 0$. A customer abandons the queue if his patience expires while waiting to be served. With the above assumptions, station I is an $M/M/N_I/K_I + M$. Marginally, station O operates like a $GI/M/N_O + M$ queue where the arrival process is the overflow process A_O. Considered jointly, the arrival process to station O depends on the evolution of station I.

State Descriptors. We let $Q_I(t)$ and $Z_I(t)$ be, respectively, the number of customers in queue and in service in station I at time t. We denote by $X_I(t) := Q_I(t) + Z_I(t)$ (where := stands for equality in definition) the corresponding total number of customers in station I (in service or waiting) at time t. Similarly, $Q_O(t)$, $Z_O(t)$, and $X_O(t) := Z_O(t) + Q_O(t)$ are the corresponding processes for station O. We let $V_I(t)$ and $V_O(t)$ denote, respectively, the offered wait at time t in stations I and O. The offered wait is the time that an infinitely patient customer, arriving at time t, would have to wait before entering service; see Mandelbaum and Zeltyn (2009). The corresponding virtual waits for a customer arriving at time t, until he enters service or abandons, are then given by $W_I(t) := V_I(t) \wedge \tau$ and $W_O(t) := V_O \wedge \tau$, where τ is an exponential random variable with rate θ that is independent of the other random variables and stands for the customer's patience. The virtual waiting time for a customer arriving to the system at time t then depends on whether that arriving customer is overflowed or not, and is given by

$$\begin{aligned} W(t) = {} & W_I(t)\mathbb{1}\{X_I(t) < N_I + K_I\} \\ & + W_O(t)\mathbb{1}\{X_I(t) = N_I + K_I\}. \end{aligned} \tag{1}$$

3.1. A Motivating Example—Call Center Outsourcing

We start by considering the setting in Figure 2(a), i.e., we consider one in-house pool having a dedicated service pool operated by an outsourcer. We assume that the capacity of the in-house pool is fixed and equal to N_I and the threshold is specified to be K_I. The firm is interested in satisfying a constraint of the form $\mathbb{E}[f(W(t))] \leqslant \alpha$ that applies to all customers—served in house or overflowed.[1]

Using (1), we have that

$$\begin{aligned} \mathbb{E}[f(W(t))] = {} & \mathbb{E}[f(W_I(t))\mathbb{1}\{X_I(t) < N_I + K_I\}] \\ & + \mathbb{E}[f(W_O(t))\mathbb{1}\{X_I(t) = N_I + K_I\}]. \end{aligned} \tag{2}$$

Given the capacity and threshold of the in-house call center, one can compute the first element on the right-hand side of (2). Say it is equal to $\beta \leqslant \alpha$. To guarantee that the global QoS target is met, the outsourcer then has to solve

$$\begin{aligned} \min_{N_O} \quad & C_s^O(N_0) \\ \text{s.t.} \quad & \mathbb{E}[f(W_O(t))\mathbb{1}\{X_I(t) = N_I + K_I\}] \leqslant \alpha - \beta, \\ & N_O \in \mathbb{Z}_+. \end{aligned} \tag{3}$$

Here, $C_s^O(\cdot)$ is the capacity cost function for the outsourcer, and $\mathbb{Z}_+$ is the set of nonnegative integers. The QoS constraint in (3) places a bound on a performance metric of the customer waiting time. Note that the constraint depends on the joint distribution of stations O and I. If the queues were pooled, as in Figure 2(b), one would be solving (3) with the original constraint $\mathbb{E}[f(W(t))] \leqslant \alpha$, and this would be an optimization problem over a single-class multipool queueing system (known as the inverted-V model) as studied, for example, in Armony (2005). To compare the settings, we need to be able to solve (3).

In practice, the outsourcer would rarely solve a problem as in (3). In fact, it is more likely that the in-house call center, given its parameters $(\lambda, \mu_I, N_I, K_I)$, would calculate the expected steady-state blocking probability $p_b := \mathbb{P}\{X_I(\infty) = N_I + K_I\}$ and subsequently require from the outsourcer to satisfy the constraint $\mathbb{E}[f(W_O(t))] \leqslant (\alpha - \beta)/p_b$. In the special case in which the constraint is on the average wait ($f(x) = x$) and the system is stationary (i.e, has the same distribution for each t), this simplification is, in fact, correct. Indeed, by virtue of Little's law, we then have that $\mathbb{E}[W(t)] = (1 - p_b)\mathbb{E}[W_I(t)] + p_b\mathbb{E}[W_O(t)]$ for each t. However, such a simplification is unlikely to provide the desired result for more general QoS metrics or for nonstationary settings. Specifically, choosing N_O to be the optimal solution to

$$\begin{aligned} \min_{N_O} \quad & C_s^O(N_0) \\ \text{s.t.} \quad & \mathbb{E}[f(W_O(t))] \leqslant \frac{\alpha - \beta}{p_b}, \\ & N_O \in \mathbb{Z}_+, \end{aligned} \tag{4}$$

does not guarantee that the global constraint $\mathbb{E}[f(W(t))] \leqslant \alpha$ is met. Moreover, even if the solution to this simplified problem is feasible with respect to the global constraint, the replacement of (3) with (4) may lead to an increase in costs. In other words, the question raised is whether by obtaining information about the joint distribution (that allows to solve (4)), the outsourcer can reduce capacity costs compared to (4).

Information may carry more value in settings as in Figure 1(b), where the outsourcer serves multiple customer classes and can use this information to determine the optimal prioritization of customers. For concreteness, assume that, exactly as in Figure 1(b), the outsourcer is serving two input streams from in-house pools 1 and 2, having α_1 and α_2 as their QoS targets, and having capacity and threshold parameters $N_{i,I}$ and $K_{i,I}$, such that $\beta_i := \mathbb{E}[f(W_{i,I}(t))\mathbb{1}\{X_{i,I}(t) < N_{i,I} + K_{i,I}\}]$, $i = 1, 2$ (where we added the superscript i to denote the respective in-house call center). Let $W_{1,O}(t)$ and $W_{2,O}(t)$ be the virtual waiting times at the outsourcer for input streams 1 and 2. Let Π denote a family of admissible prioritization policies. A prioritization policy $\pi \in \Pi$ specifies which customer class a newly available agent should serve, given that there are customers waiting in both queues. We add a superscript π to the processes to denote their dependence on the prioritization policy. Then, in the nonpooled system (depicted in Figure 2(c)), the outsourcer's problem (3) becomes

$$\begin{aligned} \min_{N_0} \quad & C_s^O(N_0) \\ \text{s.t.} \quad & \mathbb{E}[f(W_{i,O}^{\pi}(t))\mathbb{1}\{X_{i,I}^{\pi}(t) = N_i + K_i\}] \leqslant \alpha_i - \beta_i, \\ & \qquad\qquad i = 1, 2, \\ & N \in \mathbb{Z}_+, \pi \in \Pi. \end{aligned} \tag{5}$$

One may be interested in two comparisons: First, one may examine how information sharing allows for better prioritization rules and, in turn, lower capacity costs; second, one can study the impact of pooling by comparing the nonpooled system in Figure 2(c) with the pooled system in Figure 2(d). The latter is a two-class, three-agent group system, with one pool serving both classes, often referred to as the M model of SBR.

Our performance analysis will facilitate comparisons as those discussed above. Specifically, returning to the notation of the simpler setting in Figure 1(a), we will show (see Theorems 4.3 and 4.4) that the head count processes, $X_I(t)$ and $X_O(t)$, exhibit asymptotic independence, which further implies asymptotic independence of the waiting times, namely,

$$\mathbb{E}[f(W(t))] \approx \mathbb{E}[f(W_I(t))]\mathbb{P}\{X_I(t) < N_I + K_I\} \\ + \mathbb{E}[f(W_O(t))]\mathbb{P}\{X_I(s) = N_I + K_I\}. \qquad (6)$$

The asymptotic independence allows the replacement of constraint in (3) by the simpler constraint

$$\mathbb{E}[f(W_O(t))]\mathbb{P}\{X_I(t) = N_I + K_I\} \leqslant \alpha - \beta.$$

Moreover, we will prove a pointwise stationarity result by which, for each $t > 0$ (and not only in stationarity), $\mathbb{P}\{X_I(t) = N + K\}$ can be approximated by the steady-state probability of blocking in the corresponding $M/M/N_I/K_I + M$ queue.

Notably, even with this independence, the problem (3) is nontrivial because the input stream to the outsourcer's queue (the overflow process) is a renewal process with a complicated interarrival time distribution. We will provide an approximation for the overflow process via limit theorems; see Theorem 4.1. Our approximation is characterized explicitly and in a simple way via the parameters λ, μ_I and N_I, K_I of the in-house call center. The overflow approximation will allow us to study the value of real-time state information in the context of the optimization problem (5). We will show that the benefit of such information towards the optimal cost in (5) is negligible; see §EC.2.

We prove our results under the condition that the amount of overflow is nonnegligible, namely, under the condition that $(\mu_I N_I + \theta K_I)/\lambda < 1$. This is a special case of what is referred to in the queueing heavy-traffic literature as a resource-pooling condition. The assumption of nonnegligible overflow is formalized in the next section.

3.2. Heavy-Traffic Scaling and Main Assumptions

We consider a sequence of systems, indexed by the arrival rate λ, and study the properties of the sequence as λ grows. To make the dependence on the index explicit we add the superscript λ to all quantities and processes. The service rates μ_I and μ_O and the abandonment rate θ are held fixed, and we omit the superscript from these. Then, N_I^λ, N_O^λ, and K_I^λ stand, respectively, for the staffing levels in stations I and O, and the maximal buffer space in station I in the λth system. These three quantities are assumed to be nonnegative and to satisfy the following assumption.

Assumption 1 (A Resource-Pooling Condition). *The sequence* $\{(N_I^\lambda, K_I^\lambda)\}$ *satisfies*

(1) $\lim_{\lambda\to\infty}((\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda) = \nu < 1$ *as* $\lambda \to \infty$, *and*

(2) $N_O^\lambda = R_O^\lambda + \varsigma\sqrt{R_O^\lambda} + o(\sqrt{R_O^\lambda}) - \infty < \varsigma < \infty$, *where*

$$R_O^\lambda = (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)/\mu_O.$$

where, for a family of numbers $\{a^\lambda; \lambda \geqslant 0\}$, $a^\lambda = o(f(\lambda))$ if $\limsup_{\lambda\to\infty} |a^\lambda/f(\lambda)| = 0$. The first item in Assumption 1 is the formalization of the resource-pooling condition. It requires that the fraction of incoming calls that have to be overflowed (out of the total arrival rate) is nonnegligible. Indeed, because $\theta K_I^\lambda + \mu_I N_I^\lambda$ is the maximum rate of departures from station I (via service completions or abandonment), the volume of overflowed calls will be at least $\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda$. Observe that we do not impose additional scaling restrictions on the threshold beyond the requirement that, together with the staffing, the resource-pooling condition is satisfied.

The quantity R_O^λ can be thought of as the offered load to station O. Item (2) in Assumption 1 then requires that station O is staffed according to the so-called "square root safety staffing rule." In fact, a weaker condition suffices for our analysis, namely, that

$$\liminf_{\lambda\to\infty} \frac{N_O^\lambda}{R_O^\lambda} \geqslant 1, \qquad (7)$$

or, in words, that station O has sufficient capacity to serve a majority (but not necessarily all) of the overflowed calls. The square-root safety staffing is one particular choice that satisfies (7). We impose the more restrictive square-root rule to make our statements cleaner. Remark 4.5 explains how our results are extended to the general case.

Scaled Processes. We introduce the following scaled processes:

$$\hat{X}_I^\lambda(t) := \frac{X_I^\lambda(t) - (N_I^\lambda + K_I^\lambda)}{\sqrt{\lambda}}, \quad \hat{Q}_O^\lambda(t) := \frac{Q_O^\lambda(t)}{\sqrt{\lambda}},$$

$$\hat{X}_O^\lambda(t) := \frac{X_O^\lambda(t) - R_O^\lambda}{\sqrt{\lambda}},$$

and

$$\hat{A}_O^\lambda(t) := \frac{A_O^\lambda(t) - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)t}{\sqrt{\lambda}}.$$

Per our previous discussion, it is natural to center $X_O^\lambda(t)$ around the offered load R_O^λ and center $A_O^\lambda(t)$ about its first-order estimate $(\lambda_I - \mu_I N_I^\lambda - \theta K_I^\lambda)t$. We will show that this centering indeed gives rise to meaningful limits. As mentioned in the introduction, for our asymptotic independence results we will consider the (unscaled) process X_I^λ rather that its scaled version defined above.

Some Notational Conventions. Following standard conventions we use $\mathbb{Z}_+$ to denote the nonnegative integers, and use $\mathbb{R}$ and $\mathbb{R}_+$ to denote, respectively, the real numbers and the nonnegative real numbers. For an integer $d \geqslant 1$, we let $\mathbb{R}^d$ denote all d-dimensional vectors with components in $\mathbb{R}$ and let $\|\cdot\|$ be the usual Euclidean norm on $\mathbb{R}^d$.

We use $\overset{\mathrm{d}}{=}$ to denote equality in distribution and $\Rightarrow$ to denote convergence in distribution (i.e., weak convergence of random variables or random processes). For a family of random variables $\{Y^\lambda\}$ in $\mathbb{R}^d$, we write $Y^\lambda \Rightarrow Y$ when the sequence of random variables Y^λ converges in distribution to a limit random variable Y.

We remove the time index from processes when referring to the whole process rather than its value at a specific time point. For example, we write X_I^λ for the process $(X_I^\lambda(t),\ t \geqslant 0)$. We let e denote the identity function, namely, $e(t) = t$ for all $t \geqslant 0$.

We let $\mathscr{D}^d := \mathscr{D}^d[0,\infty)$ be the space of functions that are right-continuous with left limits (RCLL) from $[0,\infty)$ to $\mathbb{R}^d$ (when $d = 1$ we remove the superscript), endowed with the usual Skorohod J_1 topology. All underlying processes are assumed to be constructed as RCLL functions. If $\{x^\lambda\}$ is a sequence of $\mathscr{D}^d$-valued processes, we will write $x^\lambda \Rightarrow x$ to denote convergence in distribution in $\mathscr{D}^d[0,\infty)$. We will write that the convergence is in $\mathscr{D}^d(0,\infty)$ when the convergence holds on compact subsets of $(0,\infty)$ (i.e., excluding the point 0). Because all our established limits are continuous, convergence in any of the common nonuniform metrics on $\mathscr{D}^d$ is equivalent to uniform convergence.

Finally, following standard notation, for a family of numbers $\{a^\lambda; \lambda \geqslant 0\}$ we write $a^\lambda = O(f(\lambda))$ if $\limsup_{\lambda\to\infty} |a^\lambda/f(\lambda)| < \infty$ and write $a^\lambda = o(f(\lambda))$ if $\limsup_{\lambda\to\infty} |a^\lambda/f(\lambda)| = 0$. In particular, $a^\lambda = o(1)$ if $a^\lambda \to 0$ as $\lambda \to \infty$. Analogously, for a sequence G^λ of random variables we write $G^\lambda = O_P(f(\lambda))$ if the sequence $\{\|G^\lambda\|/f(\lambda)\}$ is tight (see Billingsley 1968). We say that $G^\lambda = o_P(f(\lambda))$ whenever $\|G^\lambda\|/f(\lambda) \Rightarrow 0$. For a sequence of stochastic processes $\{Y^\lambda\}$, we say that $Y^\lambda = o_p(f(\lambda))$ if, for each T, the sequence of random variables $G^\lambda := \sup_{0\leqslant s\leqslant T} \|Y^\lambda(s)\|$ satisfies $G^\lambda = o_P(f(\lambda))$.

4. Main Results

In this section we state our main results. Theorem 4.1 is concerned with a Brownian approximation for the overflow process. Theorem 4.3 is concerned with the asymptotic independence of stations I and O and Corollaries 4.4 and 4.5 are concerned with the implications of asymptotic independence to the virtual waiting time and related averages. Throughout, Assumption 1 holds, and ν is as defined in item (1) of that assumption.

A key role in our results is played by the process $D_I^\lambda = \{D_I^\lambda(t), t \geqslant 0\}$ defined for each t by

$$D_I^\lambda(t) := N_I^\lambda + K_I^\lambda - X_I^\lambda(t). \tag{8}$$

This process captures the difference between the number of customers present in station I, X_I^λ, and the maximum space in this station, $N_I^\lambda + K_I^\lambda$. Hence, D_I^λ is a nonnegative process taking integer values in $[0, N_I^\lambda + K_I^\lambda]$. We refer to D_I^λ as the *availability process* because a customer enters station I if $D_I^\lambda(t) > 0$ and is overflowed otherwise. The amount of time on $[0,t]$ in which customers cannot enter station I is then given by the process C_I^λ defined for each t by

$$C_I^\lambda(t) := \int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\}\, ds. \tag{9}$$

4.1. Limit Approximations for the Overflow Process

Theorem 4.1 (FCLT for the Overflow Process). *Suppose that Assumption* 1 *holds and that*

$$\hat{X}_I^\lambda(0) \quad\Rightarrow\quad \hat{X}_I(0) \quad \textit{as } \lambda \to \infty. \tag{10}$$

Then

$$(\hat{A}_O^\lambda, \hat{X}_I^\lambda, C_I^\lambda) \Rightarrow (\sigma B_O, 0e, (1-\nu)e) \quad \textit{in } \mathscr{D} \textit{ as } \lambda \to \infty,$$

where $\sigma = \sqrt{1+\nu}$ *and* B_O *is a standard Brownian motion.*

Remark 4.1 (Implications). It follows from Theorem 4.1 that, under the resource-pooling condition, the overflow process satisfies

$$A_O^\lambda = (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)e + \sqrt{\lambda}\sigma B_O + o_P(\sqrt{\lambda}).$$

It is useful to note that the same approximation applies to a renewal process with mean interarrival time $1/(\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)$ and squared coefficient of variation (SCV) for the interarrival times given by

$$\frac{\lambda\sigma^2}{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda} \approx \frac{\sigma^2}{(1-\nu)} \geqslant 1,$$

where ν is as in Assumption 1. Hence, Theorem 4.1 can be interpreted as stating that the overflow process is asymptotically equivalent to that renewal process.

Observe that as ν approaches 0, the SCV approaches 1, which is the SCV for a Poisson process. This is to be expected because as ν approaches 0, almost all calls are overflowed, so that the overflow process becomes practically equal to the exogenous Poisson arrival process A^λ. If, on the other hand, ν approaches 1 (which corresponds to negligible overflow), the coefficient of variation grows proportionally to $1/(1-\nu)$. In short, the greater the overflow, the smaller the corresponding variability relative to the mean.

Recall that the process X_I^λ evolves as the number of customers in an $M/M/N_I^\lambda/K_I^\lambda + M$ queue. In particular, for each λ, $X_I^\lambda(t) \Rightarrow X_I^\lambda(\infty)$ as $t \to \infty$, where the limit $X_I^\lambda(\infty)$ has the steady-state distribution of a $M/M/N_I^\lambda/K_I^\lambda + M$ queue with parameters λ, μ_I, θ. The following result is obtained as a corollary of Theorem 4.1 after showing that the scaled sequence of random variables $\hat{X}_I^\lambda(\infty)$ indeed converges as $\lambda \to \infty$. For the following, we let p_b^λ be the steady-state probability of blocking in this queue. From PASTA it holds that

$$p_b^\lambda := \mathbb{P}\{X_I^\lambda(\infty) = N_I^\lambda + K_I^\lambda\}.$$

Corollary 4.2. *Suppose that Assumption* 1 *holds. If* $X_I^\lambda(0) \overset{\mathrm{d}}{=} X_I^\lambda(\infty)$, *then condition* (10) *is satisfied and the result of Theorem* 4.1 *holds. Moreover, the sequence* $\{p_b^\lambda\}$ *satisfies*

$$p_b^\lambda = \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\lambda} + o\left(\frac{1}{\sqrt{\lambda}}\right) = (1-\nu) + o(1). \quad (11)$$

One expects the long-run rate of overflows to be equal to $\lambda \mathbb{P}\{X_I^\lambda(\infty) = N_I^\lambda + K_I^\lambda\}$. Theorem 4.1 and Corollary 4.2 show that this rate actually holds for each $t > 0$. This "instantaneous steady-state" result is a consequence of an averaging principle (AP), as explained in the following remark.

Remark 4.2 (An AP). Focusing first on long-run averages, Corollary 4.2 shows that $1-\nu$ is approximately the steady-state (and, in turn, the long-run) fraction of time that station I is full (and the process D_I^λ spends at state 0). That is, for each λ, we have that

$$\begin{aligned}\mathbb{P}\{X_I^\lambda(\infty) = N_I^\lambda + K_I^\lambda\} &= \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}\{D_I^\lambda(s) = 0\}\, ds \\ &= 1 - \nu + o(1),\end{aligned}$$

with the $o(1)$ term converging to zero as λ grows large. The *uniform* convergence of C_I^λ to $(1-\nu)e$ implies something stronger. This convergence holds on any time interval $[t_0, t_1)$, $t_0 < t_1$ and *without any time and/or space scaling*. In other words, on any time interval, no matter how small, the average availability coincides with the long-run average one. This phenomenon, in which the cumulative process C_I^λ is replaced in the limit by its (deterministic) long-run average behavior, is an instance of the AP.

4.2. Asymptotic Independence

Theorem 4.1 shows that $\hat{X}_I^\lambda \Rightarrow 0$ in a suitable sense. This can be interpreted as a state-space collapse (SSC) result whereby the two-dimensional network is reduced to a one-dimensional limit. We will later show (see Lemma EC.1.2) that there is, in fact, a joint convergence of station I and O in the sense $(\hat{X}_I^\lambda, \hat{X}_O^\lambda) \Rightarrow (\hat{X}_I, \hat{X}_O)$ over compact intervals of $(0,\infty)$ where $\hat{X}_I \equiv 0e$. Hence, the sequence $\{(\hat{X}_I^\lambda, \hat{X}_O^\lambda)\}$ exhibits a trivial form of independence under diffusion scaling. This trivialization is a consequence of scaling the centered X_I^λ and X_O^λ by the common factor $\sqrt{\lambda}$ rather than scaling each by its "natural" scale that will produce nontrivial limits.

The natural scaling factor for X_O^λ is $\sqrt{\lambda}$. Indeed, marginally, X_O^λ evolves as an $GI/M/N_O^\lambda + M$ queue in the Halfin-Whitt regime, for which diffusion limits have been established; see, e.g., Whitt (2005) and Dai et al. (2010). However, this is not true for X_I^λ. In fact, as we show in our proofs (see Lemma EC.1.2), X_I^λ "lives" in a neighborhood of $o(\sqrt{\lambda})$ around $N_I^\lambda + K_I^\lambda$ so that, in particular, D_I^λ is a process of magnitude $o(\sqrt{\lambda})$. Thus, the limit of $\hat{X}_I^\lambda$ gives no valuable information for the prelimit. The following example illustrates further why using a common scaler can "wash away" the dependency structures.

Example 1. Consider two sequences of random variables, $\{X^n\}_{n\geqslant 1}$ and $\{Y^n\}_{n\geqslant 1}$, where $Y^n = 1$ with probability (w.p.) 1/2, and $Y^n = 0$ otherwise. Let $X^n = \sqrt{n}$ if $Y^n > 0$ and $X^n = 0$ otherwise. Consider first the (sequence of) scaled variables $\hat{X}^n := X^n/\sqrt{n}$ and $\hat{Y}^n := Y^n/\sqrt{n}$. Because $\hat{Y}^n$ converges to the deterministic limit 0, $(\hat{X}^n, \hat{Y}^n) \Rightarrow (\hat{X}, \hat{Y})$ (see, e.g., Theorem 11.4.5 in Whitt 2002), where $\hat{X} = 1$ w.p. 1/2 and $\hat{X} = 0$ otherwise, and $\hat{Y} = 0$ w.p.1. Clearly, the limit is such that $\hat{X}$ is independent of $\hat{Y}$. However, the dependency between $\hat{X}^n$ and $\hat{Y}^n$ does not diminish as n grows. In particular, $1/2 = \mathbb{P}\{\hat{X}^n > 0, \hat{Y}^n > 0\} \neq \mathbb{P}\{\hat{X}^n > 0\} \mathbb{P}\{\hat{Y}^n > 0\} = 1/4$, for all n, no matter how large. To capture the dependency in the limit, one has to consider instead the sequence $\{(\hat{X}^n, Y^n)\}$ (with Y^n not scaled). In that case, for each n, $(\hat{X}^n, Y^n)$ is equal to $(1,1)$ with probability 1/2 and equal to $(0,0)$ otherwise.

Based on these observations, we pursue a refined analysis of the system, in which each of the processes is scaled by its natural scaling, so that nontrivial limits emerge. This will allow us to prove a stronger asymptotic independence between D_I^λ and $\hat{X}_O^\lambda$ that will also imply the asymptotic independence of the waiting times in the different stations. Our notion of independence is implicitly defined within the following theorem.

Theorem 4.3 (Asymptotic Independence). *Suppose that Assumption* 1 *holds and that*

$$(\hat{X}_I^\lambda(0), \hat{X}_O^\lambda(0)) \;\Rightarrow\; (\hat{X}_I(0), \hat{X}_O(0)) \quad \text{as } \lambda \to \infty. \quad (12)$$

Then D_I^λ *and* $\hat{X}_O^\lambda$ *are asymptotically independent for all* $t > 0$. *That is, for* $q \in \mathbb{R}$ *and* $d \in \mathbb{Z}_+$,

$$\begin{aligned}&\mathbb{P}\{\hat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d\} \\ &\quad = \mathbb{P}\{\hat{X}_O^\lambda(t) > q\}\mathbb{P}\{D_I^\lambda(t) = d\} + o(1), \quad t > 0.\end{aligned}$$

Also, for all such q *and* d,

$$\begin{aligned}&\mathbb{P}\{\hat{Q}_O^\lambda(t) > q, D_I^\lambda(t) = d\} \\ &\quad = \mathbb{P}\{\hat{Q}_O^\lambda(t) > q\}\mathbb{P}\{D_I^\lambda(t) = d\} + o(1), \quad t > 0.\end{aligned}$$

Remark 4.3 (Intuition). The asymptotic independence of D_I^λ and $\hat{X}_O^\lambda$ is driven by a separation between the time scales of the (unscaled) process D_I^λ and the (scaled) process $\hat{X}_O^\lambda$. The process D_I^λ approaches steady state almost instantaneously, so that for fixed $t, \epsilon > 0$ and all λ large enough, $D_I^\lambda(t+\epsilon)$ is "almost" independent of the "initial state" at time t, $D_I^\lambda(t)$. To prove this instantaneous steady-state limiting result we will show that the excursions of D_I^λ above 0 (which correspond to excursions of X_I^λ below $N_I^\lambda + K_I^\lambda$), are similar to the positive excursions of a very fast $M/M/1$ queue with traffic intensity $\nu < 1$. As λ grows, this $M/M/1$ queue will have an increasing number of busy cycles over any interval $[t, t+\epsilon)$ so that, in the limit, it converges to its steady state instantaneously; see (EC.5) in Theorem EC.1.4.

The asymptotic independence will then follow from the fact that the steady state of an ergodic Markov chain is independent of its initial condition.

The time scale of $\hat{X}_O^\lambda$ is “slower.” Specifically, the process $\hat{X}_O^\lambda(t)$ corresponds to a diffusion-scaled $GI/M/N_O^\lambda+M$ queue in the Halfin-Whitt regime and hence converges to a continuous process; see Theorem 3.1 of Whitt (2005). In turn, over a small interval of size ϵ, $\hat{X}_O^\lambda$ “hardly” moves, so that $\hat{X}_O^\lambda(t+\epsilon)\approx\hat{X}_O^\lambda(t)$. It follows that because $D_I^\lambda(t+\epsilon)$ is “almost independent” of both $D_I^\lambda(t)$ and $\hat{X}_O^\lambda(t)$, it is also “almost independent” of $\hat{X}_O^\lambda(t+\epsilon)$. The proof of the asymptotic independence result is a formalization of this intuition.

To state the results for the waiting-time metrics, we introduce the following notation: we let w_k^λ be the waiting time of the kth customer to arrive (whether overflowed or not). We let $w_{k,O}^\lambda$ be the waiting time of the kth customer that is overflowed upon arrival and $w_{k,I}^\lambda$ be the waiting time of the kth customer to enter station I. Note that $w_k^\lambda=w_{l,O}^\lambda$ for some integer l if the kth customer to arrive was overflowed. Similarly, $w_k^\lambda=w_{l,I}^\lambda$ for some integer l if the kth customer to arrive was not overflowed. Finally, we let $A_I^\lambda(t)$ be the number of customers admitted to station I by time t, i.e., $A_I^\lambda(t):=A^\lambda(t)-A_O^\lambda(t)$.

We focus on the case in which K_I^λ is of the order of $\sqrt{\lambda}$. Because station O uses a square-root safety staffing, one expects that both $W_I(t)=O(1/\sqrt{\lambda})$ and $W_O(t)=O(1/\sqrt{\lambda})$. Hence, to get meaningful results, as is typical in critically loaded many-server queues, the waiting times are scaled up by a factor of $\sqrt{\lambda}$. Accordingly, we let $\hat{W}^\lambda(t):=\sqrt{\lambda}W^\lambda(t)$, and we similarly define $\hat{W}_I^\lambda(t)=\sqrt{\lambda}W_I^\lambda(t)$ and $\hat{W}_O^\lambda(t)=\sqrt{\lambda}W_O^\lambda(t)$.

Corollary 4.4. *Let* f: $\mathbb{R}_+\to\mathbb{R}_+$ *be a bounded continuous function and assume that* $K_I^\lambda/\sqrt{\lambda}\to\bar{K}_I\geqslant 0$ *as* $\lambda\to\infty$. *Then, under the conditions of Theorem* 4.3, *it holds for all* $t>0$ *that*

$$\mathbb{E}[f(\hat{W}^\lambda(t))]=\mathbb{E}[f(\hat{W}_I^\lambda(t)](1-p_b^\lambda)+\mathbb{E}[f(\hat{W}_O^\lambda(t)]p_b^\lambda+o(1),$$

and

$$\mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A^\lambda(t)}f(\sqrt{\lambda}w_k^\lambda)\right]$$
$$=(1-p_b^\lambda)\mathbb{E}\left[\frac{1}{A_I^\lambda(t)}\sum_{k=1}^{A_I^\lambda(t)}f(\sqrt{\lambda}w_{k,I}^\lambda)\right]$$
$$+p_b^\lambda\mathbb{E}\left[\frac{1}{A_O^\lambda(t)}\sum_{k=1}^{A_O^\lambda(t)}f(\sqrt{\lambda}w_{k,O}^\lambda)\right]+o(1).$$

Corollary 4.5 (Limits for Waiting-Time Metrics). *Suppose that the conditions of Corollary* 4.4 *hold. Then, uniformly on compact subsets of* $(0,\infty)$, $(\hat{W}^\lambda,\hat{W}_I^\lambda,\hat{W}_O^\lambda)\Rightarrow(\hat{W},\hat{W}_I,\hat{W}_O)$ *as* $\lambda\to\infty$, *where* $\hat{W}_O$ *is the diffusion limit of the virtual waiting-time process in the* $GI/M/N_O^\lambda+M$ *queue, and* $\hat{W}_I\equiv\bar{K}_I/\nu$. *Moreover, for all* $t>0$,

$$\lim_{\lambda\to\infty}\mathbb{E}[f(\hat{W}^\lambda(t))]=\nu\mathbb{E}[f(\hat{W}_I(t))]+(1-\nu)\mathbb{E}[f(\hat{W}_O(t))],$$

and

$$\lim_{\lambda\to\infty}\mathbb{E}\left[\frac{1}{A^\lambda(t)}\sum_{k=1}^{A^\lambda(t)}f(\sqrt{\lambda}w_k^\lambda)\right]$$
$$=\nu\frac{1}{t}\int_0^t\mathbb{E}\Big[f(\hat{W}_I(s))]\,ds+(1-\nu)\frac{1}{t}\int_0^t\mathbb{E}[f(\hat{W}_O(s))\,ds\Big].$$

The second limit in Corollary 4.5 can be viewed as an asymptotic finite-horizon ASTA (arrivals see time averages) result.

Remark 4.4 (Discontinuous Functions f). Corollary 4.5 requires that the function $f(\cdot)$ be continuous. In fact, it suffices to require that f is such that the limit processes $\hat{W}_O$ and $\hat{W}_I$ satisfy

$$\int_0^\infty\mathbb{1}\{\{\hat{W}_I(s)\in disc\{f\}\}\,ds$$
$$=\int_0^\infty\mathbb{1}\{\hat{W}_O(s)\in disc\{f\}\}\,ds=0\quad\text{w.p. 1},$$

where $disc\{f\}$ is the set of discontinuity points of the function f. One case of special interest is $f(x)=\mathbb{1}\{x>T\}$, which corresponds to the common performance metric $\mathbb{P}\{\hat{W}^\lambda(t)>T\}$. The result of Corollary (4.5) continues to hold for this indicator function f provided that $K_I^\lambda/\sqrt{\lambda}\to\bar{K}_I\neq\nu T$, as $\lambda\to\infty$.

We conclude this section with a remark about the relaxation of item (2) in Assumption 1.

Remark 4.5 (When Station O Does Not Use a Square-Root Rule). As pointed out in §3.2, the assumption that station O uses a square-root staffing rule is not necessary, and it suffices that (7) holds. In that case, Theorem 4.3 continues to hold with the following minor modifications: let γ be such that

$$N_O^\lambda=R_O^\lambda+\varsigma(R_O^\lambda)^\gamma+o((R_O^\lambda)^\gamma).$$

Note that ς may be negative but by (7) $\varsigma<0$ necessarily implies that $\gamma<1$. Define

$$b^\lambda:=\begin{cases}R_O^\lambda & \text{if } \varsigma>0 \text{ or } \gamma\leqslant 1/2,\\ N_O^\lambda+\dfrac{\mu_O|\varsigma|(R_O^\lambda)^\gamma}{\theta} & \text{otherwise,}\end{cases}\qquad\text{and}$$

$$c^\lambda:=\begin{cases}0 & \text{if } \varsigma>0 \text{ or } \gamma\leqslant 1/2,\\ \dfrac{\mu_O|\varsigma|(R_O^\lambda)^\gamma}{\theta} & \text{otherwise.}\end{cases}$$

Then, one defines

$$\hat{X}_O^\lambda(t) = \frac{X_O^\lambda(t) - b^\lambda}{\sqrt{\lambda}}, \quad \text{and} \quad \hat{Q}_O^\lambda(t) = \frac{Q_O^\lambda(t) - c^\lambda}{\sqrt{\lambda}}.$$

With these new definitions, the proofs of all the results remain unchanged. Clearly, the staffing rule for station O does not affect Theorem 4.1, because that theorem focuses solely on station I. As for the asymptotic independence results, the proofs reveal that all that we require regarding station O, is that, given (12), the sequence of processes $\{\hat{X}_O^\lambda\}$ is C-Tight (see §15 of Billingsley 1968). Such tightness is guaranteed, for example, if the sequence $\{\hat{X}_O^\lambda\}$ converges to a continuous limit, as is indeed the case under the modified definition of $\hat{X}_O^\lambda$ and for any of the parameter combinations of γ and ς considered above. This convergence follows from the fact that station O is, in isolation, a $GI/M/N_O^\lambda + M$, and the application of existing results from the literature. Specifically, for $\gamma \leqslant 1/2$ the convergence follows, e.g., from Theorem 7.6 in Pang et al. (2007). For $\varsigma < 0$ and $1/2 < \gamma < 1$ such convergence is proved as in Theorem 2.1 of Whitt (2004). Whereas the result there is for $\gamma = 1$, similar arguments apply to any $1/2 < \gamma \leqslant 1$. Finally, if $\varsigma > 0$ and $\gamma > 1/2$, the $GI/M/N_O^\lambda + M$ queue is equivalent (asymptotically) to a $GI/M/\infty$ queue, so that the fact that there is convergence to a continuous limit follows from Whitt (1982).

Example 2 (A Numerical Example). We consider the network depicted in Figure 1(a). We use simulation to illustrate our two key asymptotic results: (i) the approximation for the overflow process in Theorem 4.1 and (ii) the asymptotic independence in Theorem 4.3.

We simulate several instances of this network varying in size (arrivals and capacity). As a base example, we consider a moderately sized network, having a total capacity of 42 servers. The largest system we consider has a total of 321 servers. To simplify the presentation of the results and choice of parameters, we assume that there are no abandonments, i.e., that $\theta = 0$, and that the service rates μ_I and μ_O are the same and equal to 1.

When increasing the capacity and the arrival rate we keep a few constants fixed (in alignment with our mathematical results). The constant ν, which represents the "fluid" proxy for the fraction of λ that can be served in house, is held fixed and equal to the value in the base case of 30/39. Also, the staffing in station O satisfies a square-root rule as in Assumption 1 with $\varsigma = 1$. The "fluid" proxy for the overflow rate is $\lambda - N_I$, so that the approximate load to station O is $R_O = \lambda - N_I$. We also keep constant the ratio $K_I/\sqrt{\lambda}$ (where K_I is, as before, the size of the buffer in station I) as well as the ratio $q_2/\sqrt{N_O}$ where q_2 is the value for which we will measure $\mathbb{P}\{Q_O(t) < q_2\}$. Both K_I and q_2 are rounded to obtain integer values.

We sample the system at a time t_S after initialization (all networks are initialized with all servers busy, but with no customers in either queue). To be consistent across instances, we let $t_S \approx 3{,}000/\lambda$ so that t_S is roughly the time it takes until 3,000 customers have arrived to the system.[2]

We created the simulation in ARENA and ran 10,000 replications for each of the four parameter combinations. The simulation output is rounded up to the 4th decimal number. The results are reported in Table 1. There are six columns in the simulation output. The value p_1 corresponds to $\mathbb{P}\{Q_I(t_S) < q_1\}$ and the value p_2 to $\mathbb{P}\{Q_I(t_S) < q_2\}$. The value reported in the column *Joint* corresponds to the joint probability $\mathbb{P}\{Q_I(t_S) < q_1, Q_O(t_S) < q_2\}$. By Theorem 4.3 we expect that, at least for large systems,

$$p_1 \cdot p_2 = \mathbb{P}\{Q_I(t_S) < q_1\}\mathbb{P}\{Q_O(t_S) < q_2\} \approx \text{Joint}. \tag{13}$$

The column Sim. σ reports the standard deviation of the random variables $A_O(t_S) - (\lambda - N_I)t_S$. The last column reports $\sigma := \sqrt{1+\nu}\sqrt{\lambda t_S}$. By Theorem 4.1 we expect that, at least for large systems,

$$\text{Sim. } \sigma \approx \text{Th. } \sigma. \tag{14}$$

The simulation output is encouraging in terms of the applicability of the result to systems of moderate sizes. The asymptotic independence (13) and the standard deviations of the overflow processes (14) that the theory predicts hold convincingly even for systems of moderate size.

Example 3 (Back to the Staffing Problems (3) and (4)). Using the setting of Example 2, we next consider the staffing problem (3) with $t = \infty$ (i.e., in steady state) and the performance function $f(x) := \mathbb{1}\{x > 0\}$. In that case, $\mathbb{E}[f(W^\lambda(\infty))] = \mathbb{P}\{W^\lambda(\infty) > 0\}$ captures the expected fraction of callers experiencing delay before being served. As in §3.1, given a fixed staffing N_I in station I, we then ask what is the minimal staffing level N_O in station O that guarantees that $\mathbb{P}\{W^\lambda(\infty) > 0\} \leqslant \alpha$. If $K_I^\lambda = 0$, a call that finds all servers busy is overflow so that $\mathbb{E}[f(W^\lambda(\infty))\mathbb{1}\{X_I^\lambda(\infty) < N_I^\lambda + K_I^\lambda\}] = 0$. It is also useful

Table 1. Simulation results.

	Input							Output					
#	λ	N_I	N_O	K	q_1	q_2	t_S	p_1	p_2	Joint	$p_1 \cdot p_2$	Sim. σ	Th. σ
1	39	30	12	5	4	6	50	0.5603	0.6431	0.3878	0.3603	8.2255	8.3066
2	78	60	23	7	6	9	39	0.5841	0.7222	0.4390	0.4218	11.3639	11.7473
3	156	120	42	10	9	12	20	0.5884	0.6731	0.4088	0.3960	16.23576	16.6132
4	312	240	81	14	13	17	10	0.67	0.588	0.4415	0.4278	23.4415	23.4946

to note that $\mathbb{1}\{W_O^\lambda(t) > 0\} = \mathbb{1}\{\hat{X}_O^\lambda(t) \geqslant 0\}$ (customers have to wait only if all servers are busy).

With this choice of the function f, problems (3) and (4) become

$$\begin{aligned} \min_{N_O}\ & C_s^O(N_0^\lambda) \\ \text{s.t.}\ & \mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0, X_I^\lambda(\infty) = N_I^\lambda + K_I^\lambda\} \leqslant \alpha, \\ & N_O \in \mathbb{Z}_+, \end{aligned} \tag{15}$$

and

$$\begin{aligned} \min_{N_O}\ & C_s^O(N_0^\lambda) \\ \text{s.t.}\ & \mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0\} \leqslant \alpha / p_b^\lambda, \\ & N_O \in \mathbb{Z}_+. \end{aligned} \tag{16}$$

In §3.1 we argued that our mathematical results will allow us to relate (15) to the simpler problem (16).

In contrast to (15), the problem in (16) is a staffing problem for a $GI/M/N$ queue for which simple asymptotic solutions exist. Let ς^* be such that $\tilde{\varsigma} = 2\varsigma^*/(1 + (1+\nu)/(1-\nu))$ solves $[1 + ((\tilde{\varsigma}\Phi(\tilde{\varsigma}))/\phi(\tilde{\varsigma}))]^{-1} = \alpha/(1-\nu)$ and $\phi(\cdot)$, $\Phi(\cdot)$ are, respectively, the standard normal density and distribution functions. Then, given our Theorem 4.1, it follows from Theorem 4 in Halfin and Whitt (1981) that the sequence $\{\hat{N}_O^\lambda\}$ defined through

$$\hat{N}_O^\lambda = \left\lceil R_O^\lambda + \varsigma^* \sqrt{R_O^\lambda} \right\rceil,$$

is asymptotically feasible for (16), i.e.,

$$\limsup_{\lambda \to \infty} \mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0\} \leqslant \alpha.$$

It is also asymptotically optimal in that any other asymptotically feasible sequence can be at most $o(\sqrt{R_O^\lambda})$ smaller than $\hat{N}_O^\lambda$.

To establish the connection between (15) and (16) we observe that with the (sequence of) staffing levels $\{\hat{N}_O^\lambda\}$, it holds that $\hat{X}_O^\lambda(\infty) \Rightarrow \hat{X}_O(\infty)$ for a well-defined limit. This again follows from our Theorem 4.1 and from Theorem 4 in Halfin and Whitt (1981). Moreover, it follows from Corollary 4.2 that $X_I^\lambda(\infty)/\sqrt{\lambda} \Rightarrow \hat{X}_I(0)$. (In fact, Theorem EC.1.4 in the e-companion shows that $\widehat{D}_I(0) = 0$.) In turn, Condition (12) is satisfied when initializing the network with its steady-state distribution, so we can apply Theorem 4.3 to conclude that

$$\begin{aligned} &\mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0, X_I^\lambda(\infty) = N_I^\lambda + K_I^\lambda\} \\ &\qquad = \mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0\}\mathbb{P}\{X_I^\lambda(\infty) = N_I^\lambda + K_I^\lambda\} + o(1) \\ &\qquad = \mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0\} p_b^\lambda + o(1) \\ &\qquad = \mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0\}(1-\nu) + o(1) \leqslant \alpha + o(1), \end{aligned}$$

where the last two equalities follow from Corollary 4.2 and the last inequality follows from our construction of the sequence $\{\hat{N}_O^\lambda\}$. Thus, the sequence $\{\hat{N}_O^\lambda\}$ of staffing levels is not only asymptotically optimal for (16) (in that it is within $o(\sqrt{R_O^\lambda})$ from the optimal), but is also asymptotically feasible for (15) in the sense that $\limsup_{\lambda\to\infty} \mathbb{P}\{\hat{X}_O^\lambda(\infty) \geqslant 0, X_I^\lambda(\infty) = N_I + K_I\} \leqslant \alpha$.

In fact, repeating the same argument for values of $\varsigma < \varsigma^*$ shows that $\hat{N}_O^\lambda$ is asymptotically optimal for (15). We omit the detailed argument and illustrate the strength of the proposed solution via a numerical experiment. For the experiment we use the target $\alpha = 0.1$. Specifically, we consider the system in Figure 1(a) with $\lambda = 312$, $\mu_I = \mu_O = 1$, $N_I = 240$ and no abandonment. Note that here $\nu = 240/312 \approx 0.77$. Using the procedure outlined above, we obtain $\varsigma^* = 1.6$, which yields (recall that $R_O = 72$) $\hat{N}_O^\lambda = \lceil 72 + 1.6\sqrt{72} \rceil = 86$. We use $t = 1{,}000$ so as to be close to steady-state. We then simulate 10,000 replications of the real system with the above parameters, and find that the joint probability satisfies $\mathbb{P}\{\hat{X}_O^\lambda(t) \geqslant 0, X_I^\lambda(t) = N_I^\lambda + K_I^\lambda\} = 0.0992$. Thus, the asymptotic independence and the overflow approximation allowed us to obtain a nearly optimal solution for (15), using relatively simple means.

5. Concluding Remarks

Motivated by call center outsourcing applications, we study an overflow network in which firms operate their own in-house service stations, but route a nonnegligible fraction of the customers to an outside provider. Our FCLT for the properly scaled overflow processes and our asymptotic independence results produce a significant reduction in complexity, which is advantageous in large systems where exact analysis is intractable. In fact, many of the heuristic approximations that were previously considered in the literature on optimization of overflow networks assume such independence of the stations as their starting point. An important contribution of our analysis is in showing that, under a resource-pooling condition, such assumptions have in fact a sound mathematical basis.

Our proofs rely on identifying a separation of time scales phenomenon in which the actual state of the in-house queue is "replaced" with its long-run average behavior, resulting in pointwise stationarity (alternatively, pointwise AP). Due to the fast oscillations of the in-house queues, the drift of the limiting overflow process is determined (at each time point) by the *deterministic* long-run fraction of time that the in-house buffer is full (an AP result) and which equals asymptotically to $1-\nu$. Hence, one can loosely argue that "the outside provider sees a steady-state long-run average behavior of the in-house systems at each time point" so that dependencies on the actual states of the in-house pools are negligible. However, as Theorem 4.1 and Remark 4.1 show, the coefficient of variation of the overflow process is greater than one would get from a Bernoulli thinning (with probability $1-\nu$) of the exogenous arrival process $A(t)$.

In the outsourcing context, the asymptotic independence simplifies the staffing decision of the outside provider. Furthermore, in a multiclass setting with SBR, the asymptotic independence implies that real-time information about the state of the in-house stations carries little benefit for the outside provider in solving his optimal control (or prioritization) problem. In fact, for both the staffing and control decisions it is sufficient for the outside provider to know, for each of the in-house call centers, its exogenous arrival rate λ and the "proxy" $\lambda - \mu_I N_I - \theta K_I$ for the overflow rate.

The outsourcing example that we used for motivation in §3.1 is simple in terms of the relationship between the in-house call center and the outsourcer. The fact that, at least under the resource-pooling condition, the queueing dynamics become tractable in the heavy-traffic limit, suggests that it may be possible to rigorously study various outsourcing and contracting schemes while taking the queueing effects explicitly into account.

Finally, whereas the overflow mechanism we considered is widely used in practice, alternative rules can also be considered. One may consider, for example, a time-based overflow rule in which customers are overflowed once their waiting times exceed some prespecified level. With such an overflow rule, the queue-length process in the in-house pool is no longer Markovian and this introduces new challenges. It is likely, however, that our key finding regarding the diminishing dependencies will continue to hold for such alternative overflow rules.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1070.

Endnotes

1. One may replace the requirement of time stable performance (i.e., for all $t \geqslant 0$) with averages over finite horizons (see e.g., Corollary 4.4). Under reasonable conditions one expects both constraints to be equivalent by PASTA.
2. The scaling of the sampling time has a strong justification within the analysis: recall that the process D_I^λ evolves as a "fast" underloaded $M/M/1$ queue that reaches steady state within a time proportional to $1/\lambda$. Thus, to capture all systems in the sequence at a similar stage of their dynamics, one has to scale the sampling-time point by λ.

Acknowledgments

This research began while the second author was a postdoctoral fellow at CWI in Amsterdam, the Netherlands. Both authors are grateful to the review team for their careful review of the paper and their numerous helpful comments.

References

Akşin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

Armony M (2005) Dynamic routing in large-scale service systems with heterogenous servers. *Queueing Systems* 51(3–4):287–329.

Avramidis AN, Chan W, L'Ecuyer P (2009) Staffing multi-skill call centers via search methods and a performance approximation. *IIE Trans.* 41(6):483–497.

Bassamboo A, Harrison JM, Zeevi A (2009) Pointwise stationary fluid models for stochastic processing networks. *Manufacturing Service Oper. Management* 11(1):70–89.

Bhulai S, Roubos D (2010) Approximate dynamic programming techniques for skill-based routing in call centers. Working paper, Vrije Universiteit, Amsterdam.

Billingsley P (1968) *Convergence of Probability Measures* (John Wiley & Sons, New York).

Chevalier P, Tabordon N (2003) Overflow analysis and cross-trained servers. *Internat. J. Production Econom.* 85(1):47–60.

Chevalier P, Van den Schrieck JC (2008) Optimizing the staffing and routing of small-size hierarchical call centers. *Production Oper. Management* 17(3):306–319.

Chevalier P, Shumsky RA, Tabordon N (2004) Routing and staffing in large call centers with specialized and fully flexible servers. Working paper, Tuck School of Business, Dartmouth College, Hanover, NH.

Dai JG, He S, Tezcan T (2010) Many-server diffusion limits for $g/ph/n+gi$ queues. *Ann. Appl. Probab.* 20(5):1854–1890.

Frankx GJ, Koole G, Pot A (2006) Approximating multi-skill blocking systems by hyperexponential decomposition. *Performance Evaluation* 63(8):799–824.

Gans N, Zhou YP (2007) Call-routing schemes for call-center outsourcing. *Manufacturing Service Oper. Management* 9(1):33–50.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

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

Gurvich I, Whitt W (2009b) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* 11(2):237–253.

Gurvich I, Whitt W (2010) Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. *Oper. Res.* 58(2):316–328.

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

Heyman DP (1987) Asymptotic marginal independence in large networks of loss systems. *Ann. Oper. Res.* 8(1):57–73.

Hunt PJ, Kurtz TG (1994) Large loss networks. *Stochastic Processes and Their Appl.* 53(2):363–378.

ICMI Press (2006) 2006 contact center outsouring report. http://www.icmi.com/files/ICMI/members/ccmt2006/ccmr06/June2006_issue.pdf.

Koçağa YL, Ward AR (2010) Admission control for a multi-server queue with abandonment. *Queueing Systems* 65(3):275–323.

Koole G, Pot A (2006) An overview of routing and staffing algorithms in multi-skill customer contact centers. Working paper, Vrije Universiteit, Amsterdam.

Koole G, Talim J (2000) Exponential approximation of multi-skill call centers architecture. *Proc. QNETs* (Ilkley, UK), 23/1–10.

Mandelbaum A, Zeltyn S (2009) Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* 57(5):1189–1205.

Massey WA, Whitt W (1996) Stationary-process approximations for the nonstationary Erlang loss model. *Oper. Res.* 44(6):976–983.

Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.

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

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

Perry O, Whitt W (2011b) An ode for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems* 1(1):59–108.

Perry O, Whitt W (2012a) A fluid limit for an overloaded X model via an averaging principle. Working paper, Columbia University, New York.

Perry O, Whitt W (2012b) Diffusion approximations for an overloaded X model via an averaging principle. Working paper, Columbia University, New York.

Pourbabai B (1987) Approximation of the overflow process from a $G/M/N/K$ queueing system. *Management Sci.* 33(7):931–938.

van Doorn EA (1984) On the overflow process from a finite Markovian queue. *Performance Evaluation* 4(4):233–240.

Whitt W (1982) On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Probab.* 14(1):171–190.

Whitt W (1984) Heavy traffic approximations for service systems with blocking. *AT&T Bell Lab. Tech. J.* 63(5):689–708.

Whitt W (1991) The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Sci.* 37(3):307–314.

Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer Series in Operations Research (Springer, New York).

Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* 50(10): 1449–1461.

Whitt W (2005) Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* 30(1):1–27.

Zhou Y-P, Ren ZJ (2011) Service outsourcing. Cochran JJ, ed. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ),

Itai Gurvich is an assistant professor of operations management in the Kellogg School of Management, Northwestern University. His research focuses on queueing theory and its application to capacity planning and real-time control of service systems.

Ohad Perry is an assistant professor in the Industrial Engineering and Management Sciences Department at Northwestern University. His research focuses on queueing models and their applications.