

# Overflow Networks : Approximations and Implications to Call-Center Outsourcing

Itai Gurvich

Kellogg School of Management, 2001 Sheridan Rd., Evanston IL 60208, i-gurvich@kellogg.northwestern.edu

Ohad Perry

Department of Industrial Engineering and Management Sciences, 2145 Sheridan Rd., Evanston IL 60208, ohad.perry@northwestern.edu

Motivated by call center co-sourcing 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 which stipulates, in our context, that the fraction of overflowed calls is non-negligible. 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.

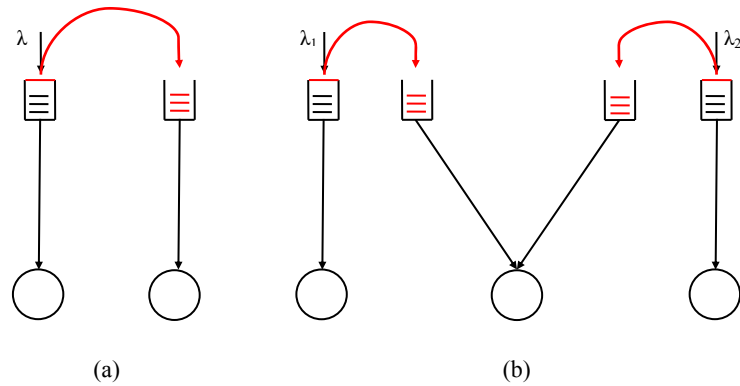
*Key words:* overflow networks, co-sourcing, heavy-traffic approximations, separation of time scales

---

## 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, while 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 co-sourcing; see the detailed discussion in Zhou and Ren [2010]. A network with co-sourcing 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 multi-class (and possibly multi-pool) configuration with Skill-Based

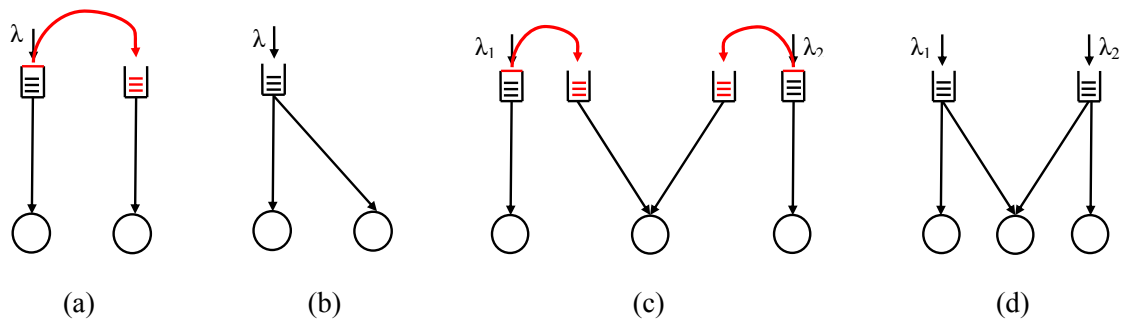
Routing (SBR) as in Figure 1(b).



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

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 pre-specified 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.



**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

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 Quality of Service (QoS) constraint that is placed on all customers – served in-house or overflowed, one can then ask what is the 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 non-coordinated 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 in-house or an 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  $\wedge$ ) 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. While 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 non-negligible co-sourcing, namely, to settings in which the capacities of the in-house centers require that a non-negligible fraction (but not all) of the calls be overflowed. The survey ICM [2006] indicates that the percentage of call centers that fall in 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 “non-negligible 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 inter-arrival time distribution that can be identified explicitly by means of Laplace transforms; see §2. We improve the 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  $t$ . Since 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 since, 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 multi-class 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, while 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 latter focuses specifically on multi-class multi-skill 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 [2010]. Most of that literature focuses on settings in which the firm outsources all of its calls. Co-sourcing 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 tradeoff 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 holds 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 hyper-exponential inter-arrival 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 multi-station 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 inter-arrival 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 of both the in-house call centers and the outsourcer, and analyze the interaction between the two under the assumption that the fraction of calls overflowed is non-negligible.

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 *Averaging Principle* (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 multi-class multi-pool systems with Skill-Based Routing (SBR), our work is closely related to the sequence of papers Perry and Whitt [2010a,b,d,c] and Perry and Whitt [2009]. 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 [2010a,b,d,c] 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 [2010b] and references therein) as 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 [2009b,a, 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 four-fold: 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 (a-typical) 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 phenomena 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).

### 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 to which we refer as station  $O$ . Station  $I$  has  $N_I$  servers and a finite waiting room of size  $K_I \geq 0$ . 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  $A = \{A(t), t \geq 0\}$  with rate  $\lambda$ . 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  $\mu_I$ . 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  $A_O = (A_O(t), t \geq 0)$  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  $\mu_O$ . 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  $\tau$  is an exponential random variable with rate  $\theta$  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

$$W(t) = W_I(t) \mathbb{1}\{X_I(t) < N_I + K_I\} + W_O(t) \mathbb{1}\{X_I(t) = N_I + K_I\}. \quad (1)$$

### 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  $\mathbb{E}[f(W(t))] \leq \alpha$  that applies to all customers – served in-house or overflowed.<sup>1</sup>

Using (1) we have that

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

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  $\beta \leq \alpha$ . To guarantee that the global QoS target is met, the outsourcer then has to solve

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

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  $\mathbb{E}[f(W(t))] \leq \alpha$  and this would be an optimization

<sup>1</sup>One may replace the requirement of time stable performance (i.e., for all  $t \geq 0$ ) with averages over finite horizons (see e.g. Corollary 4.4). Under reasonable conditions one expects both constraints to be equivalent by PASTA.

problem over a single-class multi-pool 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  $\mathbb{E}[f(W_O(t))] \leq (\alpha - \beta)/p_b$ . 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 non-stationary settings. Specifically, choosing  $N_O$  to be the optimal solution to

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

does not guarantee that the global constraint  $\mathbb{E}[f(W(t))] \leq \alpha$  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 obtaining information about the joint distribution (that allows to solve (4)), the outsourcer can reduce capacity costs compared to (3).

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  $\alpha_1$  and  $\alpha_2$  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  $\Pi$  denote a family of admissible prioritization policies. A prioritization policy  $\pi \in \Pi$  specifies which customer class should a newly available agent serve, given that there are customers waiting in both queues. We add a superscript  $\pi$  to the processes, to denote their dependence on the prioritization policy. Then, in the non-pooled system (depicted in Figure 2(c)), the outsourcer's problem (3) becomes

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

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 non-pooled 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,

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

The asymptotic independence allows to replace the constraint in (3) by the simpler constraint

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

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 non-trivial since the input stream to the outsourcer's queue (the overflow process) is a renewal process with a complicated inter-arrival-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  $\lambda, \mu_I$  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 non negligible. 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 non-negligible 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  $\lambda$ , and study the properties of the sequence as  $\lambda$  grows. To make the dependence on the index explicit we add the superscript  $\lambda$  to all quantities and processes. The service rates  $\mu_I$  and  $\mu_O$  and the abandonment rate  $\theta$  are held fixed and we omit the superscript from these. Then,  $N_I^\lambda$ ,  $N_O^\lambda$  and  $K_I^\lambda$  stand, respectively, for the staffing levels in stations  $I$  and  $O$ , and the maximal buffer space in station  $I$  in the  $\lambda^{th}$  system. These three quantities are assumed to be non-negative and to satisfy the following assumption.

ASSUMPTION 1. (**a resource pooling condition**) The sequence  $\{(N_I^\lambda, K_I^\lambda)\}$  satisfies

- (1)  $\lim_{\lambda \rightarrow \infty} \frac{\mu_I N_I^\lambda + \theta K_I^\lambda}{\lambda} = \nu < 1$  as  $\lambda \rightarrow \infty$ , and
- (2)  $N_O^\lambda = R_O^\lambda + \varsigma \sqrt{R_O^\lambda} + o(\sqrt{R_O^\lambda})$ , where  $R_O^\lambda = \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\mu_O}$ ,  $-\infty < \varsigma < \infty$ ,

where, for a family of numbers  $\{a^\lambda; \lambda \geq 0\}$ ,  $a^\lambda = o(f(\lambda))$  if  $\limsup_{\lambda \rightarrow \infty} |a^\lambda/f(\lambda)| = 0$ . 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 non-negligible. Indeed, since  $\theta K_I^\lambda + \mu_I N_I^\lambda$  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  $R_O^\lambda$  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

$$\liminf_{\lambda \rightarrow \infty} \frac{N_O^\lambda}{R_O^\lambda} \geq 1, \quad (7)$$

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:

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

and

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

Per our previous discussion, it is natural to center  $X_O^\lambda(t)$  around the offered load  $R_O^\lambda$  and center  $A_O^\lambda(t)$  about its first order estimate  $(\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)t$ . 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  $X_I^\lambda$  rather than 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 non-negative real numbers. For an integer  $d \geq 1$ , 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  $\stackrel{d}{=}$  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  $Y^\lambda$  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  $X_I^\lambda$  for the process  $(X_I^\lambda(t), t \geq 0)$ . We let  $e$  denote the identity function, namely,  $e(t) = t$  for all  $t \geq 0$ .

We let  $\mathcal{D}^d := \mathcal{D}^d[0, \infty)$  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  $\mathcal{D}^d$ -valued processes, we will write  $x^\lambda \Rightarrow x$  to denote convergence in distribution in  $\mathcal{D}^d[0, \infty)$ . We will write that the convergence is in  $\mathcal{D}^d(0, \infty)$  when the convergence holds on compact subsets of  $(0, \infty)$  (i.e., excluding the point 0). Since all our established limits are continuous, convergence in any of the common non-uniform metrics on  $\mathcal{D}^d$  is equivalent to uniform convergence.

Finally, following standard notation, for a family of numbers  $\{a^\lambda; \lambda \geq 0\}$ , we write  $a^\lambda = O(f(\lambda))$  if  $\limsup_{\lambda \rightarrow \infty} |a^\lambda/f(\lambda)| < \infty$  and write  $a^\lambda = o(f(\lambda))$  if  $\limsup_{\lambda \rightarrow \infty} |a^\lambda/f(\lambda)| = 0$ . In particular,  $a^\lambda = o(1)$  if  $a^\lambda \rightarrow 0$  as  $\lambda \rightarrow \infty$ . Analogously, for a sequence  $G^\lambda$  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  $G^\lambda := \sup_{0 \leq s \leq T} \|Y^\lambda(s)\|$  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  $\nu$  is as defined in item (1) of that assumption.

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

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

This process captures the difference between the number of customers present in station  $I$ ,  $X_I^\lambda$ , and the maximum space in this station,  $N_I^\lambda + K_I^\lambda$ . Hence,  $D_I^\lambda$  is a non-negative process taking integer values in  $[0, N_I^\lambda + K_I^\lambda]$ . We refer to  $D_I^\lambda$  as the *availability process* since 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  $C_I^\lambda$  defined for each  $t$  by

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

#### 4.1. Limit Approximations for the Overflow Process

THEOREM 4.1. (**FCLT for the overflow process**) Suppose that Assumption 1 holds and that

$$\widehat{X}_I^\lambda(0) \Rightarrow \widehat{X}_I(0) \quad \text{as } \lambda \rightarrow \infty. \quad (10)$$

Then

$$\left( \widehat{A}_O^\lambda, \widehat{X}_I^\lambda, C_I^\lambda \right) \Rightarrow (\sigma B_O, 0e, (1 - \nu)e) \quad \text{in } \mathcal{D} \text{ as } \lambda \rightarrow \infty,$$

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 inter-arrival time  $1/(\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)$  and squared coefficient of variation (SCV) for the inter-arrival times given by

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

where  $\nu$  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  $\nu$  approaches 0, the SCV approaches 1, which is the SCV for a Poisson process. This is to be expected since, as  $\nu$  approaches 0, almost all calls are overflowed, so that the overflow process becomes practically equal to the exogenous Poisson arrival process  $A^\lambda$ . If, on the other hand,  $\nu$  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  $X_I^\lambda$  evolves as the number of customers in an  $M/M/N_I^\lambda/K_I^\lambda + M$  queue. In particular, for each  $\lambda$ ,  $X_I^\lambda(t) \Rightarrow X_I^\lambda(\infty)$  as  $t \rightarrow \infty$ , 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  $\lambda, \mu_I, \theta$ . The following result is obtained as a corollary of Theorem 4.1 after showing that the scaled sequence of random variables  $\widehat{X}_I^\lambda(\infty)$  indeed converges as

$\lambda \rightarrow \infty$ . For the following, we let  $p_b^\lambda$  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  $X_I^\lambda(0) \stackrel{d}{=} X_I^\lambda(\infty)$ , 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  $D_I^\lambda$  spends at state 0). That is, for each  $\lambda$ , we have that

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

with the  $o(1)$  term converging to zero as  $\lambda$  grows large. The *uniform* convergence of  $C_I^\lambda$  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  $C_I^\lambda$  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  $\widehat{X}_I^\lambda \Rightarrow 0$  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  $(\widehat{X}_I^\lambda, \widehat{X}_O^\lambda) \Rightarrow (\widehat{X}_I, \widehat{X}_O)$  over compact intervals of  $(0, \infty)$  where  $\widehat{X}_I \equiv 0e$ . Hence, the sequence  $\{(\widehat{X}_I^\lambda, \widehat{X}_O^\lambda)\}$  exhibits a

trivial form of independence under diffusion scaling. This trivialization is a consequence of scaling the centered  $X_I^\lambda$  and  $X_O^\lambda$  by the common factor  $\sqrt{\lambda}$ , rather than scaling each by its “natural” scale that will produce non-trivial limits.

The natural scaling factor for  $X_O^\lambda$  is  $\sqrt{\lambda}$ . Indeed, marginally,  $X_O^\lambda$  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], Dai et al. [2010]. However, this is not true for  $X_I^\lambda$ . In fact, as we show in our proofs (see Lemma EC.1.2),  $X_I^\lambda$  “lives” in a neighborhood of  $o(\sqrt{\lambda})$  around  $N_I^\lambda + K_I^\lambda$  so that, in particular,  $D_I^\lambda$  is a process of magnitude  $o(\sqrt{\lambda})$ . Thus, the limit of  $\widehat{X}_I^\lambda$  gives no valuable information for the pre limit. The following example illustrates further why using a common scaler can “wash away” the dependency structures.

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

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

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

$$\mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} = \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q \right\} \mathbb{P} \left\{ D_I^\lambda(t) = d \right\} + o(1), \quad t > 0.$$

Also, for all such  $q$  and  $d$ ,

$$\mathbb{P} \left\{ \widehat{Q}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} = \mathbb{P} \left\{ \widehat{Q}_O^\lambda(t) > q \right\} \mathbb{P} \left\{ D_I^\lambda(t) = d \right\} + o(1), \quad t > 0.$$

**REMARK 4.3. (intuition)** The asymptotic independence of  $D_I^\lambda$  and  $\widehat{X}_O^\lambda$  is driven by a separation between the time scales of the (unscaled) process  $D_I^\lambda$  and the (scaled) process  $\widehat{X}_O^\lambda$ . The process  $D_I^\lambda$  approaches steady-state almost instantaneously so that, for fixed  $t, \epsilon > 0$  and all  $\lambda$  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  $D_I^\lambda$  above 0 (which correspond to excursions of  $X_I^\lambda$  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  $\lambda$  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  $\widehat{X}_O^\lambda$  is “slower”. Specifically, the process  $\widehat{X}_O^\lambda(t)$  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  $\epsilon$ ,  $\widehat{X}_O^\lambda$  “hardly” moves, so that  $\widehat{X}_O^\lambda(t + \epsilon) \approx \widehat{X}_O^\lambda(t)$ . It follows that, since  $D_I^\lambda(t + \epsilon)$  is “almost independent” of both  $D_I^\lambda(t)$  and  $\widehat{X}_O^\lambda(t)$ , it is also “almost independent” of  $\widehat{X}_O^\lambda(t + \epsilon)$ . 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  $w_k^\lambda$  be the waiting time of the  $k^{th}$  customer to arrive (whether overflowed or not). We let  $w_{k,O}^\lambda$  be the waiting time of the  $k^{th}$  customer that is overflowed upon arrival and  $w_{k,I}^\lambda$  be the waiting time of the  $k^{th}$  customer to enter station  $I$ . Note that  $w_k^\lambda = w_{l,O}^\lambda$  for some integer  $l$  if the  $k^{th}$  customer to arrive was overflowed. Similarly,

$w_k^\lambda = w_{l,I}^\lambda$  for some integer  $l$  if the  $k^{th}$  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  $K_I^\lambda$  is of the order of  $\sqrt{\lambda}$ . Since 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  $\widehat{W}^\lambda(t) := \sqrt{\lambda}W^\lambda(t)$  and we similarly define  $\widehat{W}_I^\lambda(t) = \sqrt{\lambda}W_I^\lambda(t)$  and  $\widehat{W}_O^\lambda(t) = \sqrt{\lambda}W_O^\lambda(t)$ .

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

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

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  $(0, \infty)$ ,  $(\widehat{W}^\lambda, \widehat{W}_I^\lambda, \widehat{W}_O^\lambda) \Rightarrow (\widehat{W}, \widehat{W}_I, \widehat{W}_O)$  as  $\lambda \rightarrow \infty$ , where  $\widehat{W}_O$  is the diffusion limit of the virtual waiting time process in the  $GI/M/N_O^\lambda + M$  queue, and  $\widehat{W}_I \equiv \bar{K}_I/\nu$ . Moreover, for all  $t > 0$ ,

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

and

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

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  $\widehat{W}_O$  and  $\widehat{W}_I$  satisfy

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

where  $\text{disc}\{f\}$  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  $\mathbb{P}\{\widehat{W}^\lambda(t) > T\}$ . The result of Corollary (4.5) continues to hold for this indicator function  $f$  provided that  $K_I^\lambda/\sqrt{\lambda} \rightarrow \bar{K}_I \neq \nu T$ , as  $\lambda \rightarrow \infty$ . ■

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  $\gamma$  be such that

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

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

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

Then, one defines

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

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, as 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  $\{\widehat{X}_O^\lambda\}$  is C-Tight (see Section 15 of Billingsley [1968]). Such tightness is guaranteed, for example, if the sequence  $\{\widehat{X}_O^\lambda\}$  converges to a continuous limit, as is indeed the case under the modified definition of  $\widehat{X}_O^\lambda$  and for any of the parameter combinations of  $\gamma$  and  $\varsigma$  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  $\gamma \leq 1/2$  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  $1/2 < \gamma \leq 1$ . Finally, if  $\varsigma > 0$  and  $\gamma > 1/2$ , the  $GI/M/N_O^\lambda + M$  queue is equivalent (asymptotically) to an  $GI/M/\infty$  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  $\mu_I$  and  $\mu_O$  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  $\nu$ , which represents the “fluid” proxy for the fraction of  $\lambda$  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  $t_S \approx 3000/\lambda$  so that  $t_S$  is roughly the time it takes until 3000 customers have arrived to the system.<sup>2</sup>

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 6 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,

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

<sup>2</sup> The scaling of the sampling time has a strong justification within the analysis: recall that the process  $D_I^\lambda$  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  $\lambda$ .

The column Sim.  $\sigma$  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,

$$\text{Sim. } \sigma \approx \text{Th. } \sigma. \quad (14)$$

**Table 1** Simulation Results

#	Input							Output					
	$\lambda$	$N_I$	$N_O$	$K$	$q_1$	$q_2$	$t_S$	$p_1$	$p_2$	Joint	$p_1 \cdot p_2$	Sim $\sigma$	Th. $\sigma$
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

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  $\mathbb{P}\{W^\lambda(\infty) > 0\} \leq \alpha$ . 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 to note that  $\mathbb{1}\{W_O^\lambda(t) > 0\} = \mathbb{1}\{\widehat{X}_O^\lambda(t) \geq 0\}$  (customers have to wait only if all servers are busy).

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

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

and

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

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  $\zeta^*$  be such that  $\tilde{\zeta} = 2\zeta^*/(1 + \frac{1+\nu}{1-\nu})$  solves  $[1 + \frac{\tilde{\zeta}\Phi(\tilde{\zeta})}{\phi(\tilde{\zeta})}]^{-1} = \alpha/(1 - \nu)$  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  $\{\widehat{N}_O^\lambda\}$  defined through

$$\widehat{N}_O^\lambda = \lceil R_O^\lambda + \zeta^* \sqrt{R_O^\lambda} \rceil,$$

is asymptotically feasible for (16), i.e.  $\limsup_{\lambda \rightarrow \infty} \mathbb{P}\{\widehat{X}_O^\lambda(\infty) \geq 0\} \leq \alpha$ . It is also asymptotically optimal in that any other asymptotically feasible sequence can be at most  $o(\sqrt{R_O^\lambda})$  smaller than  $\widehat{N}_O^\lambda$ .

To establish the connection between (15) and (16) we observe that with the (sequence of) staffing levels  $\{\widehat{N}_O^\lambda\}$ , it holds that  $\widehat{X}_O^\lambda(\infty) \Rightarrow \widehat{X}_O(\infty)$  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  $X_I^\lambda(\infty)/\sqrt{\lambda} \Rightarrow \widehat{X}_I(0)$ . (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

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

where the last two equalities follow from Corollary 4.2 and the last inequality follows from our construction of the sequence  $\{\widehat{N}_O^\lambda\}$ . Thus, the sequence  $\{\widehat{N}_O^\lambda\}$  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  $\limsup_{\lambda \rightarrow \infty} \mathbb{P}\{\widehat{X}_O^\lambda(\infty) \geq 0, X_I^\lambda(\infty) = N_I + K_I\} \leq \alpha$ .

In fact, repeating the same argument for values of  $\zeta < \zeta^*$  shows that  $\widehat{N}_O^\lambda$  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  $\zeta^* = 1.6$  which yields (recall that  $R_O = 72$ )  $\widehat{N}_O^\lambda = \lceil 72 + 1.6\sqrt{72} \rceil = 86$ . We use  $t = 1000$  so as to be close to steady-state. We then simulate 10000 replications of the real system

with the above parameters, and find that the joint probability satisfies  $\mathbb{P}\{\widehat{X}_O^\lambda(t) \geq 0, X_I^\lambda(t) = N_I^\lambda + K_I^\lambda\} = 0.0992$ . 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 non-negligible 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 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 multi-class 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  $\lambda$  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 pre-specified 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.

## Acknowledgments

The authors are grateful to the review team for their careful review of the paper and their numerous helpful comments.

## References

2006. Call center management review: 2006 contact center outsourcing report. <http://www.callcenternews.com/resources/statistics.shtml>.
- Akşin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations 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, A.N., W. Chan, P. L'Ecuyer. 2009. Staffing multi-skill call centers via search methods and a performance approximation. *IIE Transactions* **41**(6) 483–497.
- Bassamboo, A., J.M. Harrison, A. Zeevi. 2009. Pointwise stationary fluid models for stochastic processing networks. *Manufacturing & Service Operations Management* **11**(1) 70–89.
- Bhulai, S., D. Roubos. 2010. Approximate dynamic programming techniques for skill-based routing in call centers. Working paper, Vrije Universiteit Amsterdam.
- Billingsley, P. 1968. *Convergence of Probability Measures*. J. Wiley & Sons, New York.
- Chevalier, P., R.A. Shumsky, N. Tabordon. 2004. Routing and staffing in large call centers with specialized and fully flexible servers. Working paper, Tuck School of Business, Hanover, New Hampshire.
- Chevalier, P., N. Tabordon. 2003. Overflow analysis and cross-trained servers. *International Journal of Production Economics* **85**(1) 47–60.
- Chevalier, P., J. C. Van den Schrieck. 2008. Optimizing the staffing and routing of small-size hierarchical call centers. *Production and Operations Management* **17**(3) 306–319.
- Dai, JG, S. He, T. Tezcan. 2010. Many-server diffusion limits for  $g/\text{ph}/n+g_i$  queues. *Annals of Applied Probability* **20**(5) 1854–1890.

- Frankx, G. J., G. Koole, A. Pot. 2006. Approximating multi-skill blocking systems by hyperexponential decomposition. *Performance Evaluation* **63** 799–824.
- Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.
- Gans, N., Y. P. Zhou. 2007. Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management* **9**(1) 33–50.
- Gurvich, I., W. Whitt. 2009a. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research* **34**(2) 363–396.
- Gurvich, I., W. Whitt. 2009b. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management* **11**(2) 237–253.
- Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. *Operations Research* **58**(2) 316–328.
- Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567–588.
- Heyman, DP. 1987. Asymptotic marginal independence in large networks of loss systems. *Annals of Operations Research* **8**(1) 57–73.
- Hunt, PJ, TG Kurtz. 1994. Large loss networks. *Stochastic Processes and their Applications* **53**(2) 363–378.
- Koçağa, Y.L., A.R. Ward. 2010. Admission control for a multi-server queue with abandonment. *Queueing Systems* **65**(3) 275 – 323.
- Koole, G., A. Pot. 2006. An overview of routing and staffing algorithms in multi-skill customer contact centers. Working paper, VU University, Amsterdam, Netherlands.
- Koole, G., J. Talim. 2000. Exponential approximation of multi-skill call centers architecture. *Proceedings of QNETs* **23** 1–10.
- Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research* **57**(5) 1189–1205.
- Massey, W.A., W. Whitt. 1996. Stationary-process approximations for the nonstationary erlang loss model. *Operations Research* **44**(6) 976–983.
- Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** 193–267.
- Perry, O., W. Whitt. 2009. Responding to unexpected overloads in large-scale service systems. *Management Science* **55**(8) 1353–1367.
- Perry, O., W. Whitt. 2010a. A fluid approximation for service systems responding to unexpected overloads. Working paper, Columbia University, New York, NY.
- Perry, O., W. Whitt. 2010b. A fluid limit for an overloaded  $x$  model via an averaging principle. Working paper, Columbia University, New York, NY.
- Perry, O., W. Whitt. 2010c. Gaussian approximations for an overloaded  $x$  model via an averaging principle. Working paper, Columbia University, New York, NY.
- Perry, O., W. Whitt. 2010d. An ode for an overloaded  $x$  model involving a stochastic averaging principle. Working paper, Columbia University, New York, NY.
- Pourbabai, B. 1987. Approximation of the overflow process from a  $G/M/N/K$  queueing system. *Management Science* **33**(7) 931–938.
- van Doorn, E.A. 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. *Advances in Applied Probability* **14**(1) 171–190.
- Whitt, W. 1984. Heavy traffic approximations for service systems with blocking. *AT&T Bell Lab. Tech. J* **63** 689–708.
- Whitt, W. 1991. The pointwise stationary approximation for  $M_t/M_t/s$  queues is asymptotically correct as the rates increase. *Management Science* **37**(3) 307–314.

- Whitt, W. 2002a. *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer Series in Operations Research, New York.
- Whitt, W. 2002b. *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer-Verlag, New York. .
- Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50**(10) 1449–1461.
- Whitt, W. 2005. Heavy-traffic limits for the  $G/H_2^*/n/m$  queue. *Mathematics of Operations Research* **30**(1) 1–27.
- Zhou, Y-P., Z. J. Ren. 2010. Service outsourcing. James J. Cochran, ed., *Wiley Encyclopedia of Operations Research and Management Science*. Wiley.

**This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.**

## Proofs and extensions

This e-companion is divided into two sections. In §EC.1 we prove all the results that appears in the body of the paper. In section EC.2 we consider the extension of the base model to a multi-class setting as the one in Figure 1(b) of the paper.

### EC.1. Proofs

We start by introducing some additional notational conventions. For two random variables  $X$  and  $Y$  we write  $X \leq_{st} Y$  when stochastic ordering holds in the standard sense. Namely, when  $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$  for every non-negative non-decreasing function  $f(\cdot)$  for which the expectations are defined. If  $X$  and  $Y$  are two  $\mathcal{D}$ -valued stochastic processes we will write  $\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0\}$  to denote the fact that the processes are ordered. In other words, that there exists a construction of the sample paths of  $X$  and  $Y$  such that, almost surely,  $X(t) \leq Y(t)$  for all  $t \geq 0$ .

The rest of the section is divided into two subsections. Subsection EC.1.1 establishes important properties of the (sequence of) availability processes  $D_I^\lambda$ , as defined in (8). Building on these, we then proceed to §EC.1.2 where we prove the main results sated in §4. The proofs of several auxiliary lemmas are relegated to §EC.1.3.

#### EC.1.1. The Availability Process

We first establish upper and lower bounds (in appropriate probabilistic sense) for the process  $D_I^\lambda$ , by relating it to simple  $M/M/1$  queues; see Lemmas EC.1.1 and EC.1.3. These bounds are then used to establish the pointwise stationarity of  $D_I^\lambda$  – see Theorem EC.1.4. A FCLT via an AP for an appropriately scaled version of the related cumulative process  $C_I^\lambda$  in (9) is stated and proved in Theorem EC.1.6.

We start with a useful comparison result. For the following, let  $Q_\epsilon^+ = (Q_\epsilon^+(t), t \geq 0)$  have the law of the number of customers in an  $M/M/1$  queue with arrival rate  $\nu + \epsilon$  and service rate 1, where  $\epsilon$  is taken to be small enough ensuring that the process  $Q_\epsilon^+$  is ergodic, i.e., that

$$\rho_\epsilon^+ := \nu + \epsilon < 1 \quad \text{for all } \epsilon \text{ small enough,} \quad (\text{EC.1})$$

where  $\rho_\epsilon^+$  is the utilization of this  $M/M/1$  queue. We add the subscript  $d$  to denote the initial value at time 0. That is,  $Q_{\epsilon,d}^+(t)$  corresponds to the queue length at time  $t$  of the  $M/M/1$  queue initialized (at time  $t = 0$ ) in state  $d$ . The initial condition can be random.

Note that  $D_I^\lambda$  is a Birth-and-Death (BD) process with death rate  $\lambda$  and birth rate in state  $d$  equal to  $\mu_I(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(K_I^\lambda - d)^+$ . Since  $\mu_I N_I^\lambda + \theta K_I^\lambda = \nu\lambda + o(\lambda)$  we have that

$$\mu_I(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(K_I^\lambda - d)^+ \leq \lambda(\nu + \epsilon) = \lambda\rho_\epsilon^+ \quad (\text{EC.2})$$

for all  $\lambda$  large enough and all  $d \in \mathbb{Z}_+$ . Hence, if we scale the birth (arrival) and death (service) rates of  $Q_\epsilon^+$  by  $\lambda$ , then both  $Q_\epsilon^+$  and  $D_I^\lambda$  have the same death rates, but the birth rates of  $Q_\epsilon^+$  are larger than those of  $D_I^\lambda$ . Scaling the rates of  $Q_\epsilon^+$  is tantamount to scaling its time argument by a factor of  $\lambda$ . We can thus prove the following ordering result using standard coupling arguments for BD processes (see, e.g., Lemma 1 in Whitt [1991]). The detailed proof is omitted. To simplify notation, we let

$$Y_d := D_I^\lambda(0) \quad \text{and} \quad Y_q := Q_\epsilon^+(0). \quad (\text{EC.3})$$

**LEMMA EC.1.1. (upper bound for  $D_I^\lambda$ )** Fix  $\epsilon > 0$  and  $\lambda$  large enough so that (EC.2) holds, and assume that  $Y_d \leq_{st} Y_q$ . Then,

$$\{D_I^\lambda(t), t \geq 0\} \leq_{st} \{Q_{\epsilon, Y_q}^+(\lambda t), t \geq 0\}.$$

The above ordering allows us to upper bound the sequence  $D_I^\lambda$  by a single (time scaled)  $M/M/1$  queue. The following auxiliary result is proved in §EC.1.3.

**LEMMA EC.1.2. (SSC in diffusion limit)** Suppose that (10) holds. Then, for all  $\eta, T > 0$ ,

$$\lim_{a \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq s \leq T} D_I^\lambda(s) \geq a\sqrt{\lambda} \right\} = 0 \quad \text{and} \quad \lim_{a \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{\eta \leq s \leq T} D_I^\lambda(s) \geq a \log \lambda \right\} = 0.$$

Consequently,

$$\lambda^{-1/2}(Q_I^\lambda - K_I^\lambda) \Rightarrow 0 \quad \text{in } \mathcal{D}[0, \infty) \text{ as } \lambda \rightarrow \infty.$$

Note that the final conclusion of the lemma implies, in particular, the convergence  $\widehat{X}_I^\lambda \Rightarrow 0e$  in Theorem 4.1. We interpret this result as a state-space collapse result. It shows that, in the diffusion limit the state in

station  $I$  is constant so that the station  $O$  captures the state of the network. However, as explained in the introduction and further discussed in §4.2, this SSC result is not sufficient for our needs (see Example 1).

In addition to  $Q_\epsilon^+(t)$ , which serves as a sample-path stochastic-order upper bound for  $D_I^\lambda$ , we introduce a process,  $Q_\epsilon^-$ , which corresponds to the queue-length process in a  $M/M/1$  queue with service rate 1 and arrival rate  $\nu - \epsilon$  and will serve as lower bound. The process  $Q_\epsilon^-$  provides a weaker bound than the sample-path stochastic-order upper bound provided by  $Q_\epsilon^+$ . That weaker bound is, however, sufficient for our needs. Below we choose  $\epsilon$  sufficiently small so that both  $\rho_\epsilon^+ < 1$  (as defined in (EC.1)) and  $\epsilon < \nu$ . As before, we add a subscript to make explicit the dependency on the initial condition at time 0. Recall the definition of  $Y_d$  in (EC.3).

**LEMMA EC.1.3. (lower bound for  $D_I^\lambda$ )** Assume that (10) holds and that  $Y_d^\lambda = O_P(\sqrt{\lambda})$ . Then, given  $T > 0$ , there exists a non-negative sequence  $\varepsilon_T^\lambda$  such that  $\varepsilon_T^\lambda \rightarrow 0$  as  $\lambda \rightarrow \infty$  and so that for all  $d_1 \in \mathbb{Z}_+$ ,

$$\mathbb{P} \{ D_I^\lambda(t) \geq d_1 \} \geq \mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^-(\lambda t) \geq d_1 \right\} - \varepsilon_T^\lambda, \quad \text{for all } t \in [0, T].$$

Note that the conditions of the lemma are satisfied, in particular, if  $Y_d^\lambda = b\sqrt{\lambda}$  for all  $\lambda$  and a non-random constant  $b \geq 0$ .

Since  $Q_\epsilon^-$  and  $Q_\epsilon^+$  have birth and death rates that do not scale with  $\lambda$ , one expects that  $Q_{\epsilon, Y_d^\lambda}^-(\lambda t)$  and  $Q_{\epsilon, Y_d^\lambda}^+(\lambda t)$  are close, in a sense, to their steady-state random variables  $Q_\epsilon^+(\infty)$  and  $Q_\epsilon^-(\infty)$  for all  $\lambda$  large enough and all  $t > 0$ . Since these steady-state random variables are also “close” to each other (for small values of  $\epsilon$ ), one expects the same to hold for the process  $D_I^\lambda$ , i.e., that  $D_I^\lambda(t)$  has approximately the distribution of  $Q_\epsilon^-(\infty)$  (or  $Q_\epsilon^+(\infty)$ ) for  $\lambda$  large enough and for all  $t > 0$ . The following theorem formalizes this intuition in showing that  $D_I^\lambda(t)$  converges to a local steady state instantaneously (pointwise, for each  $t > 0$ ) as  $\lambda \rightarrow \infty$ .

**THEOREM EC.1.4. (pointwise stationarity)** Fix a sequence  $Y_d^\lambda = O_P(\sqrt{\lambda})$ . Then, for all  $t_1 > t_0$  and all  $d_1 \in \mathbb{Z}_+$ ,

$$\mathbb{P} \left\{ D_I^\lambda(t_1) \geq d_1 \mid D_I^\lambda(t_0) = Y_d^\lambda \right\} \Rightarrow \nu^{d_1} \quad \text{in } \mathbb{R} \text{ as } \lambda \rightarrow \infty. \quad (\text{EC.4})$$

In particular, fixing  $t_0 = 0$  and assuming (10), we have for all  $t > 0$  that,

$$D_I^\lambda(t) \Rightarrow Q_b(\infty) \quad \text{in } \mathbb{R} \text{ as } \lambda \rightarrow \infty, \quad (\text{EC.5})$$

where  $\mathbb{P}\{Q_b(\infty) \geq d_1\} = \nu^{d_1}$ .

Note that the convergence in (EC.4) is convergence in distribution, since, for each fixed  $\lambda$ , the conditional probabilities are random variables rather than numbers.

To prove Theorem EC.1.4 we will need the following lemma, whose proof appears in §EC.1.3.

LEMMA EC.1.5. Let  $Y_d^\lambda = O_p(\sqrt{\lambda})$ . Then, for any  $t > 0$ ,

$$\lim_{\lambda \rightarrow \infty} \mathbb{P}\left\{Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1\right\} = \mathbb{P}\{Q_\epsilon^+(\infty) \geq d_1\} \quad \text{and} \quad \lim_{\lambda \rightarrow \infty} \mathbb{P}\left\{Q_{\epsilon, Y_d^\lambda}^-(\lambda t) \geq d_1\right\} = \mathbb{P}\{Q_\epsilon^-(\infty) \geq d_1\}.$$

*Proof of Theorem EC.1.4:* Since the process  $D_I^\lambda$  is Markovian it suffices to prove (EC.4) for  $t_0 = 0$  and  $t := t_1 - t_0$ . We start by fixing  $\delta > 0$ . We will show that, given  $\epsilon, \delta > 0$ , it holds for all sufficiently large  $\lambda$ , that

$$\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda\right\} \geq \mathbb{P}\{Q_\epsilon^-(\infty) \geq d_1\} - \delta\right\} \geq 1 - \epsilon, \quad (\text{EC.6})$$

and

$$\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda\right\} \leq \mathbb{P}\{Q_\epsilon^+(\infty) \geq d_1\} + \delta\right\} \geq 1 - \epsilon. \quad (\text{EC.7})$$

The steady-state distribution of  $Q_\epsilon^+$  is continuous in  $\epsilon$  in the sense that  $Q_\epsilon^+(\infty) \Rightarrow Q_b(\infty)$  and  $Q_\epsilon^-(\infty) \Rightarrow Q_b(\infty)$  as  $\epsilon \downarrow 0$ , and where  $Q_b(\infty)$  has the steady-state distribution of a  $M/M/1$  with service rate 1 and arrival rate  $\nu$ ; see e.g. Lemma 9.8 in Perry and Whitt [2010b] (where the statement is proved for a more general quasi-birth-and-death (QBD) process. Note that  $\alpha$  denotes the steady state distribution in that reference). In particular, for all sufficiently small  $\epsilon$

$$\left|\mathbb{P}\{Q_\epsilon^+(\infty) \geq d_1\} - \mathbb{P}\{Q_b(\infty) \geq d_1\}\right| \leq \delta \quad \text{and} \quad \left|\mathbb{P}\{Q_\epsilon^-(\infty) \geq d_1\} - \mathbb{P}\{Q_b(\infty) \geq d_1\}\right| \leq \delta.$$

Plugging this into (EC.6) and (EC.7) we then have that

$$\liminf_{\lambda \rightarrow \infty} \mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda\right\} \geq \mathbb{P}\{Q_b(\infty) \geq d_1\} - 2\delta\right\} \geq 1 - \epsilon$$

and

$$\liminf_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \right\} \leq \mathbb{P} \{ Q_b(\infty) \geq d_1 \} + 2\delta \right\} \geq 1 - \varepsilon.$$

Since  $\delta$  and  $\varepsilon$  are arbitrary we may conclude that  $\mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \right\}$  converges in probability (and thus in distribution) to the constant  $\mathbb{P} \{ Q_b(\infty) \geq d_1 \}$  and, because the distribution is discrete, the convergence is, in fact, uniform over compact subsets of  $\mathbb{Z}_+$ . In passing we note that the arguments to prove (EC.6) and (EC.7) can be repeated for any sequence of initial conditions  $Y_d^\lambda = O(\sqrt{\lambda})$ .

The remainder of the proof is dedicated to establishing (EC.6) and (EC.7), starting with the former. Note that

$$\begin{aligned} & \mathbb{P} \left\{ \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \right\} \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta \right\} \\ &= \mathbb{P} \left\{ \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \right\} \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta; Y_d^\lambda > b\sqrt{\lambda} \right\} \\ & \quad + \mathbb{P} \left\{ \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \right\} \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta; Y_d^\lambda \leq b\sqrt{\lambda} \right\} \\ & \geq \mathbb{P} \left\{ \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \right\} \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta; Y_d^\lambda \leq b\sqrt{\lambda} \right\}. \end{aligned}$$

From the monotonicity of  $D_I^\lambda$  in its initial condition, see, e.g., Chapter 9 in Ross [1996], we have that almost surely

$$\mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \right\} \geq \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = 0 \right\}, \quad (\text{EC.8})$$

and we note that, while the LHS of the inequality is a random variable, the probability on the RHS is a constant. We thus have

$$\begin{aligned} & \mathbb{P} \left\{ \mathbb{P} \{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \} \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta; Y_d^\lambda \leq b\sqrt{\lambda} \right\} \quad (\text{EC.9}) \\ &= \mathbb{P} \left\{ \mathbb{P} \{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda \} \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta \mid Y_d^\lambda \leq b\sqrt{\lambda} \right\} \mathbb{P} \{ Y_d^\lambda \leq b\sqrt{\lambda} \} \\ & \geq \mathbb{1} \left\{ \mathbb{P} \{ D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = 0 \} \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta \right\} \mathbb{P} \{ Y_d^\lambda \leq b\sqrt{\lambda} \} \\ & \geq \mathbb{1} \left\{ \mathbb{P} \{ Q_{\varepsilon,0}^-(\lambda t) \geq d_1 \} - \varepsilon_T^\lambda \geq \mathbb{P} \{ Q_\varepsilon^-(\infty) \geq d_1 \} - \delta \right\} \mathbb{P} \{ Y_d^\lambda \leq b\sqrt{\lambda} \}. \end{aligned}$$

Here, the first inequality follows from (EC.8) and noting that, conditioned on  $D_I^\lambda(0) = 0$ , the conditional probability becomes a constant, specifically, it is equal to either 0 or 1, so that it can be replaced

with the indicator. Using Lemma EC.1.5 we have, for all  $\lambda$  large enough, and for a fixed  $\delta > 0$ , that  $\mathbb{P}\{Q_{\epsilon,0}^-(\lambda t) \geq d_1\} - \varepsilon_T^\lambda \geq \mathbb{P}\{Q_\epsilon^-(\infty) \geq d_1\} - \delta$  (because  $\varepsilon_T^\lambda \rightarrow 0$  as  $\lambda \rightarrow \infty$ ), so that the indicator in (EC.9) is in fact 1 for all  $\lambda$  large enough. Also, by the assumption of the theorem, there exists  $b > 0$  such that for any given  $\varepsilon > 0$  and for all  $\lambda$  large enough,  $\mathbb{P}\{Y_d^\lambda \leq b\sqrt{\lambda}\} \geq 1 - \varepsilon$ . Hence, it follows from (EC.9) that for all  $\lambda$  large enough,

$$\mathbb{P}\left\{\mathbb{P}\left\{D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda\right\} \geq \mathbb{P}\{Q_\epsilon^-(\infty) \geq d_1\} - \delta\right\} \geq 1 - \varepsilon,$$

This establishes (EC.6).

The arguments for establishing (EC.7) are very similar. We replace (EC.9) with

$$\begin{aligned} & \mathbb{P}\left\{\mathbb{P}\{D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = Y_d^\lambda\} \leq \mathbb{P}\{Q_\epsilon^+(\infty) \geq d_1\} + \delta; Y_d^\lambda \leq b\sqrt{\lambda}\right\} \\ & \geq \mathbb{1}\left\{\mathbb{P}\{D_I^\lambda(t) \geq d_1 \mid D_I^\lambda(0) = b\sqrt{\lambda}\} \leq \mathbb{P}\{Q_\epsilon^-(\infty) \geq d_1\} + \delta\right\} \mathbb{P}\{Y_d^\lambda \leq b\sqrt{\lambda}\} \\ & \geq \mathbb{1}\left\{\mathbb{P}\{Q_{\epsilon,b\sqrt{\lambda}}^+(\lambda t) \geq d_1\} \leq \mathbb{P}\{Q_\epsilon^-(\infty) \geq d_1\} + \delta\right\} \mathbb{P}\{Y_d^\lambda \leq b\sqrt{\lambda}\}. \end{aligned}$$

Here, the first inequality follows again from the monotonicity of the process  $D_I^\lambda$  and from the fact that, conditioned on  $D_I^\lambda(0)$  being equal to a constant, the conditional probabilities become constants. The last inequality above follows from Lemma EC.1.1. Together with Lemma EC.1.5, (EC.7) is established.  $\blacksquare$

Next, we establish stochastic-process limits for the cumulative (in)availability process as defined in equation (9). We define the related scaled and centered process

$$\widehat{C}_I^\lambda(t) = \sqrt{\lambda} \left( C_I^\lambda(t) - \frac{(\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)}{\lambda} t \right), \quad t \geq 0.$$

Theorem EC.1.6 below shows that this scaling leads to meaningful limits, by establishing a FCLT for  $\widehat{C}_I^\lambda$  via the AP discussed in Remark 4.2. We note that, even though the AP is related to the pointwise stationarity in Theorem EC.1.4, it is not directly implied by it (nor does the AP imply Theorem EC.1.4). Both phenomena are closely related being consequences of the separation of time scales that is caused by the fast oscillations of  $D_I^\lambda$ .

**THEOREM EC.1.6. (FCLT via the AP)** Suppose that the conditions of Theorem 4.1 hold. Then  $\widehat{C}_I^\lambda \Rightarrow \check{\sigma}B$  in  $\mathcal{D}[0, \infty)$  as  $\lambda \rightarrow \infty$ , where  $B$  is a standard Brownian motion and  $\check{\sigma}^2 = 2\nu$ .

The fact that (time-scaled) cumulative processes associated with regenerative processes converge to Brownian limits under proper scalings of time and space is not new; see Glynn and Whitt [1993]. The important thing to observe here is that, in contrast to the results in Glynn and Whitt [1993], we do not scale time to get the Brownian limit. As the proof reveals, this is due to the fast oscillations of the process  $D_I^\lambda$ , which completes  $O(\lambda)$  (regenerative) cycles over any finite time interval. See also Perry and Whitt [2010c].

The proof of Theorem EC.1.6 builds on two steps: (i) transformation of the fast oscillations of the process  $D_I^\lambda$  over intervals of length  $[0, T)$  to oscillations of a related “slowed-down” process  $D_I^{s,\lambda}$  over intervals of length  $[0, \lambda T)$  (see (EC.10) below), and (ii) applying arguments in the spirit of Glynn and Whitt [1993] in the proof of the FCLT for the “slower” process  $D_I^{s,\lambda}$ .

We first define the “slowed down” process  $D_I^{s,\lambda}$ . To that end, recall that the process  $D_I^\lambda$  has the probability law of a single-server queue with state-dependent arrival rates. Specifically, the service rate is  $\lambda$ , while the arrival rate is  $\mu(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(K_I^\lambda - d)^+$  when in state  $d \in \{0, 1, \dots, N_I^\lambda + K_I^\lambda\}$ . We then define the “slowed down” processes for each  $\lambda$

$$D_I^{s,\lambda}(t) := D_I^\lambda(t/\lambda), \quad t \geq 0. \quad (\text{EC.10})$$

Then  $D_I^{s,\lambda}$  has the probability law of a state-dependent  $M/M/1$  queue with service rate  $\lambda/\lambda = 1$  and state-dependent arrival rate  $\lambda^{-1}(\mu(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(K_I^\lambda - d)^+)$ . Moreover,

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

By Assumption 1, the arrival rate of  $D_I^{s,\lambda}$  is bounded by  $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda = \nu + o(1)$  where  $\nu < 1$  so that  $D_I^{s,\lambda}$  is (for all  $\lambda$  large enough) a stable queue. We define  $\rho_d^\lambda$  to be the steady-state probability that the server is busy in this state-dependent  $M/M/1$  queue.

The following theorem restates Theorem EC.1.6 in terms of the slowed-down processes  $D_I^{s,\lambda}$ .

**THEOREM EC.1.7.** Under the conditions of Theorem 4.1,

$$\sqrt{\lambda} \left( \frac{1}{\lambda} \int_0^{\lambda \cdot} \mathbb{1}\{D_I^{s,\lambda}(u) = 0\} du - (1 - \rho_d^\lambda) e \right) \Rightarrow \tilde{\sigma} B \quad \text{in } \mathcal{D}[0, \infty) \text{ as } \lambda \rightarrow \infty,$$

where  $\tilde{\sigma}^2 = 2\nu$ . Furthermore,  $\sqrt{\lambda}((1 - \rho_d^\lambda) - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)/\lambda) \rightarrow 0$  as  $\lambda \rightarrow \infty$ , so that

$$\sqrt{\lambda} \left( \frac{1}{\lambda} \int_0^{\lambda \cdot} \mathbb{1}\{D_I^{s,\lambda}(u) = 0\} du - \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\lambda} e \right) \Rightarrow \tilde{\sigma} B \quad \text{in } \mathcal{D}[0, \infty) \text{ as } \lambda \rightarrow \infty.$$

**Proof:** To formally define the underlying regenerative process let  $\tau_{-1}^\lambda = 0$ , and for  $k \geq 0$ , define recursively

$$\tau_k^\lambda := \inf \{ u \geq \tau_{k-1}^\lambda : D_I^{s,\lambda}(u-) \neq 0, D_I^{s,\lambda}(u) = 0 \}.$$

For  $k \geq 0$ , define

$$\Psi_k^\lambda := \int_{\tau_{k-1}^\lambda}^{\tau_k^\lambda} (\mathbb{1} \{ D_I^{s,\lambda}(u) = 0 \} - (1 - \rho_d^\lambda)) du.$$

For  $k \geq 1$ ,  $T_k^\lambda := \tau_k^\lambda - \tau_{k-1}^\lambda$  is then the length of the  $k^{\text{th}}$  busy cycle (the busy cycle consists of the busy period and the idle period). Since  $\{T_k^\lambda, k \geq 1\}$  are IID, we will use  $T_1^\lambda$  to denote a general random variable with the distribution of a busy cycle.

Let  $\{\xi_k^\lambda, k \geq 1\}$  be IID exponential random variables with rate  $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$ , corresponding to the time spent at state 0 during a busy cycle and define  $\bar{T}_k^\lambda := \tau_k^\lambda - \tau_{k-1}^\lambda - \xi_k^\lambda$ . Then  $\{\bar{T}_k^\lambda, k \geq 1\}$  are IID with each having the distribution of the busy period of this state-dependent  $M/M/1$  queue. We can thus write

$$\Psi_k^\lambda = \xi_k^\lambda \rho_d^\lambda - (1 - \rho_d^\lambda) \bar{T}_k^\lambda, \quad k \geq 1, \quad (\text{EC.11})$$

where  $\Psi_0^\lambda = 0$  by definition. We define the corresponding (possibly delayed) renewal process

$$R^\lambda(t) := \sup \{ k \geq 0 : \tau_k^\lambda \leq t \}.$$

Since  $\Psi_0^\lambda = 0$ , we can write

$$\frac{1}{\lambda} \int_0^{\lambda t} \mathbb{1} \{ D_I^{s,\lambda}(u) = 0 \} du - (1 - \rho_d^\lambda)t = \sum_{k=1}^{R^\lambda(\lambda t)} \frac{\Psi_k^\lambda}{\lambda} + \frac{1}{\lambda} \int_{\tau_{R^\lambda(\lambda t)}^\lambda}^{\lambda t} (\mathbb{1} \{ D_I^{s,\lambda}(u) = 0 \} - (1 - \rho_d^\lambda)) du. \quad (\text{EC.12})$$

From here the argument follows very closely the standard proof of the FCLT for regenerative processes. Some care is needed because the IID random variables  $\{\Psi_k^\lambda\}_{k \geq 1}$  are indexed by both  $k$  and  $\lambda$ . This entails replacing some parts of the standard proof with an FCLT for triangular arrays. It also entails establishing bounds to identify the asymptotic variance terms.

First, using (EC.11), we write

$$\sqrt{\lambda} \sum_{k=1}^{R^\lambda(\lambda t)} \frac{\Psi_k^\lambda}{\lambda} = \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} \rho_d^\lambda \xi_k^\lambda - \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} (1 - \rho_d^\lambda) \bar{T}_k^\lambda$$

$$\begin{aligned}
&= \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} \left( \rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} \right) - \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} (1 - \rho_d^\lambda) (\bar{T}_k^\lambda - \mathbb{E} [\bar{T}_k^\lambda]) \\
&\quad + \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda t)} \left( \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} - (1 - \rho_d^\lambda) \mathbb{E} [\bar{T}_k^\lambda] \right). \tag{EC.13}
\end{aligned}$$

The next two lemmas identify asymptotic expressions of the expectation and variance terms of the cycle-related random variables. These expressions are applied to establish an FCLT for the partial sums in (EC.13).

The proofs of Lemmas EC.1.8 and EC.1.9 appear in §EC.1.3.

LEMMA EC.1.8. As  $\lambda \rightarrow \infty$ ,

$$\left( \sum_{k=1}^{[\lambda \cdot]} \text{Var} \left( \frac{\xi_k^\lambda}{\sqrt{\lambda}} \right), \sum_{k=1}^{[\lambda \cdot]} \text{Var} \left( \frac{\bar{T}_k^\lambda}{\sqrt{\lambda}} \right), \frac{R^\lambda(\lambda \cdot)}{\lambda} \right) \Rightarrow \left( \frac{1}{\nu^2} e, \frac{1+\nu}{(1-\nu)^3} e, \nu(1-\nu)e \right) \text{ in } \mathcal{D}[0, \infty). \tag{EC.14}$$

Consequently, as  $\lambda \rightarrow \infty$ ,

$$\left( \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda \cdot)} \left( \rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} \right), \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda \cdot)} (1 - \rho_d^\lambda) (\bar{T}_k^\lambda - \mathbb{E} [\bar{T}_k^\lambda]) \right) \Rightarrow \left( \sqrt{\nu(1-\nu)} B^1, \sqrt{\nu(1+\nu)} B^2 \right) \tag{EC.15}$$

in  $D^2[0, \infty)$ , where  $B^1$  and  $B^2$  are two independent standard Brownian motions.

LEMMA EC.1.9. Under the conditions of Theorem 4.1,

$$\sqrt{\lambda} \left( \frac{\rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} - (1 - \rho_d^\lambda) \mathbb{E} [\bar{T}_1^\lambda] \right) \rightarrow 0 \text{ as } \lambda \rightarrow \infty.$$

In turn,

$$\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{R^\lambda(\lambda \cdot)} \left( \frac{\rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} - (1 - \rho_d^\lambda) \mathbb{E} [\bar{T}_1^\lambda] \right) \Rightarrow 0 \text{ in } D[0, \infty) \text{ as } \lambda \rightarrow \infty. \tag{EC.16}$$

Proceeding with the proof of the theorem, if

$$\frac{1}{\lambda} \int_{\tau_{R^\lambda(\lambda t)}}^{\lambda t} (\mathbb{1} \{D_I^{s,\lambda}(u) = 0\} - (1 - \rho_d^\lambda)) du \Rightarrow 0 \text{ as } \lambda \rightarrow \infty, \tag{EC.17}$$

then the theorem is established by replacing the sum in (EC.12) with its version in (EC.13) and then replacing the right-hand side (RHS) in (EC.13) with (EC.15) and (EC.16), and noting the sum of the variance of

the Brownian motions  $B^1$  and  $B^2$  in (EC.15) is  $\tilde{\sigma}^2 = 2\nu$ . It thus remains only to establish (EC.17). We first note that for all  $u \geq 0$ ,  $|\mathbb{1}\{D_I^{s,\lambda}(u) = 0\} - (1 - \rho_d^\lambda)| \leq 2$  so that, to show (EC.17), it suffices to show that

$$\frac{2}{\sqrt{\lambda}} \left( \tau_{R^\lambda(\lambda t)+1}^\lambda - \tau_{R^\lambda(\lambda t)}^\lambda \right) \Rightarrow 0 \quad \text{as } \lambda \rightarrow \infty.$$

To that end, from the convergence of the variance terms in Lemma EC.1.8, it follows that  $\mathbb{E}[(T_k^\lambda)^2] \leq C$  for all  $\lambda$  large enough and a constant  $C$  that does not depend of  $\lambda$ . By Markov's inequality,  $\mathbb{P}\{T_k^\lambda > \epsilon\sqrt{\lambda}\} \leq C/(\epsilon^2\lambda)$  for all  $\lambda$  large enough. In turn, for any constant  $M > 0$ ,

$$\lambda^{-1/2} \max_{1 \leq k \leq \lambda M} T_k^\lambda \Rightarrow 0 \quad \text{as } \lambda \rightarrow \infty. \quad (\text{EC.18})$$

(See e.g. the proof of Lemma 2.3.1 in the online appendix to Whitt [2002a].) It follows from (EC.14) and (EC.18) that for any fixed  $T > 0$  and for  $M_T$  to be specified shortly,

$$\mathbb{P}\left\{ \sup_{0 \leq t \leq T} \frac{1}{\sqrt{\lambda}} \max_{1 \leq k \leq R^\lambda(\lambda t)+1} T_k^\lambda > \epsilon \right\} \leq \mathbb{P}\left\{ \frac{1}{\sqrt{\lambda}} \max_{1 \leq k \leq \lambda M_T} T_k^\lambda > \epsilon \right\} + \mathbb{P}\left\{ \sup_{0 \leq t \leq T} R^\lambda(\lambda t) + 1 > \lambda M_T \right\}.$$

Indeed, the first element in the RHS above converges to 0 by (EC.18), and the second element converges to 0 by (EC.14), provided that  $M_T > \nu(1 - \nu)T$ . Hence  $T_{R^\lambda(\lambda t)+1}^\lambda/\sqrt{\lambda} \Rightarrow 0$  in  $\mathcal{D}[0, \infty)$  as  $\lambda \rightarrow \infty$ . This concludes the proof of the theorem.  $\blacksquare$

## EC.1.2. Proofs of Main Results

In this section we prove the main results stated in §4 in the order of their appearance.

*Proof of Theorem 4.1:* The overflow process is the number of arrivals by time  $t$  that find  $X_I^\lambda$  equal to  $N_I + K_I$  upon arrival or, equivalently, find  $D_I^\lambda = 0$  when they arrive. Let  $t_k^\lambda$  be the arrival time of the  $k^{\text{th}}$  customer to arrive (the  $k^{\text{th}}$  jump of the exogenous arrival process  $A^\lambda(t)$ ). Then,  $A_O^\lambda(t) = \sum_{k=1}^{A^\lambda(t)} \mathbb{1}\{D_I^\lambda(t_k^\lambda -) = 0\}$  or, in simpler form,  $A_O^\lambda(t) = \int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} dA^\lambda(s)$ . In turn, the scaled process (see §3.2) satisfies

$$\begin{aligned} \widehat{A}_O^\lambda(t) &= \frac{\int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} dA^\lambda(s) - \lambda \int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} ds}{\sqrt{\lambda}} \\ &\quad + \frac{\lambda \int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} ds - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)t}{\sqrt{\lambda}} \end{aligned} \quad (\text{EC.19})$$

We treat each of the elements on the RHS of (EC.19) separately. For the first element define

$$\widehat{M}^\lambda(t) := \frac{1}{\sqrt{\lambda}} \left( \int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} dA^\lambda(s) - \lambda \int_0^t \mathbb{1}\{D_I^\lambda(s-) = 0\} ds \right)$$

$$= \frac{1}{\sqrt{\lambda}} \left( \int_0^t \mathbb{1} \{D_I^\lambda(s-) = 0\} d(A^\lambda(s) - \lambda s) \right).$$

Note that the process  $\widehat{M}^\lambda(t)$  is a square integrable martingale with respect to the filtration  $\mathcal{F}^\lambda = (\mathcal{F}_t^\lambda)_{t \geq 0}$ , where  $\mathcal{F}_t^\lambda = \sigma \{(D_I^\lambda(s), Q_O^\lambda(s), A^\lambda(s)); s \leq t\}$ ; having a predictable quadratic variation process  $\langle \widehat{M}^\lambda \rangle(t) = \int_0^t \mathbb{1} \{D_I^\lambda(s-) = 0\} ds$ ; see, e.g., Lemma 3.2 in Pang et al. [2007]. By Theorem EC.1.6 we have that

$$\frac{\lambda \int_0^\cdot \mathbb{1} \{D_I^\lambda(s-) = 0\} ds - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda) e}{\lambda} \Rightarrow 0 \quad \text{in } \mathcal{D}[0, \infty) \text{ as } \lambda \rightarrow \infty.$$

By Assumption 1,  $(\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda)/\lambda \rightarrow (1 - \nu)$  as  $\lambda \rightarrow \infty$ , so that

$$\langle \widehat{M}^\lambda \rangle(\cdot) = \int_0^\cdot \mathbb{1} \{D_I^\lambda(s-) = 0\} ds \Rightarrow (1 - \nu) e \quad \text{in } \mathcal{D}[0, \infty) \text{ as } \lambda \rightarrow \infty.$$

From here it follows by the Martingale FCLT (see, e.g., Theorem 8.1 in Pang et al. [2007]) that

$$\frac{\int_0^\cdot \mathbb{1} \{D_I^\lambda(s-) = 0\} dA^\lambda(s) - \lambda \int_0^\cdot \mathbb{1} \{D_I^\lambda(s-) = 0\} ds}{\sqrt{\lambda}} \Rightarrow \sqrt{(1 - \nu) B^1} \quad \text{in } \mathcal{D}[0, \infty) \text{ as } \lambda \rightarrow \infty.$$

For the second element on the RHS of (EC.19) we have by Theorem EC.1.6, that

$$\frac{\lambda \int_0^\cdot \mathbb{1} \{D_I^\lambda(s) = 0\} ds - (\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda) e}{\sqrt{\lambda}} \Rightarrow \sqrt{2\nu} B^2 \quad \text{in } \mathcal{D}[0, \infty) \text{ as } \lambda \rightarrow \infty.$$

The convergence of  $\widehat{A}_O^\lambda$  now follows from these last two limits and from the continuity of the addition mapping at continuous limits.

Finally, by Lemma EC.1.2 we have that  $\widehat{X}_I^\lambda \Rightarrow 0$  in  $\mathcal{D}(0, \infty)$  as  $\lambda \rightarrow \infty$ . The marginal convergence of  $\widehat{A}_O^\lambda$ ,  $\widehat{X}_I^\lambda$  and  $\langle \widehat{M}^\lambda \rangle$  now implies their joint convergence because  $\widehat{X}_I^\lambda$  and  $\langle \widehat{M}^\lambda \rangle$  have deterministic limits; see e.g. Theorem 11.4.5 in Whitt [2002a]. ■

*Proof of Corollary 4.2:* Recall that  $X_I^\lambda(\infty)$  has the distribution of the steady-state number of customers in an  $M/M/N_I^\lambda/K_I^\lambda + M$  queue with arrival rate  $\lambda$ ,  $N_I^\lambda$  servers, service rate  $\mu$  and waiting room of size  $K_I^\lambda$ . Also

$$\widehat{X}_I^\lambda(t) = \frac{X_I^\lambda(t) - (N_I^\lambda - K_I^\lambda)}{\sqrt{\lambda}} = -\frac{D_I^\lambda(t)}{\sqrt{\lambda}} \geq 0.$$

We will show that

$$\mathbb{E}[|\widehat{X}_I^\lambda(\infty)|] \rightarrow 0 \quad \text{as } \lambda \rightarrow \infty, \tag{EC.20}$$

so that, in particular,  $\widehat{X}_I^\lambda(\infty) \Rightarrow 0$ . This would imply that, initializing station  $I$  at time  $t = 0$  with its steady-state distribution,  $\{\widehat{X}_I^\lambda(0)\}$  satisfies (10) so that the convergence of the scaled overflow process follows from Theorem 4.1. In fact, (11) also follows from (EC.20). Indeed, since inflow=outflow in steady-state we have  $\lambda(1 - p_b^\lambda) = \mu\mathbb{E}[Z_I^\lambda(\infty)] + \theta\mathbb{E}[Q^\lambda(\infty)]$  where  $Z_I^\lambda(\infty)$  has the steady-state distribution of the number of busy servers in station  $I$ . By work conservation in station  $I$  we have that  $Z_I^\lambda(\infty) = X_I^\lambda(\infty) \wedge N_I^\lambda$  and  $Q^\lambda(\infty) = (X_I^\lambda(\infty) - N_I^\lambda)^+$ . Then, by (EC.20),

$$\mathbb{E}[Z_I^\lambda(\infty)] = N_I^\lambda - \mathbb{E}[(X_I^\lambda(\infty) - N_I^\lambda)^-] = N_I^\lambda - o(\sqrt{\lambda})$$

and

$$\mathbb{E}[Q^\lambda(\infty)] = \mathbb{E}[(X_I^\lambda(\infty) - N_I^\lambda)^+] = K_I^\lambda + o(\sqrt{\lambda})$$

and, in turn, that  $\lambda(1 - p_b^\lambda) = \mu_I N_I^\lambda + \theta K_I^\lambda + o_p(\sqrt{\lambda})$ . Equation (11) is now obtained by dividing by  $\lambda$ .

To conclude the proof it remains to prove (EC.20). This, we claim, follows immediately from Lemma EC.1.1. Indeed, from that lemma we have that for every  $\epsilon > 0$  we can find  $\lambda$  large enough, so that  $D_I^\lambda(\infty) \leq_{st} Q_\epsilon^+(\infty)$ , where  $Q_\epsilon^+(\infty)$  has the steady-state distribution of an  $M/M/1$  queue with utilization  $\rho_\epsilon^+ < 1$ , for  $\rho_\epsilon^+$  in (EC.1). Hence,

$$\mathbb{E}[D_I^\lambda(\infty)] \leq_{st} \mathbb{E}[Q_\epsilon^+(\infty)] = O_p(1) = o_p(\sqrt{\lambda}).$$

In turn, after dividing by  $\sqrt{\lambda}$  we get  $\mathbb{E}[|\widehat{X}_I^\lambda(\infty)|] = \mathbb{E}[D_I^\lambda(\infty)/\sqrt{\lambda}] = o(1)$ . This concludes the proof. ■

*Proof of Theorem 4.3* Define for every  $\lambda$  the filtration  $\mathcal{F}^\lambda = (\mathcal{F}_t^\lambda)_{t \geq 0}$  by

$$\mathcal{F}_t^\lambda = \sigma \{D_I^\lambda(s), X_O^\lambda(s); \quad s \leq t\},$$

and consider the filtered probability space  $(\Omega, \mathbb{F}^\lambda, (\mathcal{F}_t^\lambda)_{t \geq 0}, \mathbb{P})$  (where  $\mathbb{F}^\lambda$  is the  $\sigma$ -algebra over  $\Omega$ , and  $\mathcal{F}_t^\lambda \subset \mathbb{F}^\lambda$  for all  $t \geq 0$ ).

First we claim that, from the pointwise stationarity of  $D_I^\lambda$  (see Theorem EC.1.4), it follows that for each  $t > 0$  and  $\delta \in (0, t)$ , and  $d \in \mathbb{Z}_+$ ,

$$\mathbb{P} \left\{ D_I^\lambda(t) = d \mid \mathcal{F}_{t-\delta}^\lambda \right\} \Rightarrow \mathbb{P} \{ Q_b(\infty) = d \}, \text{ as } \lambda \rightarrow \infty. \quad (\text{EC.21})$$

In words,  $D_I^\lambda(t)$  is asymptotically independent of  $\mathcal{F}_{t-\delta}^\lambda$ . Indeed, for all  $0 \leq s < t$ ,  $D_I^\lambda(t) - D_I^\lambda(s)$  depends only on the arrivals on the interval  $(s, t]$ , the abandonments and the service completions on that interval, all of which, conditioned on  $D_I^\lambda(s)$ , are independent of  $\mathcal{F}_s$ . Note also that  $D_I^\lambda$  is Markov, so that  $\mathbb{E} \left[ \mathbb{1} \{D_I^\lambda(t) = d\} \mid \mathcal{F}_{t-\delta}^\lambda \right] = \mathbb{E} \left[ \mathbb{1} \{D_I^\lambda(t) = d\} \mid D_I^\lambda(t-\delta) \right]$ . By Lemma EC.1.2,  $\sup_{0 \leq s \leq t-\delta} D_I^\lambda(u) = O_P(\sqrt{\lambda})$ , so that the (EC.21) follows from Theorem EC.1.4.

We next show how the statement of Theorem 4.3 follows from (EC.21). For all strictly positive  $T, \epsilon$  and  $\delta$  define

$$E_T := E_T(\lambda, \delta, \epsilon) = \left\{ \omega \in \Omega : \sup_{s, t \leq T: |t-s| \leq \delta} |\widehat{X}_O^\lambda(t) - \widehat{X}_O^\lambda(s)| \leq \epsilon \right\}.$$

We denote by  $E_T^c$  the complement of  $E_T$ , i.e.,  $E_T^c := \{\omega \in \Omega : \omega \notin E_T\}$ . (We omit the dependency of  $\widehat{X}_O^\lambda$  on  $\omega$  to simplify notation, with the understanding that  $\widehat{X}_O^\lambda(t) = \widehat{X}_O^\lambda(t, \omega)$ .) From the  $\mathcal{C}$ -Tightness of  $\widehat{X}_O^\lambda$  (see e.g. Theorem 3.1 of Whitt [2005] for the convergence of the underlying sequence of  $GI/M/N_O^\lambda + M$  queues to a continuous limit), it follows that

$$\lim_{\delta \rightarrow 0} \liminf_{\lambda \rightarrow \infty} \mathbb{P} \{E_T(\lambda, \delta, \epsilon)\} = 1. \quad (\text{EC.22})$$

We write

$$\begin{aligned} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} &= \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d, \mathbb{1} \{E_T\} \right\} \\ &+ \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d, \mathbb{1} \{E_T^c\} \right\}. \end{aligned} \quad (\text{EC.23})$$

From (EC.22) it then follows that

$$\mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} = \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d, \mathbb{1} \{E_T\} \right\} + o(1). \quad (\text{EC.24})$$

Note that for all  $\omega \in E_T$ ,  $t \leq T$  and  $q \in \mathbb{R}^+$

$$\left\{ \widehat{X}_O^\lambda(t-\delta) > q - \epsilon \right\} \subset \left\{ \widehat{X}_O^\lambda(t) > q \right\}. \quad (\text{EC.25})$$

By the same argument that leads to (EC.24) we have that

$$\mathbb{P} \left\{ \widehat{X}_O^\lambda(t-\delta) > q - \epsilon, D_I^\lambda(t) = d, \mathbb{1} \{E_T\} \right\} = \mathbb{P} \left\{ \widehat{X}_O^\lambda(t-\delta) > q - \epsilon, D_I^\lambda(t) = d \right\} + o(1),$$

so that, by (EC.25),

$$\begin{aligned} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d, \mathbb{1} \{E_T\} \right\} &\geq \mathbb{P} \left\{ \widehat{X}_O^\lambda(t - \delta) > q - \epsilon, D_I^\lambda(t) = d \right\} + o(1) \\ &= \mathbb{E} \mathbb{E} \left[ \mathbb{1} \left\{ \widehat{X}_O^\lambda(t - \delta) > q - \epsilon \right\} \mathbb{1} \{D_I^\lambda(t) = d\} \middle| \mathcal{F}_{t-\delta}^\lambda \right] + o(1) \\ &= \mathbb{E} \left[ \mathbb{1} \left\{ \widehat{X}_O^\lambda(t - \delta) > q - \epsilon \right\} \right] \mathbb{E} \left[ \mathbb{1} \{D_I^\lambda(t) = d\} \middle| \mathcal{F}_{t-\delta}^\lambda \right] + o(1). \end{aligned}$$

The last equality above follows from the fact that,  $\widehat{X}_O^\lambda(t - \delta)$  is measurable with respect to  $\mathcal{F}_{t-\delta}^\lambda$ . Using (EC.24) and applying (EC.21), we have that for all  $\epsilon > 0$ ,

$$\begin{aligned} \liminf_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} &\geq \lim_{\delta \rightarrow 0} \liminf_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t - \delta) > q - \epsilon \right\} \mathbb{P} \left\{ D_I^\lambda(t) = d \middle| \mathcal{F}_{t-\delta}^\lambda \right\} \\ &= \lim_{\delta \rightarrow 0} \mathbb{P} \left\{ \widehat{X}_O(t - \delta) > q - \epsilon \right\} \mathbb{P} \{Q_b(\infty) = d\} \\ &= \mathbb{P} \left\{ \widehat{X}_O(t) > q - \epsilon \right\} \mathbb{P} \{Q_b(\infty) = d\}. \end{aligned}$$

The equalities above follows from the convergence of the sequence  $\{\widehat{X}_O^\lambda\}$  to a continuous limit  $\widehat{X}_O$  (see Theorem 3.1 in Whitt [2005]) and the continuity of the projection mapping in continuous limits (see e.g. §14 of Billingsley [1968]). In fact, the limit process in Whitt [2005] is a diffusion process with continuous drift and constant diffusion coefficients. In turn, for each  $t > 0$ , the random variable  $\widehat{X}_O(t)$  has a density (see, e.g., pages 368-369 in Karatzas and Shreve [1991]) so that

$$\begin{aligned} \liminf_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} &\geq \lim_{\epsilon \rightarrow 0} \mathbb{P} \left\{ \widehat{X}_O(t) > q - \epsilon \right\} \mathbb{P} \{Q_b(\infty) = d\} \\ &= \mathbb{P} \left\{ \widehat{X}_O(t) > q \right\} \mathbb{P} \{Q_b(\infty) = d\}. \end{aligned} \tag{EC.26}$$

To prove the other direction, note that on the set  $E_T$ ,  $\left\{ \widehat{X}_O^\lambda(t) > q - \epsilon \right\} \subset \left\{ \widehat{X}_O^\lambda(t - \delta) > q \right\}$ . Arguing as before we have

$$\begin{aligned} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q - \epsilon, D_I^\lambda(t) = d \right\} &\leq \lim_{\delta \rightarrow 0} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t - \delta) > q \right\} \mathbb{P} \{D_I^\lambda(t) = d\} \\ &= \mathbb{P} \left\{ \widehat{X}_O(t) > q \right\} \mathbb{P} \{Q_b(\infty) = d\}, \end{aligned}$$

So that, upon taking limits with  $\epsilon \downarrow 0$ , we get from the  $\mathcal{C}$ -tightness of  $\widehat{X}_O^\lambda$  that

$$\limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} \leq \mathbb{P} \left\{ \widehat{X}_O(t) > q \right\} \mathbb{P} \{Q_b(\infty) = d\}. \tag{EC.27}$$

It follows from (EC.26)-(EC.27) and (EC.5) that for all  $q > 0$  and  $d \in \mathbb{Z}_+$ ,

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q, D_I^\lambda(t) = d \right\} &= \mathbb{P} \left\{ \widehat{X}_O(t) > q \right\} \mathbb{P} \{Q_b(\infty) = d\} \\ &= \lim_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \widehat{X}_O^\lambda(t) > q \right\} \mathbb{P} \{D_I^\lambda(t) = d\}. \end{aligned}$$

The independence with  $\widehat{Q}_O^\lambda(t)$  replacing  $\widehat{X}_O^\lambda(t)$  follows from the above argument noting that  $\widehat{Q}_O^\lambda(t) = [\widehat{X}_O^\lambda(t) - (N_O^\lambda - R_O^\lambda)/\sqrt{\lambda}]^+$  where, by Assumption 1,  $N_O^\lambda - R_O^\lambda = \varsigma \sqrt{(1-\nu)/\mu_O} \sqrt{\lambda} + o(\sqrt{\lambda})$ . This concludes the proof.  $\blacksquare$

*Proof of Corollary 4.4:* We first consider the virtual wait in stations  $I$  and  $O$ . To that end let  $L_I^\lambda(t)$  and  $L_O^\lambda(t)$  be the number of abandonments from station  $I$  and station  $O$ , respectively, by time  $t$ . Let  $S_I^\lambda(t)$  and  $S_O^\lambda(t)$  be the number of service completions at stations  $I$  and  $O$ , respectively by time  $t$ . Recall that  $A^\lambda(t)$  is the number of exogenous arrivals to station  $I$  by time  $t$  and  $A_O^\lambda(t)$  is the number of overflows by time  $t$  that correspond, in turn, to arrivals to station  $O$ . Following Talreja and Whitt [2009], given  $t \geq 0$  and  $u \geq 0$  we define  $S_I^{\lambda,t}(t+u)$  and  $L_I^{\lambda,t}(t+u)$  to be the service completion and abandonment by time  $t+u$  assuming that the arrival process  $A^\lambda$  is stopped at time  $t$ . Similarly we define  $S_O^{\lambda,t}(t+u)$  and  $L_O^{\lambda,t}(t+u)$ . Then, the virtual waiting time at a fixed time  $t$  satisfies

$$W_I^\lambda(t) = \inf \{ u \geq 0 : S_I^{\lambda,t}(t+u) - S_I^{\lambda,t}(t) + L_I^{\lambda,t}(t+u) - L_I^{\lambda,t}(t) \geq Q_I^\lambda(t) \}$$

$$W_O^\lambda(t) = \inf \{ u \geq 0 : S_O^{\lambda,t}(t+u) - S_O^{\lambda,t}(t) + L_O^{\lambda,t}(t+u) - L_O^{\lambda,t}(t) \geq Q_O^\lambda(t) \},$$

Due to the exponential service times and patience,  $W_I^\lambda(t)$  depends on the past only via  $X_I^\lambda(t)$  and  $Q_I^\lambda(t)$ . Similarly,  $W_O^\lambda(t)$  depends on the past only via  $X_O^\lambda(t)$  and  $Q_O^\lambda(t)$ . Since  $Q_O^\lambda(t) = [X_O^\lambda(t) - N_O^\lambda]^+$  and  $Q_I^\lambda(t) = [X_I^\lambda(t) - N_I^\lambda]^+$ , the dependence is in fact only via  $X_O^\lambda(t)$ . Theorem 4.3 shows that  $\widehat{X}_O^\lambda(t)$  is asymptotically independent of  $D_I^\lambda(t)$  so the same is true for  $\widehat{W}_O^\lambda(t)$ . Consequently, we have for every bounded continuous function  $f$  that

$$\mathbb{E} \left[ f(\widehat{W}^\lambda(t)) \middle| D_I^\lambda(t) = 0 \right] = \mathbb{E} \left[ f(\widehat{W}_O^\lambda(t)) \right] \mathbb{P} \{ D_I^\lambda(t) = 0 \} + o(1). \quad (\text{EC.28})$$

By Theorem 4.1, and under the condition of the corollary, we have that  $\widehat{Q}_I^\lambda \Rightarrow \bar{K}$  in  $\mathcal{D}[0, \infty)$ , where  $\bar{K}$  is defined in the statement of the corollary. Also, by Theorem 4.1 it holds that  $A_I^\lambda/\lambda = A^\lambda/\lambda - A_O^\lambda/\lambda \Rightarrow \nu e$  in  $\mathcal{D}[0, \infty)$  and it is easy to show that, under the condition of the corollary,  $S_I^\lambda/\lambda + L_I^\lambda/\lambda \Rightarrow \nu e$ . Applying Theorem 3.1 in Talreja and Whitt [2009]<sup>3</sup> we have that

$$\widehat{W}_I^\lambda \Rightarrow \widehat{W}_I = \bar{K}/\nu \quad \text{in } \mathcal{D}[0, \infty). \quad (\text{EC.29})$$

<sup>3</sup> We note that Theorem 3.1 in Talreja and Whitt [2009] is not stated for queues with finite waiting room and, moreover, it requires that both  $S_I^\lambda/\lambda$  and  $L_I^\lambda/\lambda$  converge and not just the sum. Nevertheless, it is easy to verify that the result does apply here.

Consequently,

$$\mathbb{E} \left[ f(\widehat{W}_I^\lambda(t)) \mathbb{1}\{D_I^\lambda(t) > 0\} \right] = \mathbb{E}[f(\widehat{W}_I^\lambda(t))] \mathbb{P}\{D_I^\lambda(t) > 0\} + o(1). \quad (\text{EC.30})$$

Combining (EC.28) and (EC.30) we have that

$$\begin{aligned} \mathbb{E} \left[ f(\widehat{W}^\lambda(t)) \right] &= \mathbb{E} \left[ f(\widehat{W}^\lambda(t)) \mathbb{1}\{D_I^\lambda(t) > 0\} \right] \mathbb{P}\{D_I^\lambda(t) > 0\} \\ &\quad + \mathbb{E} \left[ f(\widehat{W}^\lambda(t)) \mathbb{1}\{D_I^\lambda(t) = 0\} \right] \mathbb{P}\{D_I^\lambda(t) = 0\} \\ &= \mathbb{E} \left[ f(\widehat{W}_I^\lambda(t)) \right] \mathbb{P}\{D_I^\lambda(t) > 0\} + \mathbb{E} \left[ f(\widehat{W}_O^\lambda(t)) \right] \mathbb{P}\{D_I^\lambda(t) = 0\} + o(1). \end{aligned} \quad (\text{EC.31})$$

Noting that, by Theorem EC.1.4 and Corollary 4.2 we have  $\mathbb{P}\{D_I^\lambda(t) = 0\} = p_b^\lambda + o(1)$  for all  $t > 0$ , we conclude that

$$\mathbb{E} \left[ f(\widehat{W}^\lambda(t)) \right] = \mathbb{E} \left[ f(\widehat{W}_I^\lambda(t)) \right] (1 - p_b^\lambda) + \mathbb{E} \left[ f(\widehat{W}_O^\lambda(t)) \right] p_b^\lambda + o(1).$$

We turn to prove the second part of the corollary. To that end, write

$$\mathbb{E} \left[ \frac{1}{A^\lambda(t)} \sum_{k=1}^{A^\lambda(t)} f(\sqrt{\lambda} w_k^\lambda) \right] = \mathbb{E} \left[ \frac{1}{A^\lambda(t)} \sum_{k=1}^{A_I^\lambda(t)} f(\sqrt{\lambda} w_{k,I}^\lambda) \right] + \mathbb{E} \left[ \frac{1}{A^\lambda(t)} \sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda} w_{k,O}^\lambda) \right] \quad (\text{EC.32})$$

Considering first the second element on the right hand side, note that

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

By Theorem 4.1 and the strong law for renewal processes we have that, for each  $t > 0$ ,  $|A_O^\lambda(t)/A^\lambda(t) - p_b^\lambda| \Rightarrow 0$ , and the convergence also holds in expectation since  $A_O^\lambda(t)/A^\lambda(t) \leq 1$  for each  $t > 0$ . Using the fact that  $f$  is bounded we have that

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

and, in turn, that

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

A similar analysis applies then to the first element on the right hand side of (EC.32). ■

*Proof of Corollary 4.5:* Note that by the functional strong law of large numbers applied to the Poisson process  $A^\lambda$  we have that  $A^\lambda/\lambda \Rightarrow e$  in  $\mathcal{D}[0, \infty)$ . By Theorem 4.1 we have that  $A_O^\lambda/\lambda \Rightarrow (1-\nu)e$  in  $\mathcal{D}[0, \infty)$ . In turn, we also have that  $A_I^\lambda/\lambda \Rightarrow \nu e$ . Finally, by Theorem 5.2 of Talreja and Whitt [2009] we also have that  $\widehat{W}_O^\lambda \Rightarrow \widehat{W}_O = [\widehat{X}_O]^+/(1-\nu)$  where  $\widehat{X}_O$  is the limit for the scaled and centered head-count process in the associated sequence of  $GI/M/N_O^\lambda + M$  queues as in Theorem 3.1 of Whitt [2005]. In fact, Theorem 5.2 in Talreja and Whitt [2009] is for the  $M/M/N + M$  queue but the result applies identically for the  $GI/M/N + M$  replacing the appropriate limits for the head-count process of the  $M/M/N + M$  queue (Theorem 5.1 there) with those of the  $GI/M/N + M$  in Whitt [2005].

Combining all of the above and since all but  $\widehat{W}_O^\lambda$  converge to non-random limits we can conclude (see Theorem 11.4.5 in Whitt [2002a]) the joint convergence

$$\left( \frac{A_I^\lambda}{\lambda}, \frac{A_O^\lambda}{\lambda}, \widehat{W}_I^\lambda, \widehat{W}_O^\lambda \right) \Rightarrow \left( \nu e, (1-\nu)e, \bar{K}/\nu, \widehat{W}_O \right) \quad \text{in } \mathcal{D}^4[0, \infty) \text{ as } \lambda \rightarrow \infty. \quad (\text{EC.33})$$

From here the proof of the corollary follows from Corollary 4.4 by applying results in stochastic integration.

Note first that we can write

$$\frac{1}{A_I^\lambda(t)} \sum_{k=1}^{A_I^\lambda(t)} f(\sqrt{\lambda} w_{k,I}^\lambda) = \frac{1}{A_I^\lambda(t)} \int_0^t f(\widehat{W}_I^\lambda(s-)) dA_I^\lambda(s),$$

and

$$\frac{1}{A_O^\lambda(t)} \sum_{k=1}^{A_O^\lambda(t)} f(\sqrt{\lambda} w_{k,O}^\lambda) = \frac{1}{A_O^\lambda(t)} \int_0^t f(\widehat{W}_O^\lambda(s-)) dA_O^\lambda(s).$$

We will next use the following general result (whose proof appears in §EC.1.3)

LEMMA EC.1.10. Fix  $t > 0$ . Let  $\{(\Psi^\lambda, Y^\lambda)\}$  be a sequence of processes in  $\mathcal{D}^2[0, t]$  such that, for all  $\lambda$ ,  $Y^\lambda$  is increasing with  $Y^\lambda(0) = 0$  and such that  $(\Psi^\lambda, Y^\lambda) \Rightarrow (\Psi, Y)$  as  $\lambda \rightarrow \infty$  in the  $J_1$  metric where  $Y$  is a continuous process. Let  $f: \mathbb{R} \rightarrow \mathbb{R}_+$  be such that

$$\int_0^t \mathbb{1}\{\Psi(s-) \in \text{disc}\{f\}\} dY(s) = 0 \quad \text{almost surely,} \quad (\text{EC.34})$$

where  $\text{disc}\{f\}$  is the set of discontinuity points of the function  $f$ . Then for any  $s \in [0, t]$ ,

$$\int_0^s f(\Psi^\lambda(u-)) dY^\lambda(u) \Rightarrow \int_0^s f(\Psi(u-)) dY(u), \text{ as } \lambda \rightarrow \infty.$$

Proceeding with the proof of Corollary 4.5 and fixing  $t$ , define for  $0 \leq s \leq t$ , the processes  $Y_I^\lambda(s) := A_I^\lambda(s)/A_I^\lambda(t)$  and  $Y_O^\lambda(s) = A_O^\lambda(s)/A_O^\lambda(t)$ . By the strong law for renewal process we then have that, uniformly over compact subsets of  $[0, t]$ , both  $Y_I^\lambda \Rightarrow e(\cdot)/t$  and  $Y_O^\lambda \Rightarrow e(\cdot)/t$ . Also, by the first part of the corollary, we have that  $\widehat{W}_I^\lambda \Rightarrow \widehat{W}_I$  and  $\widehat{W}_O^\lambda \Rightarrow \widehat{W}_O$ . Since the limit processes are continuous this implies convergence together in  $D^2[0, T]$  and, in particular, that each of the sequences  $\{(\widehat{W}_I^\lambda, Y_I^\lambda)\}$  and  $\{(\widehat{W}_O^\lambda, Y_O^\lambda)\}$  satisfies the conditions of Lemma EC.1.10. Moreover, since the function  $f$  in the corollary is assumed to be continuous (EC.34) holds trivially. Consequently, we conclude that

$$\frac{1}{A_I^\lambda(t)} \int_0^t f(\widehat{W}_I^\lambda(s-)) dA_I^\lambda(s) \Rightarrow \frac{1}{t} \int_0^t f(\widehat{W}_I(s-)) ds = \frac{1}{t} \int_0^t f(\widehat{W}_I(s)) ds,$$

where the equalities hold almost surely using the boundedness of  $f$ , and similarly

$$\frac{1}{A_O^\lambda(t)} \int_0^t f(\widehat{W}_O^\lambda(s-)) dA_O^\lambda(s) \Rightarrow \frac{1}{t} \int_0^t f(\widehat{W}_O(s-)) ds = \frac{1}{t} \int_0^t f(\widehat{W}_O(s)) ds.$$

Since the function  $f$  is bounded and since  $A_I^\lambda(s)/A_I^\lambda(t) \leq 1$  and  $A_O^\lambda(s)/A_O^\lambda(t) \leq 1$  for all  $s \leq t$ , the convergence also holds in expectations. The proof of the corollary is thus complete.  $\blacksquare$

### EC.1.3. Proofs of auxiliary results

*Proof of Lemma EC.1.2:* Given the ordering in Lemma EC.1.1, the claim of the lemma is implied by the following result for  $M/M/1$  queues: Fix  $\epsilon > 0$  such that (EC.1) and (EC.2) hold. Then

$$\lim_{a \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \geq a\sqrt{\lambda} \right\} = 0 \quad \text{and} \quad (\text{EC.35})$$

$$\lim_{a \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{\eta \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \geq a \log \lambda \right\} = 0, \quad \text{for all } \eta > 0. \quad (\text{EC.36})$$

To establish (EC.36) it suffices to show that for any given  $\xi > 0$ , there exist  $b(\xi)$  such that

$$\lim_{a \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{b(\xi)/\sqrt{\lambda} \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) > a \log \lambda \right\} \leq \xi. \quad (\text{EC.37})$$

At the end of the proof we will show how (EC.35) also follows from (EC.37).

Now, for  $m \in \mathbb{R}_+$ , let

$$\tau_m^\lambda(Y_d^\lambda) := \inf \left\{ t \geq 0 : Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \leq m \right\} \wedge T. \quad (\text{EC.38})$$

We have that

$$\mathbb{P} \left\{ \sup_{b(\xi)/\sqrt{\lambda} \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) > a \log \lambda \right\} \leq \mathbb{P} \left\{ \sup_{\tau_m^\lambda(Y_d^\lambda) \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) > a \log \lambda \right\} + \mathbb{P} \left\{ \tau_m^\lambda(Y_d^\lambda) > \frac{b(\xi)}{\sqrt{\lambda}} \right\}. \quad (\text{EC.39})$$

We now treat each of the elements on the Right-Hand Side (RHS) of (EC.39), starting with the first element.

We claim that

$$\mathbb{P} \left\{ \sup_{\tau_m^\lambda(Y_d^\lambda) \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) > a \log \lambda \right\} \leq \mathbb{P} \left\{ \sup_{0 \leq s \leq T} Q_{\epsilon, m}^+(\lambda s) > a \log \lambda \right\}. \quad (\text{EC.40})$$

Indeed, if  $\tau_m^\lambda(Y_d^\lambda) > T$ , the set of times  $\{s : \tau_m^\lambda(Y_d^\lambda) \leq s \leq T\}$  is empty so that the inequality holds trivially.

For the other case, letting  $\tilde{Y}_d^\lambda := Q_{\epsilon, Y_d^\lambda}^+(\lambda \tau_m^\lambda(Y_d^\lambda))$  (this is the state at the random time defined in (EC.38)

and we set it to  $\infty$  if  $\tau_m^\lambda(Y_d^\lambda) > T$ ), we have that

$$\begin{aligned} & \mathbb{P} \left\{ \sup_{\tau_m^\lambda(Y_d^\lambda) \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) > a \log \lambda; \tau_m^\lambda(Y_d^\lambda) \leq T \right\} \\ &= \mathbb{P} \left\{ \sup_{0 \leq u \leq T - \tau_m^\lambda(Y_d^\lambda)} Q_{\epsilon, Y_d^\lambda}^+(\lambda(\tau_m^\lambda(Y_d^\lambda) + u)) > a \log \lambda; \tau_m^\lambda(Y_d^\lambda) \leq T \right\} \\ &\leq \mathbb{P} \left\{ \sup_{0 \leq u \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda(\tau_m^\lambda(Y_d^\lambda) + u)) > a \log \lambda; \tau_m^\lambda(Y_d^\lambda) \leq T \right\} \\ &\stackrel{(a)}{=} \mathbb{P} \left\{ \sup_{0 \leq u \leq T} Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda u) > a \log \lambda; \tilde{Y}_d^\lambda \leq m \right\} \\ &\stackrel{(b)}{=} \mathbb{P} \left\{ \sup_{0 \leq u \leq T} Q_{\epsilon, Y_d^\lambda \wedge m}^+(\lambda u) > a \log \lambda \right\}. \end{aligned}$$

where Equality (a) uses the strong Markov property of the  $M/M/1$  queue and Equality (b) follows from the fact that, on the event  $\{\tilde{Y}_d^\lambda \leq m\}$ , we have that  $\tilde{Y}_d^\lambda = Y_d^\lambda \wedge m$ . From the well-known monotonicity of the  $M/M/1$  queue in its initial condition (see e.g. Chapter 9 of Ross [1996]), we then have

$$\mathbb{P} \left\{ \sup_{\tau_m^\lambda(Y_d^\lambda) \leq s \leq T} Q_{\epsilon, Y_d^\lambda \wedge m}^+(\lambda s) > a \log \lambda \right\} \leq \mathbb{P} \left\{ \sup_{0 \leq s \leq T} Q_{\epsilon, m}^+(\lambda s) > a \log \lambda \right\},$$

which establishes (EC.40). To bound the right hand side of (EC.40), we have by Chapter VI.4 of Asmussen [2003], in particular by Problem 4.2 and Example 4.3 there, that

$$\lim_{a \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq s \leq T} Q_{\epsilon, m}^+(\lambda s) > a \log \lambda \right\} = 0. \quad (\text{EC.41})$$

We note that in Asmussen [2003] the cycles of the  $M/M/1$  queue are started in 0, but the same applies if one considers a cycle as the time between consecutive visit to state  $m$ . We conclude from (EC.40) that

$$\lim_{a \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{\tau_m^\lambda(Y_d^\lambda) \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) > a \log \lambda \right\} = 0. \quad (\text{EC.42})$$

This covers the first element on the right hand side of (EC.39), and we turn to the second element. To that end, fix  $\xi > 0$  and note that

$$\mathbb{P} \left\{ \tau_m^\lambda(Y_d^\lambda) > \frac{b(\xi)}{\sqrt{\lambda}} \right\} \leq \mathbb{P} \left\{ \tau_m^\lambda(Y_d^\lambda) > \frac{b(\xi)}{\sqrt{\lambda}}; Y_d^\lambda \leq c\sqrt{\lambda} \right\} + \mathbb{P} \left\{ Y_d^\lambda > c\sqrt{\lambda} \right\}. \quad (\text{EC.43})$$

By the assumptions of the lemma, given  $\xi$  we can choose  $c(\xi)$  so that  $\mathbb{P} \left\{ Y_d^\lambda > c(\xi)\sqrt{\lambda} \right\} \leq \xi$ . By the monotonicity of the  $M/M/1$  queue in the initial condition we have that, on the event  $\left\{ Y_d^\lambda \leq c(\xi)\sqrt{\lambda} \right\}$ ,  $\tau_m^\lambda(Y_d^\lambda) \leq_{st} \tau_m^\lambda(c(\xi)\sqrt{\lambda})$ . Namely,

$$\mathbb{P} \left\{ \tau_m^\lambda(Y_d^\lambda) > \frac{b(\xi)}{\sqrt{\lambda}}, Y_d^\lambda \leq c(\xi)\sqrt{\lambda} \right\} \leq \mathbb{P} \left\{ \tau_m^\lambda(c(\xi)\sqrt{\lambda}) > \frac{b(\xi)}{\sqrt{\lambda}} \right\}.$$

recall that  $Q_\epsilon^+$  is an  $M/M/1$  queue with arrival rate  $\nu + \epsilon < 1$  and service rate 1. By basic results for the  $M/M/1$  queue, see e.g., Proposition 3.3.1 in Meyn [2008] (note that  $T^*(x)$  in that proposition is defined in (3.5), page 63 of that reference),

$$\mathbb{E} \left[ \tau_m^\lambda(c(\xi)\sqrt{\lambda}) \right] \leq \frac{1}{\lambda} \frac{c(\xi)\sqrt{\lambda}}{1 - \nu - \epsilon} = \frac{c(\xi)}{\sqrt{\lambda}(1 - \nu - \epsilon)},$$

where the division by  $\lambda$  in the inequality above is due to the time being scaled by a factor  $\lambda$ . Then by Markov's inequality,

$$\mathbb{P} \left\{ \tau_m^\lambda(c\sqrt{\lambda}) > \frac{b(\xi)}{\sqrt{\lambda}} \right\} \leq \frac{c(\xi)}{b(\xi)(1 - \nu - \epsilon)}.$$

We can now choose  $b(\xi)$  large enough so that  $\mathbb{P} \left\{ \tau_m^\lambda(Y_d^\lambda) > b(\xi)/\sqrt{\lambda}, Y_d^\lambda \leq c(\xi)\sqrt{\lambda} \right\} \leq \xi$  for all  $\lambda$  large enough. Plugging this into (EC.43) we then have that

$$\limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \tau_m^\lambda(Y_d^\lambda) > \frac{b(\xi)}{\sqrt{\lambda}} \right\} \leq 2\xi. \quad (\text{EC.44})$$

Since for each  $\xi > 0$  we can find such  $b(\xi)$ , we can plug (EC.44) together with (EC.42) into (EC.39) to conclude the proof of (EC.37). We turn now to prove (EC.35).

To that end, we note that

$$\begin{aligned} \mathbb{P} \left\{ \sup_{0 \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \geq M\sqrt{\lambda} \right\} &\leq \mathbb{P} \left\{ \sup_{b(\xi)/\sqrt{\lambda} \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \geq M/2\sqrt{\lambda} \right\} \\ &+ \mathbb{P} \left\{ \sup_{0 \leq s \leq b(\xi)/\sqrt{\lambda}} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \geq M/2\sqrt{\lambda} \right\}. \end{aligned}$$

The first element on the RHS was treated in (EC.37). For the second element we can use a crude bound.

Note that for all  $s \leq b(\xi)/\sqrt{\lambda}$ ,  $Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \leq_{st} Y_d^\lambda + \mathcal{N}(\lambda(\nu + \epsilon)b(\xi)/\sqrt{\lambda})$ , where  $\mathcal{N}(\cdot)$  is a unit rate Poisson process (i.e, the state at time  $s$  is smaller than the initial condition plus all the arrivals up to time  $b(\xi)/\sqrt{\lambda} \geq s$ ). Since both  $Y_d^\lambda = O_P(\sqrt{\lambda})$  and  $\mathcal{N}(\lambda(\nu + \epsilon)b(\xi)/\sqrt{\lambda}) = O_P(\sqrt{\lambda})$  it follows that

$$\lim_{M \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq s \leq b(\xi)/\sqrt{\lambda}} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \geq M/2\sqrt{\lambda} \right\} = 0.$$

■

*Proof of Lemma EC.1.3:* Let  $\tilde{\tau}_M^\lambda(Y_d^\lambda) := \inf \left\{ t \geq 0 : D_I^\lambda(t) \geq M\sqrt{\lambda} \right\} \wedge T$ . Recall that  $D_I^\lambda$  is a state-dependent BD process with state space  $d \in \{0, 1, \dots, N_I^\lambda + K_I^\lambda\}$ , having death rate  $\lambda$  and birth rate (in state  $d$ )  $\mu(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(K_I^\lambda - d)^+$  so that for all  $d \leq M\sqrt{\lambda} + 1$

$$\mu(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(K_I^\lambda - d)^+ \geq \lambda(\nu - \epsilon).$$

The process  $D_I^\lambda(t)$  can be constructed on the same sample space with the  $M/M/1$  queue,  $Q_{\epsilon, Y_d^\lambda}^-$  so that

$$D_I^\lambda(t) \geq Q_{\epsilon, Y_d^\lambda}^-(\lambda t), \text{ for all } t \leq \tilde{\tau}_M^\lambda(Y_d^\lambda).$$

The comparison does not necessarily hold after  $\tilde{\tau}_M^\lambda(Y_d^\lambda)$ . The simple coupling argument is omitted and we refer the reader to the proof of Lemma EC.1.11 where a very similar argument is used. Note now that for  $t \leq T$ ,

$$\begin{aligned} \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1 \right\} &= \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1, T \leq \tilde{\tau}_M^\lambda(Y_d^\lambda) \right\} + \mathbb{P} \left\{ D_I^\lambda(t) \geq d_1, T > \tilde{\tau}_M^\lambda(Y_d^\lambda) \right\} \\ &\geq \mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^-(\lambda t) \geq d_1, T \leq \tilde{\tau}_M^\lambda(Y_d^\lambda) \right\} \\ &\geq \mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^-(\lambda t) \geq d_1 \right\} - \mathbb{P} \left\{ \tilde{\tau}_M^\lambda(Y_d^\lambda) < T \right\}. \end{aligned}$$

To conclude the proof we only need to show that for all  $t \leq T$ ,

$$\lim_{M \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \{ \tilde{\tau}_M^\lambda(Y_d^\lambda) < T \} = 0. \quad (\text{EC.45})$$

This, however, follows by noting that

$$\begin{aligned} \lim_{M \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \{ \tilde{\tau}_M^\lambda(Y_d^\lambda) < T \} &\leq \lim_{M \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq s \leq T} D_I^\lambda(s) \geq M\sqrt{\lambda} \right\} \\ &\leq \lim_{M \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq s \leq T} Q_{\epsilon, Y_d^\lambda}^+(\lambda s) \geq M\sqrt{\lambda} \right\} = 0, \end{aligned}$$

where the second inequality follows from Lemma EC.1.1 and the last equality follows from (EC.36).  $\blacksquare$

*Proof of Lemma EC.1.5:* We prove the result only of  $Q_\epsilon^+$ . The proof is identical for  $Q_\epsilon^-$ . To that end, let  $\tau_m^\lambda(Y_d^\lambda)$  be defined as in (EC.38) and note that

$$\mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1 \right\} = \mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1, \tau_m^\lambda(Y_d^\lambda) \leq b \right\} + \mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1, \tau_m^\lambda(Y_d^\lambda) > b \right\}.$$

Since, for any  $b > 0$ ,  $\limsup_{\lambda \rightarrow \infty} \mathbb{P} \{ \tau_m^\lambda(Y_d^\lambda) > b \} = 0$  (see equation (EC.44) and the argument leading to it), we have that

$$\mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1 \right\} = \mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1, \tau_m^\lambda(Y_d^\lambda) \leq b \right\} + o(1),$$

so that it suffices to consider  $\mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1, \tau_m^\lambda(Y_d^\lambda) \leq b \right\}$ . By the strong Markov property of the  $M/M/1$  queue we can write

$$\begin{aligned} \mathbb{P} \left\{ Q_{\epsilon, Y_d^\lambda}^+(\lambda t) \geq d_1, \tau_m^\lambda(Y_d^\lambda) \leq b \right\} &= \mathbb{P} \left\{ Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \geq d_1, \tau_m^\lambda(Y_d^\lambda) \leq b \right\} \\ &= \mathbb{P} \left\{ Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \geq d_1 \right\} + o(1), \end{aligned}$$

where  $\tilde{Y}_d^\lambda$  is the state at the random time  $\tau_m^\lambda(Y_d^\lambda)$ , i.e.,  $\tilde{Y}_d^\lambda = Q_{\epsilon, Y_d^\lambda}^+(\tau_m^\lambda(Y_d^\lambda))$ . The last equality follows again from the fact that  $\mathbb{P} \{ \tau_m^\lambda(Y_d^\lambda) > b \} \rightarrow 0$ . Now,

$$\begin{aligned} \mathbb{P} \left\{ Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \geq d_1 \right\} &= \mathbb{P} \left\{ Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \geq d_1, \tilde{Y}_d^\lambda \leq 2m \right\} \\ &\quad + \mathbb{P} \left\{ Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \geq d_1, \tilde{Y}_d^\lambda > 2m \right\}. \end{aligned}$$

Since  $\mathbb{P}\{\tau_m^\lambda(Y_d^\lambda) > b\} \rightarrow 0$  we also have that  $\mathbb{P}\{\tilde{Y}_d^\lambda > 2m\} \rightarrow 0$ . In turn, it suffices to consider the first element on the RHS above. Since  $\tau_m^\lambda(Y_d^\lambda) \leq T$  by definition and since we consider the event on which the initial condition  $\tilde{Y}_d^\lambda \leq 2m$  (which does not change with  $\lambda$ ) we have that as  $\lambda \rightarrow \infty$ ,

$$Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \Rightarrow Q_\epsilon^+(\infty),$$

and in turn that

$$\mathbb{P}\left\{Q_{\epsilon, \tilde{Y}_d^\lambda}^+(\lambda(t - \tau_m^\lambda(Y_d^\lambda))) \geq d_1, \tilde{Y}_d^\lambda \leq 2m\right\} \rightarrow \mathbb{P}\{Q_\epsilon^+(\infty) \geq d_1\}.$$

This concludes the proof. ■

*Proof of Lemma EC.1.8* We start with the proof of (EC.14). The first part is straightforward. Indeed, since  $\xi_k^\lambda$  is exponential with rate  $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$  we have that

$$\text{Var}(\xi_k^\lambda/\sqrt{\lambda}) = \frac{1}{\lambda} \frac{\lambda^2}{(\mu_I N_I^\lambda + \theta K_I^\lambda)^2},$$

so that, for each  $t > 0$ ,

$$\lim_{\lambda \rightarrow \infty} \sum_{k=1}^{\lceil \lambda t \rceil} \text{Var}(\xi_k^\lambda/\sqrt{\lambda}) = \lim_{\lambda \rightarrow \infty} \frac{\lceil \lambda t \rceil}{\lambda} \frac{\lambda^2}{(\mu_I N_I^\lambda + \theta K_I^\lambda)^2} = \frac{1}{\nu^2} t, \quad (\text{EC.46})$$

where we used the fact that, by our assumptions  $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda \rightarrow \nu$ . Note that, since both the pre-limit functions above are monotone increasing and since the limit is continuous, the pointwise convergence implies uniform convergence. For the second part of (EC.14) we have the following auxiliary result whose proof appears at the end of this section.

LEMMA EC.1.11. As  $\lambda \rightarrow \infty$ ,

$$\sqrt{\lambda} \left( \mathbb{E}[\bar{T}_1^\lambda] - \left(1 - \frac{\mu_I N_I^\lambda + \theta K_I^\lambda}{\lambda}\right)^{-1} \right) \rightarrow 0, \quad (\text{EC.47})$$

$$\sqrt{\lambda} \left( (1 - \rho_d^\lambda) - \frac{\lambda - \mu_I N_I^\lambda - \theta K_I^\lambda}{\lambda} \right) \rightarrow 0, \quad (\text{EC.48})$$

and

$$\text{Var}[\bar{T}_1^\lambda] \rightarrow \frac{1 + \nu}{(1 - \nu)^3}, \quad (\text{EC.49})$$

and consequently

$$\sum_{k=1}^{\lceil \lambda \cdot \rceil} \text{Var} \left( \frac{\bar{T}_k^\lambda}{\sqrt{\lambda}} \right) \Rightarrow \frac{1+\nu}{(1-\nu)^3} e, \quad (\text{EC.50})$$

in  $\mathcal{D}[0, \infty)$ .

Using (EC.46) and (EC.50) we have immediately that

$$\sum_{k=1}^{\lceil \lambda \cdot \rceil} \text{Var} \left( \frac{T_k^\lambda}{\sqrt{\lambda}} \right) \Rightarrow \left( \frac{1+\nu}{(1-\nu)^3} + \frac{1}{\nu^2} \right) e, \quad (\text{EC.51})$$

in  $\mathcal{D}[0, \infty)$ , which proves (EC.14). We can now apply the FCLT for double arrays (see e.g. Theorem 2.3.9 in the internet supplement to Whitt [2002a]) to conclude that, as  $\lambda \rightarrow \infty$ ,

$$\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lceil \lambda \cdot \rceil} (T_k^\lambda - \mathbb{E}[T_k^\lambda]) \Rightarrow \check{\sigma} B \quad \text{as } \lambda \rightarrow \infty,$$

in  $\mathcal{D}[0, \infty)$  where  $\check{\sigma}^2 = \frac{1+\nu}{(1-\nu)^3} + \frac{1}{\nu^2}$ . By Theorem 1 of Iglehart and Whitt [1971] the convergence of the partial sums implies convergence of the corresponding renewal process so that

$$\frac{1}{\sqrt{\lambda}} (R^\lambda(\lambda \cdot) - \lambda / \mathbb{E}[T_k^\lambda] \cdot) \Rightarrow \sqrt{\nu(1-\nu)} \check{\sigma} B \quad \text{as } \lambda \rightarrow \infty,$$

in  $\mathcal{D}[0, \infty)$ . From here it also follows directly that, uniformly on compact subsets,

$$R^\lambda(\lambda \cdot) / \lambda - \cdot / \mathbb{E}[T_k^\lambda] \Rightarrow 0 \quad \text{as } \lambda \rightarrow \infty,$$

in  $\mathcal{D}[0, \infty)$ . Since  $\mathbb{E}[T_k^\lambda] \rightarrow 1/\nu + 1/(1-\nu) = 1/(\nu(1-\nu))$ , we have that  $R(\lambda \cdot) / \lambda \Rightarrow \nu(1-\nu)e$  in  $\mathcal{D}[0, \infty)$ .

The convergence in (EC.14) now follows from the marginal convergence of the components because all limits are non-random (see Theorem 11.4.5 in Whitt [2002a]). We now use (EC.46), (EC.50) and (EC.48) and apply the FCLT for double arrays to each of the two (independent) processes

$$\frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} \left( \rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} \right), \quad \text{and} \quad \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} (1 - \rho_d^\lambda) (\bar{T}_k^\lambda - \mathbb{E}[\bar{T}_k^\lambda])$$

to obtain the joint convergence of

$$\left( \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} \left( \rho_d^\lambda \xi_k^\lambda - \frac{\lambda \rho_d^\lambda}{\mu_I N_I^\lambda + \theta K_I^\lambda} \right), \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{\lambda \cdot} (1 - \rho_d^\lambda) (\bar{T}_k^\lambda - \mathbb{E}[\bar{T}_k^\lambda]), \frac{R(\lambda \cdot)}{\lambda} \right) \Rightarrow \left( B^1, \sqrt{\frac{1+\nu}{1-\nu}} B^2, \nu(1-\nu)e \right),$$

in  $\mathcal{D}[0, \infty)$  where  $B^1$  and  $B^2$  are standard independent Brownian motions. Equation (EC.15) now follows from the random time change theorem; see e.g. §17 of Billingsley [1968]. ■

*Proof of Lemma EC.1.9* The first part follows from (EC.47) and (EC.48) after basic algebraic manipulations. Equation (EC.16) then follows since  $R^\lambda(\lambda)/\lambda \Rightarrow \nu(1-\nu)e$  as proved within the proof of Lemma EC.1.8. ■

*Proof of Lemma EC.1.10:* Consider first the (sequence of) processes  $M^\lambda(s)$ ,  $0 \leq s \leq t$  defined by

$$M^\lambda(s) := \int_0^s \Psi^\lambda(u-) dY^\lambda(u), \quad 0 \leq s \leq t,$$

and let  $M(s) = \int_0^s \Psi(u-) dY(u)$ . Then, from Theorem 0 in ? it follows, under certain conditions to be discussed shortly, that  $M^\lambda \Rightarrow M$  in  $\mathcal{D}[0, t]$ . The conditions in that theorem refer to a certain UT property of the process  $Y^\lambda$  (see the discussion in ? immediately after the statement of Theorem 0 there). However, this condition holds trivially when (as in our case)  $Y^\lambda$  is, for each  $\lambda$ , an increasing process and such that  $Y^\lambda \Rightarrow Y$ , which implies, in particular, that the sequence  $Y^\lambda$  is stochastically bounded. In addition, the martingale condition in Theorem 0 in ? holds trivially, as one can always use the self filtration of the process  $(\Psi^\lambda, Y^\lambda)$ , for each  $\lambda$ . The process  $\Psi^\lambda$  is then adapted to this filtration and, being an increasing process,  $Y^\lambda$  is a semi-martingale with respect to this filtration. Thus, it indeed holds that  $M^\lambda \rightarrow M$ .

We next show that the convergence holds if one replaces the integrand  $\Psi^\lambda(s-)$  with  $f(\Psi^\lambda(s-))$ , where  $f$  satisfies (EC.34). Note that, for all  $\lambda$ ,  $t > 0$  and Borel measurable set  $\mathcal{B}$ , we have that

$$\mathcal{M}^\lambda(\mathcal{B}) := \int_0^t \mathbb{1}\{\Psi^\lambda(s-) \in \mathcal{B}\} dY^\lambda(s)$$

is a (random measure) on  $\mathbb{R}$ ; see e.g. ?. By the first part of this proof it holds that

$$\int_0^t f(\Psi^\lambda(s-)) dY^\lambda(s) \Rightarrow \int_0^t f(\Psi(s-)) dY(s),$$

for every continuous function  $f : \mathbb{R} \rightarrow \mathbb{R}_+$ . Thus, the sequence of random measures  $\mathcal{M}^\lambda(\cdot)$  converges weakly to  $\mathcal{M}(\cdot)$ , where

$$\mathcal{M}(\mathcal{B}) = \int_0^t \mathbb{1}\{X(s-) \in \mathcal{B}\} dY(s).$$

The condition (EC.34) implies that  $\mathcal{M}(\cdot)$  assigns zero measure to the family of discontinuity points of  $f$  so that, from the generalized version of continuous mapping theorem (see, e.g., Theorem 3.4.3 in Whitt [2002b]), we deduce the convergence

$$\int_0^t f(\Psi^\lambda(s-)) dY^\lambda(s) \Rightarrow \int_0^t f(\Psi(s-)) dY(s),$$

for all functions  $f$  that satisfy condition (EC.34). ■

*Proof of Lemma EC.1.11* Recall that  $\bar{T}^\lambda$  has the distribution of the length of a busy period of the process  $D_I^{s,\lambda}(\cdot)$ , i.e.,  $\bar{T}^\lambda = T^\lambda - \xi^\lambda$  for  $T^\lambda$  that has the distribution of the busy cycle and  $\xi^\lambda$  that has the distribution of the idle period. To obtain bounds for the busy period we will couple the process  $D_I^{s,\lambda}(\cdot)$  with two queues.

Specifically, let  $\tilde{T}^\lambda$  have the distribution of a busy period in an  $M/M/1$  queue with service rate 1 and arrival rate  $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$ . Let  $\tilde{\rho}^\lambda$  be the utilization in this  $M/M/1$  queue. Let  $\check{T}^\lambda$  have the distribution of a busy period in an  $M/M/1$  queue with service rate 1 and arrival rate  $(\mu_I N_I^\lambda/\lambda + \theta K_I^\lambda - 2K \log(\lambda))/\lambda$  (note that for all  $\lambda$  large enough this is a positive number and it is also strictly smaller than 1). Let  $\check{\rho}^\lambda$  be the utilization in this  $M/M/1$  queue and let  $\tilde{M}^\lambda$  be a random variable with the distribution of the busy-period maximum in such an  $M/M/1$  queue.

We will next use a coupling argument to show that

$$\mathbb{E} \left[ \tilde{T}^\lambda \mathbb{1} \left\{ \tilde{M}^\lambda \leq K \log(\lambda) \right\} \right] \leq \mathbb{E} [\bar{T}^\lambda] \leq \mathbb{E} [\check{T}^\lambda] \quad (\text{EC.52})$$

and similarly that

$$\mathbb{E} \left[ (\tilde{T}^\lambda)^2 \mathbb{1} \left\{ \tilde{M}^\lambda \leq K \log(\lambda) \right\} \right] \leq \mathbb{E} [(\bar{T}^\lambda)^2] \leq \mathbb{E} [(\check{T}^\lambda)^2]. \quad (\text{EC.53})$$

Let  $\tilde{Q}^\lambda(t)$  be the queue length of the  $M/M/1$  queue with service rate 1 and arrival rate  $(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$  and let  $\check{Q}^\lambda(t)$  be defined similarly with the arrival rate  $(\mu_I N_I^\lambda/\lambda + \theta K_I^\lambda - 2K \log(\lambda))/\lambda$ . Then, we start at time 0 all the queues with the server busy and zero customers in queue. Namely, set  $\tilde{Q}^\lambda(0) = \check{Q}^\lambda(0) = D_I^{s,\lambda}(0) = 1$ . Then, we will couple them until the first of them hits 0. We generate “events” using a Poisson process with rate  $b^\lambda := 1 + (\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda$ . We determine whether an event is an arrival or a service completion using the same sequence of uniform  $[0, 1]$  random variables for all the three queues (for each event we use a different uniform random variable and these uniforms are IID). The  $k^{\text{th}}$  event trigger an arrival in  $D_I^{s,\lambda}$  when in state  $d$  if  $U_k \in [0, ((\mu(N_I^\lambda - (d - K_I^\lambda)^+) + \theta(d \wedge K_I^\lambda))/\lambda)/b^\lambda)$ . It is a service completion if  $U_k \in [1 - 1/b^\lambda, 1]$  and  $D_I^{s,\lambda}$  stays in its state  $d$  otherwise. Similarly, an event triggers an arrival in  $\check{Q}_1^\lambda$  if  $U_k \in [0, ((\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda)/b^\lambda)$ . It trigger a service completion if  $U_k \in [1 - 1/b^\lambda, 1]$  and it stays put otherwise. Finally, an event triggers an arrival in  $\tilde{Q}_1^\lambda$  if  $U_k \in [0, ((\mu_I N_I^\lambda + \theta K_I^\lambda - 2K \log(\lambda))/\lambda)/b^\lambda)$ . It triggers a service completion if  $U_k \in [1 - 1/b^\lambda, 1]$  and it stays put otherwise.

Let  $\tilde{\tau}^\lambda = \inf \left\{ t \geq 0 : \tilde{Q}^\lambda(t) = 0 \text{ or } \tilde{Q}^\lambda(t) \geq K \log(\lambda) \right\}$ . Then with the above sample-path construction is necessarily holds that:

- (i) for all  $t \geq 0$ ,  $D_I^{s,\lambda}(t) \leq \tilde{Q}^\lambda(t)$  and
- (ii) for all  $t \leq \tilde{\tau}^\lambda$ ,  $\tilde{Q}^\lambda(t) \leq D_I^{s,\lambda}(t) \leq \tilde{Q}^\lambda(t)$ .

From here it also follows that, under this construction

$$\tilde{T}^\lambda \mathbb{1} \left\{ \tilde{M}^\lambda \leq K \log \lambda \right\} \leq \bar{T}^\lambda \leq \check{T}^\lambda$$

almost surely. Taking squares and applying expectations on both sides (ny Proposition 8.10 in Asmussen [2003] these expectations are finite), we have (EC.52) and (EC.53).

We next use (EC.52) and (EC.53) to complete the proof of the lemma. By known results for the busy period of the  $M/M/1$  queue (see e.g. Proposition 8.10 in Asmussen [2003]) we have

$$\mathbb{E} [\check{T}^\lambda] = \frac{1}{(1 - (\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda)} \text{ and } \mathbb{E} [\tilde{T}^\lambda] = \frac{1}{(1 - (\mu_I N_I^\lambda + \theta K_I^\lambda - 2K \log(\lambda))/\lambda)},$$

so that  $\sqrt{\lambda} \left( \mathbb{E} [\check{T}^\lambda] - \mathbb{E} [\tilde{T}^\lambda] \right) \rightarrow 0$ , and, to establish the first part of the lemma, it suffices to show that

$$\sqrt{\lambda} \mathbb{E} \left[ \tilde{T}^\lambda \mathbb{1} \left\{ \tilde{M}^\lambda > K \log(\lambda) \right\} \right] \rightarrow 0, \text{ as } \lambda \rightarrow \infty.$$

By Holder's inequality

$$\mathbb{E} \left[ \tilde{T}^\lambda \mathbb{1} \left\{ \tilde{M}^\lambda > K \log(\lambda) \right\} \right] \leq \sqrt{\mathbb{E} \left[ (\tilde{T}^\lambda)^2 \right]} \sqrt{\mathbb{P} \left\{ \tilde{M}^\lambda > K \log(\lambda) \right\}}.$$

It is known (see again Asmussen [2003]) that  $\mathbb{E} \left[ (\tilde{T}^\lambda)^2 \right] = (1 + \tilde{\rho}^\lambda)/(1 - \tilde{\rho}^\lambda)^3$  so that, since  $\tilde{\rho}^\lambda \rightarrow \nu < 1$  by Assumption 1, we have  $\limsup_{\lambda \rightarrow \infty} \mathbb{E} \left[ (\tilde{T}^\lambda)^2 \right] < \infty$ . Note that  $\tilde{M}^\lambda \leq_{st} M^\lambda := \sup_{n \geq 0} S_n^\lambda$  where  $S_n^\lambda$  is random walk (starting at 1) that increases by one with probability

$$\phi^\lambda = \frac{(\mu_I N_I^\lambda + \theta K_I^\lambda - 2K \log(\lambda))/\lambda}{\lambda + (\mu_I N_I^\lambda + \theta K_I^\lambda - 2K \log(\lambda))/\lambda}$$

and decreases by one with probability  $1 - \phi^\lambda$ . This is a Gambler's-ruin type problem and from basic arguments is follows that

$$\mathbb{P} \left\{ M^\lambda > K \log(\lambda) + 1 \right\} = \left( \frac{\phi^\lambda}{1 - \phi^\lambda} \right)^{2K \log \lambda};$$

see e.g. Chapter 7.3 of Ross [1996]. By Assumption 1,  $\phi^\lambda \rightarrow 1/(1+\nu) < 1$ . In turn, there exists a constant  $c < 1$  such that  $\mathbb{P}\{M^\lambda > K \log(\lambda) + 1\} \leq c^{2K \log \lambda}$  for all  $\lambda$  large enough. Recalling that  $\tilde{M}^\lambda \leq_{st} M^\lambda$  and choosing  $K$  large enough, we conclude that

$$\sqrt{\lambda} \sqrt{\mathbb{P}\{\tilde{M}^\lambda > K \log(\lambda)\}} \rightarrow 0.$$

Similar arguments are applied to the variance terms. The only difference is that when applying Holder's inequality we will use the fact that, since  $\tilde{\rho}^\lambda \rightarrow \nu$ , the sequence  $\{\mathbb{E}[(\tilde{T}^\lambda)^4], \lambda \geq 0\}$  convergence so that, in particular,  $\limsup_{\lambda \rightarrow \infty} \mathbb{E}[(\tilde{T}^\lambda)^4] < \infty$ . The fact that the moments converge follows from the convergence of the characteristic function together with its derivatives at 0 of all orders as  $\lambda \rightarrow \infty$  and  $\tilde{\rho}^\lambda \rightarrow \nu$ . The convergence of the characteristic function is easily verifiable using the explicit expressions; see Proposition 8.10 of Asmussen [2003].

To conclude the proof we show that (EC.48) follows from (EC.47). Indeed, by basic regenerative process arguments (applied to the regenerative process  $D_I^{s,\lambda}$ ) we have that

$$\rho_d^\lambda = \frac{\mathbb{E}[\bar{T}^\lambda]}{\mathbb{E}[\bar{T}^\lambda] + \mathbb{E}[\xi^\lambda]}.$$

Recalling that  $\mathbb{E}[\xi^\lambda] = \frac{1}{(\mu_I N_I^\lambda + \theta K_I^\lambda)/\lambda}$  and using (EC.47) the result by basic algebraic manipulations. ■

## EC.2. The Multi-Class Setting

Thus far we considered a setting in which the outside provider uses a dedicated pool of servers for the overflow input. With this configuration, analyzing a system with multiple in-house call centers is identical to analyzing multiple independent subsystems consisting of an in-house call center and its dedicated station at the outside provider's facility.

In this section we consider the case in which the outside provider serves multiple input streams in a common facility. Station  $O$  can then be thought of as a call center with multiple customer classes and possibly multiple agent pools. An example having a single agent pool is depicted in Figure 1(b). More specifically, each overflow stream can be thought of as a customer class. There might be different agent groups that differ by their skills set, and hence have different levels of flexibility. In such a multi-class multi-pool configuration, the outsourcer needs to determine the policy for real-time prioritization of customers.

For the system in Figure 1(b), the controller has to determine: (i) what to do with an arriving call if there is an available agent upon its arrival (admit it to service or reserve the capacity for other customer classes), and (ii) which customer to admit to service when an agent becomes available and there are customers waiting in multiple queues. We refer the reader to Gurvich and Whitt [2010] for a more detailed description of the underlying queueing network.

Fortunately, our results for the base model extend to this more complex setting. Below we highlight what are the corresponding mathematical statements and what are the assumptions that have to be imposed on the prioritization rule in station  $O$ . To formally state the results, we let  $\mathcal{M} := \{1, \dots, M\}$  be the set of in-house centers, i.e, the sources of overflows for station  $O$ . The process  $A_m$  is then the Poisson process of exogenous arrivals to (in-house) center  $m \in \mathcal{M}$ , with rate  $\lambda_m$ . We denote the aggregate arrival rate by  $\lambda = \sum_m \lambda_m$ . As before, we add the superscript  $\lambda$  to all quantities that change along the sequence of systems. We will assume that  $\lambda_m/\lambda = a_m > 0$ , i.e., the fractions  $a_m$  do not scale with  $\lambda$ .

The service rate at station  $m$  is given by  $\mu_m$  and we let  $\theta_m$  be the patience rate of the customers arriving to station  $m$ . Let  $(N_m^\lambda, K_m^\lambda)$  be the capacity and threshold pair for station  $m \in \mathcal{M}$ . The resource pooling condition (item (1) in Assumption 1) is assumed to hold for each of the in-house centers, i.e,

$$\lim_{\lambda \rightarrow \infty} \frac{\mu_m N_m^\lambda + \theta K_m^\lambda}{\lambda_m} = \nu_m < 1, \quad m \in \mathcal{M}. \quad (\text{EC.54})$$

Let  $A_{m,O}^\lambda$  be the overflow process from station  $m$ , and define the scaled process

$$\widehat{A}_{m,O}^\lambda(t) = \frac{A_{m,O}^\lambda(t) - (\lambda_m - \mu_m N_m^\lambda - \theta_m K_m^\lambda)t}{\sqrt{\lambda_m}}.$$

For state descriptors, we use  $X_{m,I}(t)$  to denote the number of customers present in station  $m \in \mathcal{M}$  at time  $t$ . In station  $O$  there is a queue for each overflow stream (for each customer class) and we let  $X_{m,O}(t)$  and  $Q_{m,O}(t)$  be, respectively, the total head count of “class- $m$ ” customers in station  $O$  at time  $t$ , and the number in queue there. As in Assumption 1, we will assume that station  $O$  uses a square-root safety staffing rule.

For in-house station  $m \in \mathcal{M}$  we define the scaled and centered process  $\widehat{X}_{m,I}^\lambda := (X_{m,I}^\lambda - (N_m^\lambda + K_m^\lambda))/\sqrt{\lambda}$ . For station  $O$  and class  $k$  we denote by  $\widehat{Q}_{m,O}^\lambda$  the scaled queue length process and by  $\widehat{X}_{m,O}^\lambda$  the scaled and centered head-count process. The square-root staffing rule, as well as the centering of the

head-count processes  $X_{m,O}^\lambda$ , need to be defined carefully in the multi-class multi-pool setting. Such details are not central for our results below and we refer the reader to §2.1 of Gurvich and Whitt [2009]. We will be assuming, analogously to (12), that the initial conditions converge, i.e, that

$$(\widehat{Q}_{m,O}^\lambda(0), \widehat{X}_{m,O}^\lambda(0), \widehat{X}_m^\lambda(0); m \in \mathcal{M}) \Rightarrow (\widehat{Q}_{m,O}(0), \widehat{X}_{m,O}(0), \widehat{X}_m(0); m \in \mathcal{M}) \text{ as } \lambda \rightarrow \infty. \quad (\text{EC.55})$$

Since the in-house call centers are independent of each other, Theorem 4.1 immediately generalizes to the multiclass setting and the proof requires no modification, i.e,

$$(\widehat{A}_1^\lambda, \dots, \widehat{A}_M^\lambda) \Rightarrow (\sigma_1 B_1, \dots, \sigma_M B_M) \text{ in } \mathcal{D}^M \text{ as } \lambda \rightarrow \infty, \quad (\text{EC.56})$$

where  $B_1, \dots, B_M$  are  $M$  independent standard Brownian motions, and  $\sigma_m^2 = 1 + \nu_m$ ,  $m \in \mathcal{M}$ .

The asymptotic independence in Theorem 4.3 continues to hold in the multi-class setting for each pair of in-house call center and outsourcer. The proof of that theorem reveals that the only requirement is that the processes in station  $O$  are  $\mathcal{C}$ -tight. (See also the intuition in Remark 4.3.) In the multi-class settings, such tightness holds, in particular, if the sequence  $\{(\widehat{X}_{1,O}^\lambda, \dots, \widehat{X}_{M,O}^\lambda)\}$  converges to a continuous limit. This is satisfied, for example, by the QIR family of controls studied in Gurvich and Whitt [2009]. The same is true for other policies that try to maintain certain proportions between the queues; see Atar [2005] and Atar et al. [2004]. However, not all routing rules satisfy this property. For example, the bang-bang type rule in Harrison and Zeevi [2004] does not produce continuous heavy-traffic limits.

If the  $\mathcal{C}$ -tightness holds, then the corresponding result is a direct extension of Theorem 4.3 – requiring again no modification of the proof, as the analysis can be applied to each  $m \in \mathcal{M}$  separately. The corresponding asymptotic independence statement is then

$$\mathbb{P} \left\{ \widehat{X}_{m,O}^\lambda(t) > q, D_{m,I}^\lambda(t) = d \right\} = \mathbb{P} \left\{ \widehat{X}_{m,O}^\lambda(t) > q \right\} \mathbb{P} \left\{ D_{m,I}^\lambda(t) = d \right\} + o(1), \quad t > 0, \quad q \in \mathbb{R}, d \in \mathbb{Z}_+,$$

where  $D_{m,I}^\lambda := N_m^\lambda + K_m^\lambda - X_{m,I}^\lambda$ . Corollaries 4.4 and 4.5 also extend to the multi-class case without any changes to their proofs.

### EC.2.1. Optimization in the Multi-class Setting

As we pointed out in the introduction, one may wish to study the value of real-time information about the state of the in-house call centers for optimization problems of the outside provider. We will now show that such information carries at most marginal benefit. For ease of presentation we focus on a relatively simple setting, but the results hold for more general models. Specifically, we focus on the case in which the outside provider has a single pool of servers serving all customer classes (overflow streams). A two-class example is depicted in Figure 1(b). The queueing network in station  $O$  is then the so-called  $V$  model; see Gurvich et al. [2008] and Atar et al. [2004].

In serving multiple input streams in one facility, the outside provider's problem is now a staffing-and-routing optimization problem. Specifically, the outside provider's problem (5) may become

$$\begin{aligned} \min \quad & C_s(N_O) + \sum_{m \in \mathcal{M}} \mathbb{E} \left[ \int_0^\infty e^{-\gamma s} C_m(\widehat{Q}_{m,O}^{\pi,\lambda}(s)) ds \right] \\ \text{s.t.} \quad & N_O \in \mathbb{Z}_+, \pi \in \Pi^\lambda, \end{aligned} \quad (\text{EC.57})$$

where  $\Pi^\lambda$  is the family of admissible routing rules, and the superscript  $\pi$  makes explicit the dependency of the queue length process on the control being employed. Normally, as in (5), the outside provider would have a constraint on the waiting time rather than holding costs, but one may think about that problem as a dualization of such constraints. In fact, it suffices to fix  $N_O$  and consider the control problem

$$\begin{aligned} \min \quad & \sum_{m \in \mathcal{M}} \mathbb{E} \left[ \int_0^\infty e^{-\gamma s} C_m(\widehat{Q}_{m,O}^{\pi,\lambda}(s)) ds \right] \\ \text{s.t.} \quad & \pi \in \Pi^\lambda. \end{aligned} \quad (\text{EC.58})$$

Assuming that (EC.57) has an optimal staffing-and-routing solution  $(N_O^{*,\lambda}, \pi^{*,\lambda})$  and that, for each  $N_O$ , (EC.58) has an optimal control  $\pi^{*,\lambda}(N_O)$ , one can solve (EC.57) by using  $\pi^{*,\lambda}(N_O)$  and optimizing only over  $N_O$ .

We further assume that the functions  $C_m(\cdot)$  are twice continuously differentiable with strictly positive second derivatives, and that these functions have at most polynomial growth, i.e, that there exist constant

$M_m$  and  $l_m$  such that  $C_m(x) \leq M_m(1 + |x|^{l_m})$ . These assumptions will allow us to interpret directly some results in Atar et al. [2004].

To show that the real-time state information does not carry significant value we first formalize the notion of information with respect to the optimization problem (EC.58). To that end, we define  $S_{m,O}^\lambda(t)$  to be the number of class- $m$  customers that departed from station  $O$  by time  $t$  after completing service, and  $L_{m,O}^\lambda(t)$  be the number of class- $m$  customers that abandoned station  $O$  by time  $t$ . We assume that, regardless of whether information is shared or not, the outside provider knows the state of his local queues, as reflected in the processes  $X_{m,O}^\lambda$  and  $Q_{m,O}^\lambda$ , as well as the local history up to the time of the decision. Namely, when making a decision at time  $t$ , the outside provider knows the evolution of the processes  $(X_{m,O}^\lambda, Q_{m,O}^\lambda)$ ,  $m \in \mathcal{M}$  up to time  $t$ , as well as that of the processes  $A_{m,O}^\lambda, S_{m,O}^\lambda$  and  $L_{m,O}^\lambda$ .

The information-sharing and no-information-sharing settings differ with regards to whether or not the outside provider has access to the real-time state information of each of the in-house queues as captured by the process  $(X_{1,I}^\lambda, \dots, X_{M,I}^\lambda)$ . Without information sharing the outside provider has to make its prioritization decisions based only on his local information. Define the filtrations

$$\mathcal{F}_t^\lambda = \sigma \left\{ A_{m,O}^\lambda(s), S_{m,O}^\lambda(s), L_{m,O}^\lambda(s), X_{m,O}^\lambda(s), Q_{m,O}^\lambda(s), Z_{m,O}^\lambda(s), X_{m,I}^\lambda(s); m \in \mathcal{M}, s \leq t \right\}.$$

and

$$\check{\mathcal{F}}_t^\lambda = \sigma \left\{ A_{m,O}^\lambda(s), S_{m,O}^\lambda(s), L_{m,O}^\lambda(s), X_{m,O}^\lambda(s), Q_{m,O}^\lambda(s), Z_{m,O}^\lambda(s); m \in \mathcal{M}, s \leq t \right\}.$$

Note that  $\mathcal{F}_t^\lambda$  contains also the information about  $X_{m,I}^\lambda$  which is not contained in  $\check{\mathcal{F}}_t^\lambda$ . We then let  $\Pi^\lambda$  be the family of policies that are ‘‘adapted’’ to  $(\mathcal{F}_t^\lambda)_{t \geq 0}$ , and  $\check{\Pi}^\lambda$  be the subset of  $\Pi^\lambda$  that is adapted to  $(\check{\mathcal{F}}_t^\lambda)_{t \geq 0}$ . We refer the reader to Atar et al. [2004] for a more formal discussion. The control problem for the outside provider can then be defined by

$$\begin{aligned} \min \quad & \sum_{m \in \mathcal{M}} \mathbb{E} \left[ \int_0^\infty e^{-\gamma s} C_m(\widehat{Q}_{m,O}^{\pi,\lambda}(s)) ds \right] \\ \text{s.t.} \quad & \check{\pi} \in \check{\Pi}^\lambda, \end{aligned} \tag{EC.59}$$

and this should be contrasted with (EC.58), where the larger family of policies  $\Pi^\lambda$  is admissible. Let  $C^\lambda(\pi, x)$  be the cost under policy  $\pi$  when the initial state is  $x$ . By state we mean the state throughout the system (in-house stations and outside provider). Let  $V^\lambda(x) = \inf_{\pi \in \Pi^\lambda} C^\lambda(\pi, x)$  and  $\check{V}^\lambda(x) = \inf_{\check{\pi} \in \check{\Pi}^\lambda} C^\lambda(\check{\pi}, x)$  be the optimal costs to (EC.58) and (EC.59), respectively.

One expects the optimal value of (EC.59) to be strictly greater than that of (EC.58), i.e., that  $V^\lambda(x) < \check{V}^\lambda(x)$ . The main result of this section shows that the reverse inequality also holds asymptotically. Namely, that in a context with resource pooling, the additional information about the in-house call centers does not carry significant benefits.

**THEOREM EC.2.1.** Assume that (EC.54) holds. Let  $x^\lambda = (\hat{Q}_{m,O}^\lambda(0), \hat{X}_{m,O}^\lambda(0), \hat{X}_{m,I}^\lambda(0) : m \in \mathcal{M})$  and suppose that the sequence  $\{x^\lambda\}$  satisfies (EC.55) where the limit  $x = \lim_{\lambda \rightarrow \infty} x^\lambda$  is nonrandom. Then,

$$\check{V}^\lambda(x^\lambda) \leq V^\lambda(x^\lambda) + o(1).$$

Our approximation of the overflow process in (EC.56) is key in establishing the above result. In fact, given that approximation, the proof of Theorem EC.2.1 (see below) is a direct corollary of the analysis and the results in Atar et al. [2004]. Theorem EC.2.1 adds to our discussion of outsourcing-scheme comparisons in §3.1, where we argued how our asymptotic-independence results allow to simplify the outsourcer problem to one that does not require knowledge of joint distributions. Here we show that even real-time information carries, at most, negligible benefit for the outside provider.

*Outline of the proof:* We provide only an outline of the argument, assuming familiarity of the reader with the terminology and notation in Atar et al. [2004]. First note that, due to the convergence of the overflow processes (which would be the input processes in the model of Atar et al. [2004]), one can follow the steps in §2.5 there to “guess” the Brownian control problem. This Brownian control problem should be interpreted in our setting as focusing on the V model in station  $O$ . Hence, it uses only local information about station  $O$  as we do in defining the family  $\check{\Pi}^\lambda$  of policies. Due to our restrictions on the cost functions  $C_m(\cdot)$ ,  $m \in \mathcal{M}$ , Assumptions 1-3 of Atar et al. [2004] hold. In turn, the requirements of Theorem 1 there are satisfied, and the Brownian control problem has an optimal Markov control policy. In fact, due to our restrictions on

the cost functions, it also follows from Proposition 3 there that the Markov control function  $h$  is locally Hölder continuous (see the statement of Theorem 2 there). Atar et al. [2004] then uses this Markov control in proposing a sequence of controls for the original sequence of queueing system (see §2.6 there).

By Theorem 2 in Atar et al. [2004] it follows that the policies constructed there are asymptotically optimal for (EC.59) that we defined. It remains to argue that these are actually asymptotically optimal for (EC.58). In our setting this follows from our convergence results for the overflow process (see Theorem 4.1) and repeating precisely the steps in the proof of asymptotic optimality in §4 of Atar et al. [2004]. In following that proof, one should note that, given the convergence of the overflow processes to a drifted Brownian motion that is independent of the in-house processes, the arguments remain unchanged when the larger information set is being employed. ■

## e-companion references

- Asmussen, S. 2003. *Applied Probability and Queues*. 2nd ed. Springer, New York.
- Atar, R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Annals of Applied Probability* **15**(4) 2606–2650.
- Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multi-class queue with many exponential servers: asymptotic optimality in heavy traffic. *Annals of Applied Probability* **14**(3) 1084–1134.
- Billingsley, P. 1968. *Convergence of Probability Measures*. J. Wiley & Sons, New York.
- Glynn, P.W., W. Whitt. 1993. Limit theorems for cumulative processes. *Stochastic Processes and their Applications* **47** 299–314.
- Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service-level differentiation in call centers with fully flexible servers. *Management Science* **54**(2) 279–294.
- Gurvich, I., W. Whitt. 2009. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research* **34**(2) 363–396.
- Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. *Operations Research* **58**(2) 316–328.
- Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime. *Operations Research* **52**(2) 243–257.
- Iglehart, D.L., W. Whitt. 1971. The equivalence of functional central limit theorems for counting processes and associated partial sums. *The Annals of Mathematical Statistics* **42**(4) 1372–1378.
- Karatzas, I., S. Shreve. 1991. *Brownian Motion and Stochastic Calculus*. 2nd ed. Springer-Verlag, New York.
- Meyn, S.P. 2008. *Control techniques for complex networks*. Cambridge Univ Pr.
- Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** 193–267.
- Perry, O., W. Whitt. 2010a. A fluid limit for an overloaded  $x$  model via an averaging principle. Working paper, Columbia University, New York, NY.
- Perry, O., W. Whitt. 2010b. Gaussian approximations for an overloaded  $x$  model via an averaging principle. Working paper, Columbia University, New York, NY.
- Ross, S.M. 1996. *Stochastic processes*. Wiley New York.

- Talreja, R., W. Whitt. 2009. Heavy-traffic limits for waiting times in many-server queues with abandonment. *Annals of Applied Probability* **19**(6) 2137–2175.
- Whitt, W. 1991. The pointwise stationary approximation for  $M_t/M_t/s$  queues is asymptotically correct as the rates increase. *Management Science* **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, New York.
- Whitt, W. 2005. Heavy-traffic limits for the  $G/H_2^*/n/m$  queue. *Mathematics of Operations Research* **30**(1) 1–27.