MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

A Logarithmic Safety Staffing Rule for Contact Centers with Call Blending

Guodong Pang
The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University
gup3@psu.edu

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

We consider large contact centers that handle two types of jobs – inbound and outbound – simultaneously, a process commonly referred to as *call blending*. Inbound work arrives to the system according to an exogenous arrival process, whereas outbound work is generated by the contact center. We assume that there is an infinite supply of outbound work to process, and that inbound calls are prioritized over the outbound calls. We propose a logarithmic safety staffing rule, combined with a threshold control policy, ensuring that agents' utilization is very close to one at all times, but that there are practically always idle agents present. Specifically, we prove that it is possible to have almost all inbound calls answered immediately upon their arrival, in addition to satisfying a target long-run throughput rate of outbound calls, with at most a negligible proportion of those calls dropped. Simulation experiments demonstrate the effectiveness and accuracy of our analysis.

Key words: contact centers; call blending; safety staffing; threshold controls; many-server queues

1. Introduction

As the number, types and complexity of the services that are offered over the phone and online keep increasing, so does the significance of contact centers, namely, call centers that handle more than one type of jobs, such as inbound and outbound calls or telephony and online chatting. In particular, most businesses place outbound calls in addition to their inbound-calls services; see Ben-Chanoch (2004) and Reynolds (2010).

In this paper we consider contact centers in which a large pool of agents handles inbound and outbound calls simultaneously, a process commonly referred to as *call blending*. We show that, due to the relatively low variability associated with outbound calls, blending makes it possible to provide pre-specified levels of service quality to inbound calls and throughput rate of outbound calls, while keeping the service utilization close to 1 at all times. Specifically, our main results demonstrate the operational advantages of blended systems over split systems where, in a split

system, inbound and outbound calls are handled by dedicated service pools, with each pool serving only one type of calls.

Overview of the Problem and Results. When inbound calls are prioritized over outbound calls in a blended system, it is possible to ensure that waiting times of inbound calls are relatively short, even if all the agents are working all the time, provided that the total number of agents in the service pool is sufficiently large relative to the arrival rate of inbound work. However, if all agents are constantly working, most inbound customers will be delayed in queue before entering service, and some will abandon. This can be avoided if one can ensure that there is idleness in the system at almost all times.

Having constant idleness is often needed from the outbound-calls perspective, so as to minimize the probability of calls from being dropped due to unavailability of agents. Specifically, many modern outbound contact centers, automatic dialers initiate outbound calls even when all agents are busy, using *predictive dialing* software, with the purpose of minimizing agents' idling times. In such cases, if a customer replies to an outbound call and there is no available agent to take care of this customer, the call is dropped immediately. Those dropped calls are costly to the contact center and cause nuisance to "abandoned" customers. Moreover, government regulations limit the allowed percentage of such calls; see Samuelson (1999). The risk for dropped outbound calls clearly increases in blended pools since, even if an outbound call is initiated when agents are available, by the time the called party replies, there may no longer remain available agents. To minimize the number of dropped calls in this automated environment, one has to guarantee that there is sufficient idleness in the system. We elaborate further in §2.2 below.

In summary, there are clear benefits for having idleness in the system at almost all times. On the other hand, as staffing constitutes the bulk operating costs of contact centers (Aksin et al. (2007), Gans et al. (2003)), management has clear incentives to operate with the minimal possible number of agents. Our main managerial insight is that the safety staffing needed in a blended system is order of magnitudes smaller than the safety staffing needed in a split system, for given operational and quality-of-service (QoS) requirements. Specifically, when the goals are to (i) satisfy certain QoS constraints for the inbound calls; (ii) maintain a pre-specified throughput rate of outbound calls; and (iii) ensure that at most a negligible proportion of outbound calls are dropped; blending can achieve these goals with a substantially smaller number of agents than the number needed for the split system.

Employing a simple threshold policy, in the spirit of Bhulai and Koole (2003) and Gans and Zhou (2003), combined with a *logarithmic safety-staffing rule* (see §2 below), we show that, possibly after a short initial time, the number-of-busy-agents process experiences stochastic fluctuations that are proportional to $\log(n)$, when there are n agents in the system, giving rise to our staffing suggestion.

Our analysis builds on simple many-server heavy-traffic arguments, and simulation experiments demonstrate the effectiveness and accuracy of the asymptotic analysis.

In ending we remark that we study the operational aspects of blending, without addressing human-related issues. It is significant that blending may require longer agent training, as well as create a more stressful working environment for the staff, since agents need to continuously switch between the two types of calls. Management should therefore weigh human factors against the operational advantages that blending offers when considering whether to employ blending or not.

1.1. Literature Review

There is a vast literature that is dedicated to staffing and control of inbound call centers. For a thorough study and literature review we refer to Gans et al. (2003) and Aksin et al. (2007), which also review more general contact centers. In contrast, the literature on the staffing and control of outbound contact centers in general, and call blending in particular, is much smaller. We review the most relevant work to ours.

Outbound Contact Centers and Blending. In Reynolds (2010), the basics of staffing and forecasting for outbound call centers and contact centers with call blending are discussed. Samuelson (1999) considers the problem of dropped outbound calls in outbound call centers (with no blending). In practice, automatic dialers are sometimes programmed to initiate calls even when all agents are busy. If a customer replies sooner than anticipated, or a call lasts for a longer time than predicated, then there may be no available agent to take care of that customer, in which case his call is dropped ("abandoned" by the system). Government regulations require that the number of dropped outbound calls be small (3% or less of the number of calling attempts), and often management wants that no outbound calls will be dropped at all. On the other hand, predictive dialing can decrease agents' idling time, if it is done correctly. To deal with these conflicting requirements, an algorithm is developed in Samuelson (1999), utilizing queueing theory and simulation, aimed at maximizing the number of calling attempts per representative, subject to an upper bound on the proportion of calls abandoned due to unavailability of agents. This algorithm is the basis for "predictive dialing" software packages, which are commonly used in modern outbound contact centers.

Bhulai and Koole (2003), Gans and Zhou (2003) and Deslauriers et al. (2007) deal with call blending directly. The first two references provide an optimal control, which is of threshold type, when the service rates of the two types of jobs are equal. Gans and Zhou (2003) go beyond this setting, and show that a threshold policy is optimal among all policies that give priority to inbound calls. Unfortunately, the computational effort required to calculate the optimal parameters is large. Our control is inspired by the first two references. We refer to §5 of Gans and Zhou (2003) for

a discussion on why asymptotic analysis is useful, even for the exact same model as theirs, but emphasize that both our model and our objectives are different than theirs. Specifically, we consider the combined problem of staffing and control, incorporating inbound-customer abandonment and dropped outbound calls, and describe the transient (time-dependent) behavior of the system, in addition to the stationary analysis.

In Deslauriers et al. (2007), five Markovian models with inbound and outbound calls are studied in contact centers having two types of agents – inbound only and blend. As in our model, customer abandonment is assumed. In their models, a dialer automatically determines when to make outbound calls and how many as a function of the system's state, using a threshold policy, motivated by Bhulai and Koole (2003). The approach taken in Deslauriers et al. (2007) is that of describing each model as a CTMC on a finite state space (there is a finite buffer for waiting inbound calls). The M_5 model in §2.5 in Deslauriers et al. (2007) is related to our model, but ours assumes an infinite state space. As the authors explain in §3, even if the state space is not very large, it is computationally costly to compute performance measures when the service rates for the two types of calls are different.

Blending has also been considered by the industry, and a number of related patents exist. In Dumas et al. (1996), unlike in our model, inbound customers may choose to be called back (and become "outbound customers") and outbound customers have the option to wait in queue (although their waiting times tend to be shorter than those of inbound customers), i.e., they are not dropped. Based on extensive simulation experiments, it is shown that blending inbound and outbound calls and employing a threshold policy, ensure that the outbound throughput rate is met while waiting times of customers are very short. It is also shown that blending the two types of calls in one pool requires significantly less agents than employing two distinct pools. The examples in that patent demonstrate that blending can reduce the number of agents by 10% to 17% in contact centers with a number of agents in the order of 10's.

Other Types of Outbound Calls. The assumption of infinite supply of outbound work that we and the papers cited above take, is not valid for all outbound contact centers. For example, Armony and Maglaras (2004a) considers a single-pool contact center offering a call-back option, which is a type of outbound work. In this model, the outbound calls are performed from a list of waiting customers that are a subset of the customers who previously called the call center. In particular, the contact center is modeled as a two-class single-pool queueing system with customer balking. It is assumed that service times of the two classes are equal, and the systems are analyzed in the Halfin-Whitt (QED) many-server heavy-traffic limiting regime. It is shown that a threshold policy that gives priority to the inbound calls, as long as the queue of call-back customers is below a certain threshold, is asymptotically optimal in these settings, and that the guaranteed waiting

times of those customers who chose the call-back option are satisfied. A similar system is considered in Armony and Maglaras (2004b), but customers are informed about their waiting time in queue as well as the time until they will be called back, if they choose this option. We also mention Armony and Gurvich (2010) and Gurvich et al (2009), which study call centers that exercise cross-selling. The cross-selling phase is initiated by the agent at the end of the requested service, and can thus be considered to be a type of outbound work.

Immediate Response to Inbound Calls. In our main results we will prove that we can have practically all inbound calls admitted to service immediately upon their arrival. Inbound call centers that provide such service levels are said to operate in the *quality driven* (QD) regime. The classical reference for a single many-server pool Markovian model with many extra agents is Iglehart (1965), which proves the asymptotic equivalence of this model to the infinite-server queue.

In Whitt (1999), dynamic staffing in a call center is considered, with the objective of immediately answering all calls when the arrival rate is time-dependent and even stochastic. The author suggests to have a large number of extra agents working on other types of jobs (not answering inbound calls), that will be available to help with the inbound calls on a short notice. In essence, there are two pools of agents: one pool of dedicated agents, and a second pool to which outbound calls can be routed when this is needed. The total number of agents needed for each future time t is computed using an infinite-server approximation, by estimating the mean and variance of the calls that will still be in progress at that time, and the number of future arrivals; see (4.1) and (4.2) in that reference.

Safety Staffing. There is a vast literature on safety staffing of many server systems in heavy traffic, emanating from the seminal paper by Halfin and Whitt (1981). Halfin and Whitt studied the $GI/M/N$ model, having a renewal arrival process and N exponential servers, as the number of servers and the arrival rate λ grow to infinity together, such that $N = \lambda/\mu + O(\sqrt{\lambda})$[1], where μ denotes the service rate of an individual server, and is kept fixed as $\lambda \to \infty$. Intuitively, λ/μ is the minimal number of servers needed to ensure that the processing rate of jobs can match the arrival rate, and the additional $O(\sqrt{\lambda})$ number of servers - the so-called "square-root safety staffing" - is needed to take care of the stochastic variability in the system, ensuring that the probability that an arrival is delayed in queue is strictly between 0 and 1 as $\lambda \to \infty$. The need for a square-root safety staffing was first observed in Erlang (1917).

The square-root safety staffing rule has been generalized to many-server systems with abandonment, see Garnett et al. (2002) and Mandelbaum and Zeltyn (2009), and, e.g., Gurvich and Whitt (2009) and references therein for multiclass models with SBR. See also Bassamboo et al. (2010) for

[1] For a real function $f(\cdot)$, $O(f(x))$ is a quantity that grows proportionally to $f(x)$: $\limsup_{x\to\infty} O(f(x))/f(x) < \infty$.

a proposed capacity-prescription safety-staffing rule, derived via a suitable newsvendor problem for the $M/M/n+M$ model with a random arrival rate.

Resource Flexibility. Finally, our analysis adds to the literature on the benefits of flexibility in resource allocation; see, e.g., Bassamboo et al. (2009) and Tsitsiklis and Xu (2012) and references therein. In our setting, the flexibility attributed to outbound calls, when blending is employed, allows the system to achieve high service level targets for both inbound and outbound calls with a minimal number of "safety staffing", giving rise to the logarithmic safety staffing rule.

2. The Model

We consider a system with a single pool of homogeneous agents and two types of calls – inbound and outbound. We will sometimes refer to inbound calls as class-1 customers, and to outbound calls as class 2 (numbered in the order of their priority: class-1 customers have priority over class-2 customers). Class 1 customers arrive to the system according to a Poisson process. If a class-1 customer is not routed to service immediately upon arrival, then he waits in queue for his turn to be served, with customers being served in the order of arrival. However, a waiting inbound customer has a finite patience, and will abandon if his waiting time in queue exceeds a random time that is exponentially distributed with mean $1/\theta_1$. Unlike class 1, we assume that there is an infinite supply of class-2 customers, so that an available agent can always serve such a customer, if that is desired. The service times of all class-i customers are assumed to be exponential random variables with mean $1/\mu_i$, $i=1,2$. Finally, service times, time to abandon, and class-1 arrivals are all mutually independent.

Since we consider a system having a large pool of agents, a large volume of incoming calls and a large throughput of outbound calls, it is appropriate to employ many-server heavy-traffic approximations. To that end, we consider a sequence of systems, indexed by superscript n. For each $n \geq 1$ we let N^n denote the number of agents in the pool, λ_1^n denote the class-1 arrival rate, and η_2^n be the *target* throughput rate of outbound calls. It is significant that these three parameters are asymptotically proportional to one another, so that neither one is negligible in the limit. More formally,

Assumption 1. (*heavy-traffic scaling*) *For strictly positive real numbers N, λ_1 and η_2 it holds that*

$$N^n/n \to N, \quad \lambda_1^n/n \to \lambda_1 \quad \textit{and} \quad \eta_2^n/n \to \eta_2 \quad \textit{as } n \to \infty.$$

In contrast to the parameters in Assumption 1, the service and abandonment rates are fixed along the sequence, i.e., are independent of n. This is consistent with the fact that the service needs of customers and their (im)patience is independent of the size of the system.

For $n \geq 1$, let $Q_1^n(t)$ denote the number of class-1 customers waiting in queue to be served, and $Z_i^n(t)$ be the number of class-i customers in service at time t, $i = 1, 2$. Let $\{I^n(t) : t \geq 0\}$ denote the idleness process, i.e., $I^n(t) := N^n - (Z_1^n(t) + Z_2^n(t))$ is the number of idle agents at time t. (We use $:=$ to denote equality by definition.)

If agents are assigned to outbound calls whenever such a call is initiated (we elaborate on this issue in §2.2 below), then given a state-dependent routing policy, the process $\{X_3^n(t) : t \geq 0\}$ defined by

$$X_3^n(t) := (Q_1^n(t), Z_1^n(t), Z_2^n(t)), \quad t \geq 0, \tag{1}$$

is a three-dimensional *continuous time Markov chain* (CTMC). Moreover, it is easily seen that X_3^n possesses a stationary distribution, regardless of the values of N^n and λ_1^n, due to customer abandonment. (For notational convenience, we will often remove the time argument t when the whole process is considered, e.g., $X_3^n := \{X_3^n(t) : t \geq 0\}$.)

Whereas η_2^n is the target throughput rate of outbound calls, which management wants to achieve, there is also an *actual* throughput rate, which depends on system's performance. With the notation above, we define the actual *instantaneous* throughput rate of outbound calls at time t to be $r_2^n(t) := \mu_2 Z_2^n(t)$. In an infinite-horizon case, the *actual throughput rate* in system $n \geq 1$ is defined as the almost-sure limit

$$r_2^n := \lim_{t \to \infty} \frac{1}{t} \int_0^t \mu_2 Z_2^n(s)\, ds = \mu_2 E[Z_2^n(\infty)], \tag{2}$$

where $Z_2^n(\infty)$ denotes a random variable that has the stationary distribution of the process Z_2^n. The second equality in (2) is due to the well-known ergodic theorem for CTMC's. Of course, one goal is to have the actual throughput rate r_2^n be equal to the target η_2^n.

We next make an assumption about the initial condition.

ASSUMPTION 2. (*initial condition*) *There exists* $a > 0$ *such that* $P(\bar{Q}_1^n(0) > a) \to 0$ *as* $n \to \infty$. [2]

It is important to allow for a general initial condition that does not converge to a specific value, as in Assumption 2, because the arrival rates of inbound calls change throughout the day, and can be assumed fixed for only certain time periods. In applications, one has to ensure that the time interval under consideration (having fixed rates) is long enough so that the actual throughput rate of outbound calls, r_2^n in (2), is sufficiently close to the desired throughput rate η_2^n. Theorem 4 below provides a simple method to guarantee that.

[2] Note that only the first component of $\bar{X}_3^n$, namely the class-1 queue, is considered in Assumption 2, since the other two components, $\bar{Z}_1^n$ and $\bar{Z}_2^n$, are both bounded w.p.1 by N^n/n which converges to a finite number N by Assumption 1. We remark that our results below hold if we replace Assumption 2 with a strictly weaker condition; see Remark 3 below. We chose to make Assumption 2 in its current form for simplicity of exposition, and since it should clearly hold in practice.

2.1. Advantages of Blending: The Logarithmic Safety Staffing Rule

Staffing of outbound pools depends on whether agents manually dial customers, or an automatic dialer is used. In the example in Reynolds (2010, page 3), 40% of agents' time is involved in look-up and dialing, leading to the conclusion that "... the inefficiencies of a manual dialing process... make the return on investment of an automated dialer significant". Whether automatic dialers are employed or not, the average time until an outbound customer replies needs to be considered for staffing purposes. To compute the workload on the pool for a given desired output, this mean response time *is added to the expected actual service time*; we again refer to Reynolds (2010). For example, if the actual mean service time is 5 minutes and the expected time until a customer replies is $1/M = 9$ seconds (0.15 minutes), then the workload is computed assuming an average service time of $1/\mu_2 = 5.15$ minutes.

Note that this implies that agents spend some of their time waiting for customers to reply to a call. Those waiting agents are not considered idle, and are referred to as *waiting agents*. Clearly, the number of waiting agents in an outbound pool is proportional n/M, when there are n agents in the pool. Hence, to achieve the desired throughput rate of calls, at least η_2^n/μ_2 agents are required for the outbound pool in a split system, where $1/\mu_2$ is computed as described above.

As was reviewed in §1.1, the dedicated pool for inbound calls must have $\lambda_1^n/\mu_1 + O(n)$ agents to ensure that practically all inbound calls are answered immediately. That is, in a split system having two dedicated service pools, the total number of agents needed to achieve our service requirements for both types of calls is $\eta_2^n/\mu_2 + \lambda_1^n/\mu_1 + O(n)$. As we will show below, the same service goals, for both types of calls, can be achieved with an additional $O(\log n)$ agents when blending is employed instead of splitting. In particular, with blending it is sufficient to have the number of agents be $N^n = \lambda_1^n/\mu_1 + \eta_2^n/\mu_2 + O(\log n)$.

We remark that less ambitious service levels are often desired for inbound calls, e.g., inbound call centers that offer good service levels often operate in the QED regime, and not in the QD regime. Our analysis will make it clear that any other service level for inbound calls, in term of waiting times in queue and abandonment, can be achieved when blending is employed, with a total number of agents that is substantially smaller than the number needed in the split setting; see §4 below.

2.2. More on Dropped Outbound Calls

As was reviewed above, automatic dialers are often employed in order to speed up the calling process and increase the outbound throughput. Nevertheless, automated dialing may lead to a nonnegligible number of dropped outbound calls, because the number of inbound arrivals during any time interval may be larger than the number of departures over that interval. In a large system

having a large volume of calls, the difference between the numbers of arrivals and departures may be substantial even over short time intervals.

Specifically, if one employs a threshold policy as in Bhulai and Koole (2003) and Gans and Zhou (2003) with a fixed threshold $K^n = K$ that does not increase with n, then it can be shown that the idleness process fluctuates between 0 and K *infinitely often* on every interval, no matter how small, as n increases indefinitely. Furthermore, there will be no idleness in the system for a non-negligible proportion of time. Therefore, even if the time until outbound customers reply is short relative to their service time, it may still be long relative to the fluctuations of the idleness process.

This suggests that the relevant time scale regarding dropped calls is associated with the behavior of the idleness process I^n, which turns out to operate in a fast time scale (in $o(1)$ scale[3]) that grows faster as n increases; see §3.2 and the proofs of Theorem 1 and Lemma 5 below. To ensure idleness throughout, the threshold K^n should therefore increase with n. Our analysis identifies the *minimal* order of size that is needed for K^n as being $O(\log(n))$.

Our Modeling Approach. To explicitly incorporate dropped calls in the model would require tracking the process of outbound calls that are yet to be replied. This added dimension presents complicated modeling and analytical issues, because the corresponding stochastic process is small relative to the processes comprising X_3^n in (1); we elaborate in Appendix B. We therefore start by considering the model described in §2, in which an outbound customer replies to a call immediately upon its initiation. In Appendix B we show that a logarithmic safety staffing is sufficient to prevent dropped calls in a more involved model which explicitly takes into account the random positive time it takes an outbound customer to reply to a call. In this case, waiting agents may start helping inbound arrivals before their assigned outbound customers reply, so that dropped calls can occur if staffing levels are not adequate.

A Simulation Example. To see the significance of the two time scales in a practical setting, we conducted a simulation for a system with the following parameters:

$$N^n = 500, \quad \lambda_1^n = 350, \quad \mu_1 = 1, \quad \mu_2 = 0.5, \quad \theta_1 = 0.5.$$

These parameters imply that the target throughput rate is $\eta_2^n = 0.5 \cdot 150 = 75$ per unit time. Here, time is directly measured in class-1 service-time units (since $\mu_1 = 1$).

We initialize the simulation with $Z_1^n(0) = 350$ and $Z_2^n(0) = 150$ and with no class-1 customers in queue. In the simulation, an outbound call is initiated at time t if and only if $I^n(t) > 5$, i.e., we are using a simple threshold policy on the idleness process with a threshold $K^n = 5$: If, when an

[3] $o(f(n))$ denotes a quantity that grows slower than $f(n)$: $\lim_{n\to\infty} o(f(n))/n = 0$. In particular, $\lim_{n\to\infty} o(1) = 0$.

agent becomes available, there are exactly 5 additional idle agent, then one of the idle agents is immediately assigned to an outbound call. (In particular, there are never more than 5 idle agents.)

Figure 1 shows one sample path of the idleness process during 5 time units. It is clear from this figure that I^n moves very fast among its states $\{0,\ldots,5\}$. That same sample path is shown in Figure 2 over a short time interval of length 0.4 time units. Observe that around the time points $t_1 = 0.7$, $t_2 = 0.75$ and $t_3 = 0.85$ the idleness process drops very quickly from five to zero. Observe also that the idleness process can spend a significant amount of time at state 0 (in which case a queue forms).

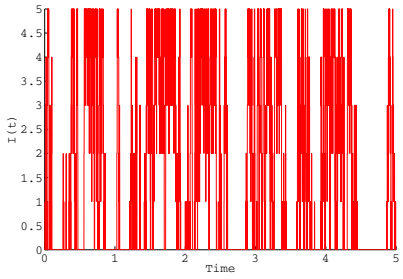

Figure 1 **Fast fluctuations of idleness process.**

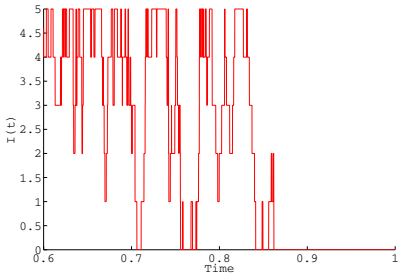

Figure 2 **Short time interval.**

3. Main Results

We consider contact centers in which high quality of service, in terms of waiting times and proportion of abandonment, is desired for the inbound calls, while a specified throughput rate of outbound calls should be maintained. Specifically, our goal is to design the system such that the number of inbound customers who have to wait before entering service is negligible and, at the same time, the desired throughput rate η_2^n is achieved with at most a negligible number of dropped outbound calls. Analogously to the inbound call-center terminology (e.g., Gans et al. (2003)), we refer to this service regime as being *quality-driven* (QD), because both customer types receive a high quality of service. We discuss extensions to other service regimes in §4 below.

The next assumption determines the staffing in the system, which follows the logarithmic safety staffing rule discussed above.

ASSUMPTION 3. (*QD logarithmic-safety-staffing rule*) *The number of agents in system $n \geq 1$ satisfies*

$$N^n := \lceil \lambda_1^n/\mu_1 + \eta_2^n/\mu_2 \rceil + K^n,$$

where K^n is a positive integer satisfying $K^n/\log(n) \to K$ as $n \to \infty$, for some $0 < K < \infty$.

Observe that Assumptions 1 and 3 imply together that $\lambda_1 < \mu_1 N$, which further implies that class-1 customers can receive excellent service, in terms of waiting times in queue, if they are given priority over class-2 customers. The following routing policy ensures that this is indeed the case.

Definition 1 (Priority Control with Idleness Threshold). Given K^n in Assumption 3 the idleness-threshold control is as follows.

- **Upon arrival of a class-1 customer**, the customer will be routed to an available agent. If no agent is available, the customer will wait in queue to be served in the order of arrival.
- **Upon completion of service**, the newly available agent will take a customer from the head of the class-1 queue. If the queue is empty and there are less than K^n other idle agents, then the newly available agent will idle and all other idling agents will remain idle. Otherwise, if there are at least K^n additional idle agents, then a class-2 customer will start to be served by one of those idling agents.

With the policy just described, it is clear that $I^n(t) \leq K^n$ for all $t \geq 0$, given the assumption that an outbound call is successful immediately, because an outbound call is initiated when an agent becomes available and there are K^n additional idle agents at that time. Therefore, the maximum possible idleness in the system is of order $\log(n)$ by our choice of K^n in Assumption 3. We henceforth refer to K^n as the "idleness threshold", or simply the "threshold" when the meaning is clear.

Additional notation. To present our results we need to introduce some more notation. We use the usual $\mathbb{R}$ and $\mathbb{R}_k$ notations to denote, respectively, the real numbers and k-dimensional vectors with components in $\mathbb{R}$, $k \geq 1$. We let $\Rightarrow$ denote convergence in distribution of random variables, or stochastic processes. In the latter case, the limits will always have continuous sample paths, and the convergence holds uniformly on compact intervals in an appropriate function space (we elaborate further in §7). Let $\|\cdot\|$ denote the Euclidean norm in $\mathbb{R}_k$ and let e be the identity function, i.e., $e(t) := t$, $t \in I$, for some interval $I \subset [0, \infty)$. Finally, as in Assumption 2, we use a 'bar' to denote "fluid-scaled" processes: for a process $Y^n := \{Y^n(t) : t \geq 0\}$, $\bar{Y}^n := Y^n/n$.

Throughout this section, Assumptions 1, 2 and 3 are assumed to hold, and the priority control in Definition 1 is employed.

3.1. Results for the Transient Period

We now present results for the transient period, as the system approaches its steady state. The proofs of those results appear in §7.

The three-dimensional process X_3^n in (1) is evidently hard to analyze, even asymptotically (see its sample-path representation in (10) below). One way to simplify the problem is to consider a system with equal service rates for both types of calls, i.e., with $\mu_1 = \mu_2$. In that case, the total output rate due to service completions depends only on the number of agents working, and is independent

of how many agents work with each class. This approach was often taken in the literature, as was reviewed in §1.1. However, when $\mu_1 \neq \mu_2$, all three components of the CTMC X_3^n are relevant for each $n \geq 1$. A significant simplification is achieved in the heavy-traffic limits using the following result, which states that X_3^n is asymptotically equivalent, after possibly some finite time T, to an essentially one-dimensional process X^n defined below. Such types of results are typically referred to as *state-space collapse* (SSC) in the literature.

Let

$$X^n(t) := (0, Z_1^n(t), N^n - Z_1^n(t)), \quad t \geq 0, \tag{3}$$

and note that X^n is completely characterized by Z_1^n, and is thus effectively one dimensional.

THEOREM 1. (*eventual SSC*) *There exists a time T, $0 \leq T < \infty$, such that $\bar{Q}_1^n \Rightarrow 0e$ uniformly over $[T, \infty)$ as $n \to \infty$. As a result, the process $\bar{X}_3^n$ is asymptotically equivalent to the process $\bar{X}^n$, for X^n in (3), in the sense that $\sup_{T \leq t < u} \|\bar{X}_3^n(t) - \bar{X}^n(t)\| \Rightarrow 0$ as $n \to \infty$, $T < u < \infty$.*

Theorem 1 implies that $\bar{Q}_1^n$ is null after time T, for all n large enough. The reason why this SSC only holds after some time T is due to the possibility that the system is initially overloaded. For example, we might have that $\bar{Q}_1^n(0) \Rightarrow q_1(0)$ as $n \to \infty$, for some $q_1(0) > 0$. Even if $q_1(0) = 0$ there is no guarantee that the conclusion in Theorem 1 holds before some time $T > 0$, because the value of $Z_1^n(0)$ might be such that the system is temporarily overloaded, so that a queue builds up; a numerical example is given in §5.2 below. In §7.4 we explain how T can be estimated. We emphasize that T is finite and relatively short, even in extreme cases. In particular, Theorem 1 does not hold only in steady state.

The next theorem strengthens Theorem 1 significantly.

THEOREM 2. *Consider time T in Theorem 1. Then, for all $t \geq T$, $P(I^n(t) > 0) \to 1$ as $n \to \infty$.*

Theorem 2 implies that $P(Q_1^n(t) > 0) \to 0$ as $n \to \infty$ for all $t > T$. That follows immediately from the fact that the idleness process $I^n(t)$ is asymptotically positive for each $t \geq T$, for T in Theorem 1. The significance of this result is clear: it implies that practically every class-1 customer will enter service immediately upon arrival, and that no outbound call will be dropped due to agents' unavailability, provided that n is large enough.

3.2. Intuition Behind the Logarithmic Safety Staffing Rule

The proofs of Theorems 1 and 2 will show that the stochastic fluctuations of the number-in-system process $Q_1^n + Z_1^n + Z_2^n$ above the idleness threshold K^n are of order $\log(n)$, and that, even though the idleness process might reach state 0 (in which case there are no idle agents), the proportion of time it spends in that state converges to zero as $n \to \infty$. In particular, we show that $O(\log(n))$ is *the minimal order* of the idleness threshold which ensures that there is idleness for almost all

$t \geq T$, asymptotically. We now provide a heuristic explanation for that result, and refer to Perry and Whitt (2011) and Gurvich and Perry (2012) for a refined analysis of similar phenomenons.

Recall that the idleness process I^n operates in a faster time scale as n grows. Equivalently, the process

$$Y^n := Q_1^n + Z_1^n + Z_2^n - (N^n - K^n),$$

which captures the fluctuations of the number-in-system process above the idleness threshold K^n, becomes faster as n increases. In contrast, the process $\bar{X}^n$ has small changes over short time intervals because, as n grows, this process approaches a continuous limit. Specifically, for large n, the process $\bar{X}^n$ is almost constant over the time interval $[t, t+\epsilon)$ when ϵ is small, i.e., $\bar{X}^n(s) \approx \bar{X}^n(t)$ for all $s \in [t, t+\epsilon)$ and for any fixed time $t \geq 0$.

Consider Y^n over the interval $[t, t+\epsilon)$. Observe that $Y^n(t)$ increases by 1 if an inbound customer arrives, which happens with rate λ_1^n, and decreases by 1 if a customer leaves the system, which happens with rate $\beta_t^n := \theta_1 Q_1^n(t) + \mu_1 Z_1^n(t) + \mu_2 Z_2^n(t)$ at time t. However, since $\bar{X}^n$ is almost fixed for small ϵ and large n, β_t^n is (approximately) fixed over $[t, t+\epsilon)$,. Therefore, Y^n is approximately distributed as an $M/M/1$ queue with arrival rate λ_1^n and service rate β_t^n over that interval. Let $Q_t^n := \{Q_t^n(u) : u \geq 0\}$ denote the queue process of an $M/M/1$ system with the same arrival and service rates. (The subscript t in β_t^n and Q_t^n stands for the fact that the system $X^n(t)$, via which β_t^n is defined, is considered to be fixed at its value at time t.)

As we will show, there exists a time $T \geq 0$ such that, for all $t \geq T$, it holds that $\lambda_1^n < \beta_t^n$ with a probability approaching 1 as n grows large. Hence, Q_t^n is, for sufficiently-large n, positive recurrent. Moreover, the process Q_t^n with rates β_t^n and λ_1^n over $[t, t+\epsilon)$ is equal in distribution to a "slowed" $M/M/1$ with rates β_t^n/n and λ_1^n/n over the time-scaled interval $[t, t+n\epsilon)$. It is well-known that the maximum of a positive-recurrent $M/M/1$ queue grows as $O(\log n)$ over intervals of length $O(n)$ as $n \to \infty$ (see Anderson (1970)), and that the queue process keeps returning to state 0. Since Q_t^n is a fast version of the slowed $M/M/1$, it follows that Q_t^n, and thus $\{Y^n(s) : t \leq s < t+\epsilon\}$, have infinitely-many fluctuation of order $\log(n)$ as $n \to \infty$ over $[t, t+\epsilon)$, no matter how small ϵ is. The same holds for an ϵ-neighborhood of any $t \geq T$.

By choosing the idleness threshold to be in logarithmic scale, we can guarantee that the probability that all agents are busy converges to 0, at least after some short time T. Moreover, our proofs will show that this holds regardless of the exact choice of the threshold; in particular, we only require that K^n grows at least as fast as $\log(n)$.

3.3. Stationarity

Due to the complexity of the system, we cannot prove convergence of the sequence of processes $\bar{X}_3^n$ as $n \to \infty$. Nevertheless, we can show that the sequence of stationary distributions $\{\bar{X}_3^n(\infty) : n \geq 1\}$

converges to a simple deterministic limit in $\mathbb{R}_3$, where $X_3^n(\infty)$ denotes a random variable in $\mathbb{R}_3$ that has the distribution of the steady state of the process X_3^n.

Let

$$x^* := (q_1^*, z_1^*, z_2^*) = (0, \lambda_1/\mu_1, N - \lambda_1/\mu_1)\,. \tag{4}$$

THEOREM 3. (*limit of stationary distributions*) $\bar{X}_3^n(\infty) \Rightarrow x^*$ *in* $\mathbb{R}_3$ *as* $n \to \infty$.

Note that Theorem 3 together with Assumptions 1 and 3 imply that $\bar{Z}_2^n(\infty) \Rightarrow \eta_2/\mu_2$, so that, for large n, $\mu_2 Z_2^n(t) \approx \eta_2^n$, i.e., the actual instantaneous throughput rate is close to its target, provided that t is large enough so that the system is sufficiently close to its steady state. As we already mentioned (see the discussion below Assumption 2), the arrival rates change slowly throughout the day and can only be assumed fixed for finite time periods. It is thus important to verify that the system approaches its desired steady state before the arrival rates change significantly. In other words, it is important to assess the speed at which the system approaches stationarity. The next theorem describes the typical asymptotic behavior of the system, showing that the convergence rate to the limiting stationary state is exponentially fast.

THEOREM 4. *There exists* $C > 0$ *such that, for every deterministic element* $x^C := (0, z_1^C, N - z_1^C)$ *of* $\mathbb{R}_3$ *satisfying* $\|x^C - x^*\| < C$, *if* $\bar{X}_3^n(0) \Rightarrow x^C$ *in* $\mathbb{R}_3$, *then* $\bar{X}_3^n \Rightarrow x := \{x(t) : t \geq 0\}$ *uniformly over* $[\delta, u]$ *as* $n \to \infty$, *for any* $0 < \delta < u < \infty$, *where* $x(t) := (0, z_1(t), N - z_1(t))$ *and*

$$z_1(t) = \lambda_1/\mu_1 + \left(z_1^C - \lambda_1/\mu_1\right) e^{-\mu_1 t}. \tag{5}$$

In particular, $x(t) \to x^*$ *as* $t \to \infty$.

By Theorems 3 and 4 the point x^* in (4) serves as a good approximation for $\bar{X}_3^n(t)$ for large n and t when the fluid-scaled system is considered. However, once the system stabilizes at the neighborhood of x^*, it becomes valuable to capture stochastic fluctuations that are of smaller order than $O(n)$ to approximate the variability of the system.

Let $\hat{X}_3^n$ denote the diffusion-scaled processes of X_3^n:

$$\hat{X}_3^n := (\hat{Q}_1^n, \hat{Z}_1^n, \hat{Z}_2^n), \tag{6}$$

$$\text{where} \quad \hat{Q}_1^n := \frac{Q_1^n}{\sqrt{n}}, \quad \hat{Z}_1^n := \frac{Z_1^n - n z_1^*}{\sqrt{n}}, \quad \hat{Z}_2^n := \frac{Z_2^n - n z_2^*}{\sqrt{n}} = -\hat{Z}_1^n.$$

Let

$$\hat{X} := (0e, \hat{Z}_1, -\hat{Z}_1), \quad \text{where} \quad \hat{Z}_1(t) = \hat{Z}_1(0) + \sqrt{2\mu_1} B(t) - \mu_1 \int_0^t \hat{Z}_1(s) ds, \quad t \geq 0, \tag{7}$$

with $\{B(t) : t \geq 0\}$ denoting is a standard Brownian motion.

Observe that $\hat{Z}_1$ is an Ornstein-Uhlenbeck (OU) process with (state-dependent) drift $m(x) = -\mu_1 x$ and infinitesimal (state-independent) variance $\sigma^2(x) = 2\mu_1$ for all $x \in \mathbb{R}$. It is well-known that the steady state distribution of this OU process is a standard normal, i.e., it is normally distributed with mean 0 and variance 1.

THEOREM 5. (*diffusion limit*) *If* $\hat{X}_3^n(0) \Rightarrow \hat{X}(0)$ *in* $\mathbb{R}_3$ *as* $n \to \infty$ *for some random variable* $\hat{X}(0)$ *of the form* $\hat{X}(0) = (0, \hat{Z}_1(0), -\hat{Z}_1(0))$, *then* $\hat{X}_3^n \Rightarrow \hat{X}$ *uniformly over compacts as* $n \to \infty$, *for* $\hat{X}$ *in* (7).

REMARK 1. The assumption $\hat{X}_3^n(0) \Rightarrow \hat{X}(0)$ as $n \to \infty$ in the statement of Theorem 5 implies that the initial conditions of the *fluid-scaled* processes $\bar{X}_3^n$ converge to the stationary state x^*, and thus, we have $P(I^n(t) > 0) \to 1$ as $n \to \infty$ for all $t > 0$, and not only after some time T which may be greater than 0, as in Theorem 2.

3.4. Implications of the Results

It follows from basic CTMC theory that $\bar{X}_3^n(t) \Rightarrow \bar{X}^n(\infty)$ as $t \to \infty$ for each $n \geq 1$. Theorem 3 shows that the stationary stochastic system is well approximated by x^* in (4), and Theorem 4 shows that convergence to stationarity is exponentially fast. Once $\bar{X}_3^n$ is close to x^*, the throughput rate of outbound calls is within $o(n)$ of the desired rate. To see this, observe that for N^n in Assumption 3 and x^* in (4),

$$N = \lim_{n\to\infty} N^n/n = \lambda_1/\mu_1 + \eta_2/\mu_2 = z_1^* + z_2^*. \tag{8}$$

Now, since $z_1(t)$ converges exponentially fast to $z_1^* = \lambda_1/\mu_1$ as $t \to \infty$ by Theorem 4, it follows from (8) and Theorem 1 that $\mu_2 z_2(t)$ converges exponentially fast to η_2. Moreover, for r^n in (2), we have the following immediate corollary to Theorem 3.

Corollary 1 *The sequence* $\{r_2^n/n : n \geq 1\}$ *converges to* η_2 *w.p.1.*

Proof. We have that

$$\frac{r_2^n}{n} = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mu_2 \bar{Z}_2^n(s) ds = \mu_2 E[\bar{Z}_2^n(\infty)] \to \mu_2 z_2^* \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty,$$

with the first limit holding w.p.1, for each $n \geq 1$ due to the ergodicity of Z_2^n, and the second limit holding due to Theorem 3 together with the bounded convergence theorem. The statement follows from the fact that $z_2^* = N - z_1^* = N - \lambda_1/\mu_1 = \eta_2/\mu_2$, where the last equality follows from (8). □

Recalling that the fixed arrival rates and staffing hold over finite time intervals in applications (which are relatively long compared to the average service times), Corollary 1 allows management to estimate whether a given time interval is sufficiently long for the required performance measures to be achieved. Together with Theorem 4, one can then approximate $\int_I \mu_2 Z_2^n(t) dt$, where I is the time

interval under consideration, and determine whether I is long enough. (If not, the staffing should be adjusted.) In addition, Theorem 4 can be used to approximate the instantaneous throughput rate of outbound calls $\mu_2 Z_2^n(t)$, for all $t \geq T$, using the expression in (5), and recalling that, by Theorem 1, $z_2(t) = 1 - z_1(t)$ for all $t \geq T$. In any case, Theorem 2 guarantees that inbound calls receive QD service levels, and that there are almost no dropped calls, long before the system can be considered to be close to its steady state.

The significance of Theorem 5 is twofold. From a theoretical perspective, this theorem is a further illustration of our argument that, asymptotically, the system operates in the QD regime from the point of view of the inbound calls. This is because the diffusion limit $(0e, \hat{Z}_1)$ of $(\hat{Q}_1^n, \hat{Z}_1^n)$ is similar to the limit in Iglehart (1965) for a sequence of $M/M/N^n$ queues with many extra agents, i.e., when $\mu_1 N^n = \lambda_1^n + O(n)$. That is true despite the fact that, in our case, almost all agents are busy all the time. From the practical perspective, Theorem 5 is a stochastic refinement to the output process of outbound calls $r_2^n(t) := \mu_2 Z_2^n(t)$, $t \geq 0$, and can be used to approximate the variability in the system when it is approximately stationary.

4. Extensions

As we mentioned in §2.1, inbound call centers typically consider the QED regime as providing sufficiently-good service levels. In that case, the number of agents in a dedicated pool to inbound calls is $\lambda_1^n/\mu_1 + \beta\sqrt{\lambda_1^n}$, where β a real-valued constant. Other inbound call centers may even desire to save substantially on staffing by letting practically all arrivals be delayed in queue, and a proportion of those arrivals abandon. Such call centers are said to operate in the *efficiency-driven* (ED) regime. We refer again to Gans et al. (2003) Aksin et al. (2007), and Mandelbaum an Zeltyn (2009).

We now consider a modification of the staffing and control policy discussed in §2 with the aim of achieving a different service level for inbound customers. For concreteness, we consider the ED regime for inbound calls, in which a given proportion γ, $0 < \gamma < 1$, of the inbound calls is to abandon. As in the first case, the desired throughput rate of outbound calls and the arrival rate of inbound calls are assumed to satisfy Assumption 1, and it is still required that almost no outbound calls be dropped. Since, as before, we seek a simple and automatic control, we want to have some idleness almost all the time.

The staffing of the system in this case is as follows.

ASSUMPTION 4. (*ED logarithmic safety-staffing rule*) *The number of agents N^n in system $n \geq 1$ satisfies $N^n = \lceil \lambda_1^n(1-\gamma)/\mu_1 + \eta_2^n/\mu_2 \rceil + K^n$ where K^n is a positive integer satisfying $K^n/\log(n) \to K$, for some $0 < K < \infty$.*

The aim of our second policy is to ensure that a proportion γ of arrivals abandons. This is achieved by forcing all arrivals to wait in queue before entering service.

Definition 2 (Priority Control with Queue and Idleness Thresholds). Given positive integers K^n and L^n, the queue-and-idleness-threshold control is as follows.

- **Upon arrival of a class-1 customer** that finds L^n customers waiting in queue, the customer that has been waiting the longest will enter service, if an agent is available. If there are less than L^n customers in queue, then no class-1 customer enters service (even if there are idle agents).
- **Upon completion of service**, the newly available agent will take the class-1 customer that has been waiting the longest if there are at least L^n customers waiting in the class-1 queue. Otherwise, if there are at least K^n additional idling agents, then one of those idle agents will start serving a class-2 customer. If together with the newly available agent the total number of idle agents is less than K^n (and $Q^n(t) < L^n$), then no new customer, from either type, will enter service.

Note that this second policy is not non-idling with respect to the inbound work. In particular, there are time instances in which some agents idle even though customers are waiting in the queue of inbound calls. Note also that this control no longer gives strict priority to inbound calls over the outbound calls.

To completely define our second control policy, we need to specify the queue threshold L^n.

Assumption 5. *The sequence of thresholds* $\{L^n : n \geq 1\}$ *satisfies*

$$L^n/q_1^n \to 1 \quad \text{as } n \to \infty, \quad \text{where } q_1^n := \lambda_1^n \gamma/\theta_1. \tag{9}$$

We remark that for the Markovian Erlang A system having n agents and operating in the ED limiting regime (i.e., n is large and the traffic intensity is strictly larger than 1), the steady-state queue length $Q_1^n(\infty)$ and abandonment probability $P^n(AB)$ satisfy $\lambda_1^n P^n(AB) \approx \theta_1 E[Q_1^n(\infty)]$, with the approximation becoming exact in the limit as $n \to \infty$; see Equation (2.23) in Theorem 2.3 in Whitt (2004). Specifying a desired steady-state abandonment probability $P^n(AB) = \gamma$, we obtain the quantity q_1^n in (9).

The control in Definition 2 is a generalization of the first priority control in Definition 1, since in the first control we take $L^n = 0$ for all $n \geq 1$. As in the first case, at most $K^n = O(\log(n))$ agents can be idle so that Z_2^n completely determines Z_1^n, and vice versa.

Let

$$\tilde{x} := (\tilde{q}_1, \tilde{z}_1, \tilde{z}_2) = (\lambda_1\gamma/\theta_1, \lambda_1(1-\gamma)/\mu_1, N - \lambda_1(1-\gamma)/\mu_1)\,.$$

Analogous results to those in §3 hold in the current case. In particular, it can be shown that there exists a $T > 0$ such that Q_1^n has fluctuation of order $O(\log(n))$ about the threshold L^n and that

the idleness process is positive for almost all $t \geq T$. The counterparts of Theorem 3 and Corollary 1 can be shown to hold. Finally, a similar result to that in Theorem 4 can be proved with the limit

$$z_1(t) = \lambda_1(1-\gamma)/\mu_1 + (z_1(0) - \lambda_1(1-\gamma)/\mu_1)\, e^{-\mu_1 t}.$$

However, a diffusion limit for the second policy is much harder to derive than the limit in Theorem 5 because the fast fluctuations of the queue process about its threshold contribute to the variability of the limit. See Perry and Whitt (2014) and Theorem 4.1 in Gurvich and Perry (2012).

5. Simulation Experiments

In this section, we conduct simulation experiments to show the effectiveness and accuracy of our approximations when applied to stochastic systems.

The idleness and queue processes. The consequences of the theorems in §3 are illustrated in Figures 3-6 below. We conducted two simulation experiments which can be viewed as two elements in a sequence of queueing systems. In particular, we take two cases, the first with $n = 200$ and the second with $n = 500$ agents, to show how the scaling affects the behavior of the systems. The parameters of the two cases are as follows:

Case 1: $n = 200$; $K^n = 13 \approx 2.5\log(n)$; $\lambda_1 = 0.7n = 140$; $\mu_1 = 1$; $\mu_2 = 0.5$; $\theta = 0.5$.

Case 2: $n = 500$; $K^n = 16 \approx 2.5\log(n)$; $\lambda_1 = 0.7n = 350$; $\mu_1 = 1$; $\mu_2 = 0.5$; $\theta = 0.5$.

Figures 3 and 4 present a short time window of the queue and idleness processes in Case 1, while Figures 5 and 6 show the same processes in Case 2. As can be seen from Figures 4 and 6, the idleness process rarely hits zero, which implies that most of the time there is at least one available agent present. Overall, the proportion of time that all agents were busy was 0.057 for $n = 200$ and 0.047 for $n = 500$, over 100 units of time. As a result, the queue is mostly nonpositive. When a queue does build up, it is relatively small and goes very quickly back to zero. This can be seen in Figures 3 and 5. It is instructive to observe how small the numbers on the vertical axes of the queue figures are relative to the size of the system.

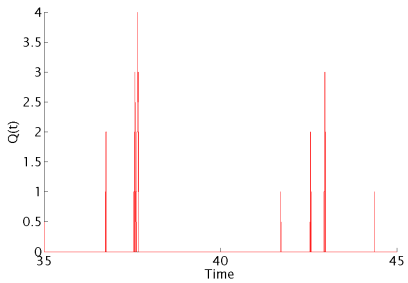

Figure 3 **Number in queue,** $n = 200$.

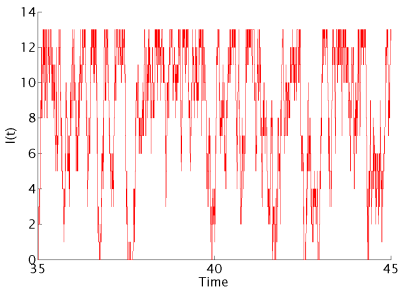

Figure 4 **Number of idle agents,** $n = 200$.

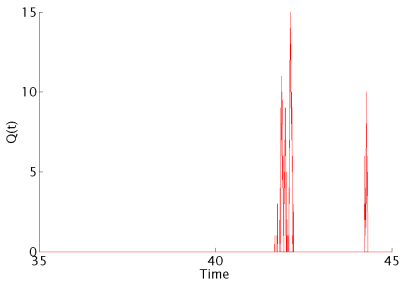

Figure 5 **Number in queue,** $n = 500$.

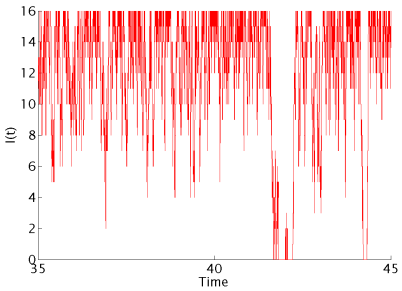

Figure 6 **Number of idle agents,** $n = 500$.

5.1. Simulations of the Routing Policy in §4

We now present simulation experiments that demonstrate the effectiveness of our approximations for the second control policy. We only simulated the case for $n = 200$, since the point on the effect of scaling was already made above. In particular, we simulated the system with 200 agents having the same arrival, service and abandonment rates as in Case 1 above. In addition, we take $K^n = 13$ as above, $\gamma = 0.1$ and $r = 0.37$, so that the threshold on the queue is $L^n = 28$.

Figures 7 and 8 show the fluctuations of the idleness process and of the queue process about their thresholds over a short time period. As can be seen, at almost all times there is at least one idle agent (overall, the proportion of time that all agents were busy was 0.065), and the queue does not deviate much from the threshold L^n.

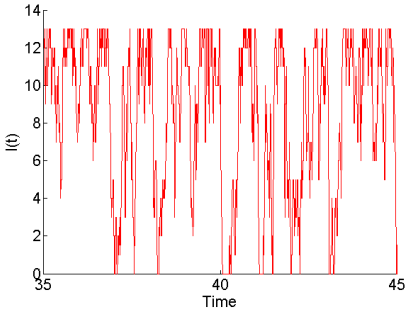

Figure 7 **Number of Idle Agents.**

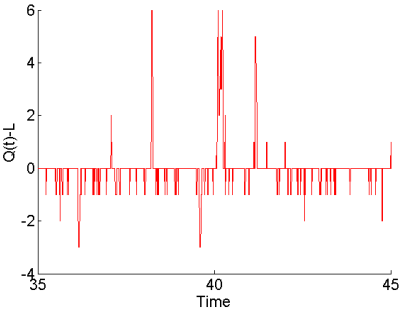

Figure 8 **Fluctuations about** L^n

5.2. The problem with the initial condition.

Recall that, due to the generality of the initial condition, it may take some time before the queue stabilizes near its target level. In particular, the SSC in Theorem 1 holds after a time T which may be strictly positive. To illustrate this point, we simulated the two systems in Cases 1 and 2 above with initial conditions that make each system overloaded initially. For Case 1 ($n = 200$), we initialized the system with $Z_1^n(0) = 50$ and $Z_2^n(0) = 150$. For Case 2 ($n = 500$), we initialized the system with $Z_1^n(0) = 125$ and $Z_2^n(0) = 375$. In both cases, the initial queue is zero: $Q_1^n(0) = 0$. We

conducted 50 independent simulation runs with the above initial conditions for each of the two cases. The results are shown in Figures 9 and 10.

Note that, for each individual sample path, the queue can become positive after time T, as in Figures 3 and 5. The average of the simulation runs, shown in Figures 9 and 10, is a strong-law type result, as are the fluid approximations, which our proofs build upon. In particular, the average behavior of the queues, as depicted in Figures 9 and 10, is an evidence to the value of the fluid approximation for the queue in Theorem 1, whereas Figures 3-6 represent the more refined analysis that is carried out in Theorem 2 and its proof.

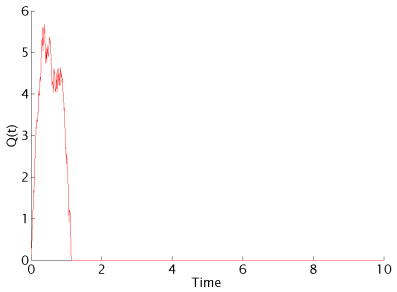

Figure 9 **Initial behavior,** $n = 200$.

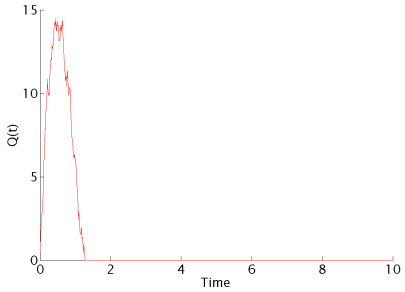

Figure 10 **Initial behavior,** $n = 500$.

6. Concluding Remarks

In this paper we demonstrated the operational advantages of call blending in contact centers that serve both inbound and outbound calls. In this setting, outbound calls are well modeled as an infinite supply of work that can be executed when management finds it convenient. The flexibility in the scheduling of outbound jobs reduces the variability in the system, which in turn decreases the number of agents that are needed to maintain desired service levels when compared to the splitting setting. In particular, we have shown that, in large systems, a logarithmic-safety-staffing rule is sufficient in order for the system to have idle agents at almost all times, while providing the desired service-levels for inbound calls and maintaining fixed throughput rate of outbound calls.

For both control policies, in §§3 and 4, we have studied the transient behavior of the system as it approaches steady state. The unique steady state of a large system was shown to be well modeled by a simple fixed point, and the limiting approximations were shown to approach that stationary point exponentially fast.

Future Research. An immediate insight is that large blended service pools can respond quickly to bursts in the arrival process by automatically, and almost instantaneously, allocating more agents to inbound calls when this is needed. This suggests that call blending can be useful in nonstationary settings in which the arrival process of inbound work is time dependent, or even stochastic. In

these more complex environments, the relatively low variability associated with outbound work can help "smooth" the increased variability associated with the inbound work. It remains to study those more challenging settings, demonstrate the robustness of blending against nonstationarity, and find the appropriate order of safety staffing that is required.

Moreover, the reduction in staffing needs in a blended pool is due to the smaller number of agents that need to process inbound work relative to a split system. It stands to reason that some further reduction in staffing needs can be attributed to outbound calls as well. Of course, any further decrease in staffing must be small (if not negligible), but this point is yet to be studied. See also §B.1.

Finally, since empirical work suggests that service and patience times are often non-exponential, it will be useful to analyze corresponding non-Markovian models as well. Assuming Non-exponential service times is required if one wants to incorporate predictive dialing into the model, since the lack of memory of the exponential distribution implies that the age of an ongoing call (namely, the time that has passed since an ongoing service began) has no implications on the remaining service time of that call.

7. Proofs for the Results in §3

The proofs of the theorems employ supporting lemmas which are proved in §A. Recall that Assumptions 1, 2 and 3 hold, and that the first priority control is employed.

Additional Notation. let $\mathcal{D}_k(I) \equiv \mathcal{D}(I, \mathbb{R}_k)$ be the space of all right-continuous $\mathbb{R}_k$-valued functions on I with limits from the left everywhere, endowed with the familiar Skorohod J_1 topology, and let $\mathcal{C}_k(I)$ be the subset of continuous functions in $\mathcal{D}_k(I)$. If I is an arbitrary compact interval, we simply write $\mathcal{D}_k$ instead of $\mathcal{D}_k(I)$, and similarly for $C_k(I)$. We let d_{J_1} denote the J_1 metric on $\mathcal{D}_k(I)$. Since all our limit processes are continuous, convergence in the J_1 topology is equivalent to uniform convergence on compact intervals.

For two vectors $x, y \in \mathbb{R}_d$, $d \geq 1$, we write $x \leq y$ if the inequality holds componentwise. For two stochastic processes X and Y in $\mathcal{D}_d$ we write $X \leq_{st} Y$ if X is stochastically smaller than Y in sample-path stochastic order, i.e., if it is possible to construct two processes $\tilde{X}$ and $\tilde{Y}$, such that $\tilde{X} \stackrel{\mathrm{d}}{=} X$ and $\tilde{Y} \stackrel{\mathrm{d}}{=} Y$, and $P(\tilde{X} \leq \tilde{Y}) = 1$, where the inequality holds componentwise for all $t \geq 0$.

For a stochastic process Y in $\mathcal{D}$ and a deterministic sequence $\{a_n : n \geq 1\}$ of real numbers we say that Y is $O_P(a_n)$, and write $Y = O_P(a_n)$, if $\|Y\|_{a_n t}/a_n$ is *stochastically bounded*, i.e., if

$$\lim_{c\to\infty} \limsup_{n\to\infty} P\left(\|Y\|_{a_n t}/a_n > c\right) = 0.$$

We say that Y is $o_P(a_n)$ if $\|Y\|_{a_n t}/a_n$ converges in probability (and thus in distribution) to 0, i.e., if $\|Y\|_{a_n t}/a_n \Rightarrow 0 \quad \text{as } n \to \infty$. If $a_n = 1$ for all $n \geq 1$, then for a sequence of stochastic processes $\{Y^n : n \geq 1\}$, $Y^n = O_P(1)$ if Y^n is stochastically bounded, and $Y^n = o_P(1)$ if $\|Y^n\|_t \Rightarrow 0$ as $n \to \infty$.

To simplify notation, we assume, without loss of generality, that $N^n = n$. (We can normalize the parameters in Assumption 1 by dividing by N.)

7.1. Preliminary Results

Our results build on the following representation of the process X_3^n in (1), using independent unit-rate Poisson processes. The proof of the following lemma is identical to that of Lemma 2.1 in Pang et al. (2007) and is omitted.

LEMMA 1. *For each $n \geq 1$, with the priority control policy of idleness threshold, the stochastic process X_3^n is a well-defined random element of $\mathcal{D}_3$, and can be represented via*

$$\begin{aligned}
Q_1^n(t) &= Q_1^n(0) + N_1^a\left(\lambda_1^n \int_0^t \mathbf{1}_{\{I^n(s)=0\}} ds\right) - N_1^r\left(\theta_1 \int_0^t Q_1^n(s) ds\right) \\
&\quad - N_1^s\left(\mu_1 \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} Z_1^n(s) ds\right) - N_2^s\left(\mu_2 \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} Z_2^n(s) ds\right) \\
Z_1^n(t) &= Z_1^n(0) + N_2^a\left(\lambda_1^n \int_0^t \mathbf{1}_{\{I^n(s)>0\}} ds\right) + N_2^s\left(\mu_2 \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} Z_2^n(s) ds\right) \\
&\quad - N_3^s\left(\mu_1 \int_0^t \mathbf{1}_{\{Q_1^n(s)=0\}} Z_1^n(s) ds\right) \\
Z_2^n(t) &= Z_2^n(0) + N_3^s\left(\mu_1 \int_0^t \mathbf{1}_{\{I^n(s-)=K^n\}} Z_1^n(s) ds\right) - N_2^s\left(\mu_2 \int_0^t \mathbf{1}_{\{I^n(s-)<K^n\}} Z_2^n(s) ds\right),
\end{aligned} \tag{10}$$

where $I^n(t) = N^n - Z_1^n(t) - Z_2^n(t)$, $t \geq 0$, and N_1^a, N_2^a, N_1^r and N_j^s, $j = 1, 2, 3$, are mutually independent unit-rate Poisson processes.

The main difficulty in analyzing (10) is due to the indicator functions appearing in the integrants of the time-changed Poisson processes, e.g., $\mathbf{1}(Q_1^n(s) > 0)$ and $\mathbf{1}(I^n(s) = 0)$. No limit theorems can be proved directly, even if the initial condition in Assumption 2 is strengthened and is assumed to converge.

Recall that, according to the routing policy, $I^n \leq K^n = O(\log n)$ so that, $\bar{I}^n := I^n/n \Rightarrow 0e$ and $\hat{I}^n := I^n/\sqrt{n} \Rightarrow 0e$ as $n \to \infty$. Nevertheless, this does not imply that $\int_0^t \mathbf{1}(I^n(s) = 0) ds \Rightarrow 0e$ as $n \to \infty$. (In fact, highly-utilized queueing systems with a relatively small idleness tend to have long time periods in which all agents are busy.)

To study the system represented by the equations in (10), we will analyze the fluctuations of the number-in-system process about the threshold. Let $D^n := \{D^n(t) : t \geq 0\}$ denote the difference between the number of customers in the system and the threshold K^n, i.e.,

$$D^n(t) = Q_1^n(t) + K^n - I^n(t), \quad t \geq 0. \tag{11}$$

We refer to the process D^n as the "difference process".

Note that, since $Q_1^n(t) \cdot I^n(t) = 0$ under the routing policy, there are two options: If $Q_1^n(t) > 0$, then $I^n(t) = 0$ and $D^n(t)$ is equal to the number of customers **above** the threshold. If $Q_1^n(t) = 0$, then $I^n(t) \geq 0$ but, since $I^n(t) \leq K^n$ for all $t > 0$ due to our routing policy, $D^n(t)$ again measures the extra number of customers above the threshold.

The control ensures that the difference process is nonnegative and can thus be rephrased in terms of the difference process:

Alternative description of the control: If $D^n(t) > 0$, then a newly available agent at time t takes his next customer from the class-1 queue if there is a customer waiting, or else becomes idle. If at time t an agent finishes service and $D^n(t) = 0$, then one of the idle agents begins serving a class-2 customer.

Let

$$X_2^n := (Q_1^n, Z_1^n, N^n - Z_1^n). \tag{12}$$

Note that X_2^n is an essentially two-dimensional process, since the third component of X_2^n, namely $Z_2^n := N^n - Z_1^n$, is completely determined by Z_1^n. The next proposition is a simple SSC statement implying that, for large n, X_3^n is approximately equivalent to X_2^n. In particular, (Q_1^n, Z_1^n) carries sufficient information (asymptotically) to characterize the process X_3^n under appropriate scalings.

Proposition 1 *$d_{J_1}(X_3^n, X_2^n)/c_n \Rightarrow 0$ in $\mathcal{D}_3([\delta, \infty))$ as $n \to \infty$ for any $\delta > 0$ and any sequence $\{c_n : n \geq 1\}$ of positive real numbers satisfying $c_n / \log n \to \infty$ as $n \to \infty$.*

Proof. The proof is a simple consequence of the fact that there is an infinite supply of class-2 jobs, so that every available server can be assigned to a job at any time t. In general, the convergence might not hold at time 0 because the idleness may initially be of order larger than $O(c_n)$, i.e., it may hold that $I^n(0)/c_n \to \infty$ if the sequence $\{c_n : n \geq 1\}$ in the statement satisfies $c_n = o(n)$ as $n \to \infty$. However, due to the infinite supply of class-2 jobs and the assumption that an outbound call is successful immediately, we have that $I^n(t)/c_n \Rightarrow 0$ for any $t > 0$, because there can be at most $K^n = O(\log n)$ idle agents according to the routing policy. In particular, $I^n/c_n \Rightarrow 0e$ or, equivalently, $\bar{Z}_1^n + \bar{Z}_2^n \Rightarrow Ne$ in $\mathcal{D}([\delta, \infty))$ as $n \to \infty$, for any $\delta > 0$. Hence, the result follows. □

The analysis hinges on the following tightness of $\bar{X}_3^n$.

Lemma 2. (*tightness*) *The sequence $\{\bar{X}_3^n : n \geq 1\}$ is $\mathcal{C}$-tight in $\mathcal{D}_3$ with each limit being Lipschitz continuous and hence differentiable almost everywhere. As a consequence, $\bar{Y}^n := \{\bar{Q}_1^n + \bar{Z}_1^n : n \geq 1\}$ is also $\mathcal{C}$-tight in $\mathcal{D}$, with Lipschitz-continuous limits.*

The proof of Lemma 2 is similar to the proofs of Theorem 5.2 and Corollary 5.1 in Perry and Whitt (2013), and is omitted. We remark that Theorem 5.2 and Corollary 5.1 in Perry and Whitt (2013) assume that the sequence of fluid-scaled processes in the initial condition converge to a

proper random variable, however, in order to show the sequence of fluid-scaled processes $\bar{X}_3^n$ is tight, the tightness requirement for the initial condition is sufficient.

$\mathcal{C}$-Tightness of stochastic processes implies stochastic boundedness of these processes. We next identify an explicit stochastic bound for $\bar{X}_3^n$ which will be needed in the proof of Theorem 1.

To construct stochastic-order bounds for all $n \geq 1$, let

$$X_{bd}^n(t) := (Q_{bd}^n(t), N^n, N^n), \quad t \geq 0, \tag{13}$$

where Q_{bd}^n denotes the number-in-system process in an $M/M/\infty$ queue having arrival rate λ_1, service rate θ_1 and the initial condition $Q_{bd}^n(0) = \max\{na, Q_1^n(0)\}$, for a in Assumption 2. Note that for all $n > N_o$, for some N_o large enough, $\bar{Q}_{bd}^n(0) = a$ by Assumption 2 (where N_o is random). Hence, without loss of generality, we assume that $\bar{Q}_{bd}^n(0) = a$. We can represent Q_{bd}^n as

$$Q_{bd}^n(t) = na + N_1^a(\lambda_1 t) - N_1^r\left(\theta_1 \int_0^t Q_{bd}^n(s)\, ds\right), \quad t \geq 0, \tag{14}$$

where N_1^a and N_1^r are the Poisson processes from the representation of Q_1^n in (10).

The proofs of Lemmas 3 and 4 appear in §A.

LEMMA 3. (*boundedness*) *For each $n \geq 1$, $\bar{X}_3^n \leq \bar{X}_{bd}^n$ in $\mathcal{D}_3$ w.p.1 and in particular,*

$$Q_1^n(t) \leq Q_{bd}^n(t), \quad \text{w.p.1, for all } t \geq 0. \tag{15}$$

As a result, $P(\bar{X}_3^n(t) \leq x_{bd}(t), \ \ t \geq 0) \to 1$ as $n \to \infty$, where $x_{bd} := (q_{bd}, 1, 1)$ and, for a in Assumption 2,

$$q_{bd}(t) = \lambda_1/\theta_1 + (a - \lambda_1/\theta_1)e^{-\theta_1 t}. \tag{16}$$

Let $x_{1,2}(0)$ and $x_{1,2}^*$ denote the restrictions of $x(0)$ and x^* to the state space of $\mathcal{S}_{1,2} := [0,\infty) \times [0,1]$, respectively, corresponding to the first two components of $x(0)$ and x^*:

$$x_{1,2}(0) := (q_1(0), z_1(0)) \quad \text{and} \quad x_{1,2}^* := (0, \lambda_1/\mu_1). \tag{17}$$

Define $X_{1,2}^n := (Q_1^n, Z_1^n)$, the first two components of X_2^n in (12), and $\bar{X}_{1,2}^n := X_{1,2}^n/n$.

LEMMA 4. *For any $\epsilon > 0$, there exists T_ϵ, $0 < T_\epsilon < \infty$, such that $\|\bar{X}_{1,2}(t) - x_{1,2}^*\| < \epsilon$ w.p.1 for all $t > T_\epsilon$, for any limit process $\bar{X}_{1,2}$ of $\bar{X}_{1,2}^n$.*

Note that the time T_ϵ in Lemma 4 is uniform across *all the limits* of $\bar{X}_2^n$. We obtain the following corollary for the three-dimensional processes $\bar{X}_2^n$ in (12).

Corollary 2 *For any $\epsilon > 0$ there exists T_ϵ, $0 < T_\epsilon < \infty$, such that $\|\bar{X}_2(t) - x^*\| < \epsilon$ w.p.1, for all $t > T_\epsilon$, for any limit process $\bar{X}_2$ of $\bar{X}_2^n$ and for x^* in (4).*

Corollary 2 implies that $\|\bar{Q}_1^n\| < \epsilon$ in $\mathcal{D}([T_\epsilon, \infty))$ as $n \to \infty$, but the queue might still be strictly positive asymptotically under other other scalings. Hence, we cannot yet conclude that I^n is asymptotically positive with probability converging to 1 as in Theorem 2.

7.2. Proofs of Theorems 1 and 2

Proof of Theorem 1 For D^n in (11), let $T_K^n := \inf\{t \geq T_\epsilon : D^n(t) = 0\}$. Observe that T_K^n is the first time after T_ϵ, at which the threshold K^n is hit.

We start by showing that there exists $T < \infty$ such that

$$\lim_{n\to\infty} P(T_K^n \leq T) = 1. \tag{18}$$

For $t \geq 0$ and $X_3^n = (Q_1^n, Z_1^n, Z_2^n)$, let

$$\beta^n(X_3^n) = \mu_1 Z_1^n + \mu_2 Z_2^n + \theta_1 Q_1^n. \tag{19}$$

Similarly, for $t \geq 0$ and $\gamma := (q_1, z_1, z_2)$, let $\beta(\gamma) = \mu_1 z_1 + \mu_2 z_2 + \theta_1 q_1$.

Note that

$$\beta^n(X_3^n)/n \Rightarrow \beta(\bar{X}_3) \quad \text{in } \mathcal{D} \quad \text{as } n \to \infty, \tag{20}$$

where the second limit in (20) holds for the converging subsequence $\bar{X}_3^n$ to $\bar{X}_3$ in $\mathcal{D}_3$ and from the continuous mapping theorem (in particular, the continuity of linear functions at continuous limits). In addition, by Corollary 2 and the continuity of $\bar{X}_3$ and the function β, for any $\epsilon > 0$ we can find $T_\epsilon \in [0, \infty)$, for which

$$\lambda_1 < \beta(\bar{X}_3(t)), \quad t \geq T_\epsilon \quad w.p.1,$$

so that $\sup_{t\geq T_\epsilon} \beta(\bar{X}_3(t)) \geq \lambda_1$. Moreover

$$\|\bar{X}_3(t) - x^*\| \leq \epsilon \quad \text{for all } t \geq T_\epsilon. \tag{21}$$

Since $\beta(x^*) > \lambda_1$, by choosing $\epsilon < \beta(x^*) - \lambda_1$ sufficiently small, we can guarantee that $\sup_{t\geq T_\epsilon} \beta(\bar{X}_3(t)) > \lambda_1$. For later purposes, we will need to consider perturbation of these rates. Specifically, we take $\xi > 0$ small enough, such that

$$\alpha_+ := \lambda_1 + \xi, \quad \beta_+ := \sup_{t\geq T_\epsilon} \beta(\bar{X}_3(t)) - \xi \quad \text{and} \quad \alpha_+ < \beta_+ \quad w.p.1. \tag{22}$$

Consider an $M/M/1$ queue with arrival rate α_+ and service rate β_+ as in (22), and let Q_+ denote the queue-length process. By (22), Q_+ is an ergodic birth-death (BD) process.

Consider the process $Q_+^n := \{Q_+^n(X_3^n, t) : t \geq 0\}$ on $\{0, 1, 2, 3, \dots\}$ which, from any state $i \geq 0$, can jump one state up with rate α_+^n, and, from any state $i \geq 1$, can jump one state down with rate β_+^n, where

$$\alpha_+^n := \lambda_1^n \quad \text{and} \quad \beta_+^n := \sup_{t\geq T_\epsilon} \beta^n(X_3^n(t)) \tag{23}$$

for $\beta^n(X_3^n(t))$ given in (19). Then, conditional on X_3^n, the process Q_+^n is an $M/M/1$ queue process with the specified rates for each $n \geq 1$.

We need the following result, whose proof appears in §A.

LEMMA 5. *If* $D^n(T_\epsilon) = Q_+^n(\bar{X}_3^n, T_\epsilon)$, *then* $D^n \leq_{st} Q_+^n$ *in* $\mathcal{D}([T_\epsilon, \infty))$, *for* T_ϵ *in* (22).

Now observe that

$$\{Q_+^n(X_3^n, t) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{Q_+^n(\bar{X}_3^n, nt) : t \geq 0\}, \tag{24}$$

where $Q_+^n(\bar{X}_3^n, \cdot)$ is, conditional on $\bar{X}_3^n$, an $M/M/1$ queue with rates $\alpha_+^n(X_3^n)/n$ and $\beta_+^n(X_3^n)/n$, for α_+^n and β_+^n in (23).

It follows from the limits in (20) and our choice of the rates of Q_+ in (22), that $P(\alpha_+^n/n < \alpha_+ \text{ and } \beta_+^n/n > \beta_+) \to 1$ as $n \to \infty$, so that, if $Q_+^n(\bar{X}_3^n, 0) = Q_+(0)$, then for any a, $0 < a < 1$, we can find N_a, such that for all $n > N_a$,

$$P(Q_+^n(\bar{X}_3^n, nt) \leq Q_+(t),\ t \geq 0) > 1 - a. \tag{25}$$

Next, employing Lemma 3 and (15), we have that

$$D^n(T_\epsilon) \leq_{st} Q_{bd}^n(T_\epsilon) \quad \text{and} \quad \bar{Q}_{bd}^n(T_\epsilon) \Rightarrow q_{bd}(T_\epsilon) \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty, \tag{26}$$

where q_{bd} is the process in (16).

Let $Q_+(0) = D^n(T_\epsilon)$ and define $T_+^n := \inf\{t \geq T_\epsilon : Q_+(t) = 0\}$. It follows from Proposition 5.5, pp. 111 in Robert (2003) and from (26), that

$$\lim_{n\to\infty} P\left(\frac{T_+^n}{n} \leq T\right) = 1,$$

where $T := q_{bd}(T_\epsilon)/(\beta_+ - \lambda_1)$. The latter limit implies that $Q_+(nt)$ hits 0 in $O_P(1)$ time, and specifically, if $T^n := \inf\{t \geq 0 : Q_+(nt) = 0\}$ with $Q_+(0) = D^n(T_\epsilon)$, then

$$\lim_{n\to\infty} P(T^n \leq T) = 1. \tag{27}$$

Thus, for a in (25),

$$P(T_K^n \leq T) \geq P(T^n \leq T) - a \to 1 - a \quad \text{as } n \to \infty, \tag{28}$$

where the first inequality follows from Lemma 5 together with (24), (25) and (27). Taking $a \to 0$ gives (18).

Finally, by Proposition 5.11 in Robert (2003), the time for an ergodic $M/M/1$ queue starting near the origin to reach level n grows exponentially with n. In particular, the time it takes Q_+ to reach level n, given that $Q_+(0) = O_P(1)$, is of the order $(\beta_+/\alpha_+)^n$. Hence, the fluctuations of $\{Q_+(nt) : t \in I\}$, where $I \subset [T, \infty)$ is a compact interval, are of order $o_P(n)$. (For a more careful treatment, see the proof of Theorem 2 below.) We thus have that $\bar{Q}_1^n \Rightarrow 0$ in $\mathcal{D}([T, \infty))$ as $n \to \infty$ which, together with Proposition 1, concludes the proof. □

Proof of Theorem 2 The result follows from the bounding arguments leading to the proof of (28). It is sufficient to show that the fluctuations of $\{Q_+(nt) : t \geq 0\}$ cannot be larger asymptotically than $O_P(\log n)$. Let $M_+(t) := \sup_{0 \leq u \leq t} Q_+(u)$. Now, Q_+ is an ergodic $M/M/1$ queue, so that, by Theorem 6 in Anderson (1970) and the Example on pp. 112 in that reference, $M_+(nt) = O_P(\log n)$. Hence, the fluctuations of D^n above the threshold K^n are also $O_P(\log n)$. Note that having the threshold be $O(\log n)$ means that we can potentially have infinitely-many time points for which $D^n(t) \geq K^n$, unless we choose K large enough. We next show that for any choice of K, the statement in the theorem holds true.

Consider an interval $[t_1, t_2]$, with $T \leq t_1 < t_2 < \infty$ for T in the statement of the theorem, and let

$$\mathcal{A}^n := \{t \in [t_1, t_2] : I^n(t) = 0\} = \{t \in [t_1, t_2] : D^n(t) = K^n\}.$$

We claim that, since D^n is stochastically dominated by the ergodic time-accelerated $M/M/1$ process $Q_+(nt)$, $P(\mathcal{A}^n) \to 0$ as $n \to \infty$. To see this, let M be a positive constant, and let, $Q_+(0) = D^n(t_1)$. Then

$$\int_{t_1}^{t_2} 1_{\{D^n(s) \leq M\}}\, ds \geq_{st} \int_0^{t_2 - t_1} 1_{\{Q_+(ns) \leq M\}}\, ds = \frac{1}{n} \int_0^{n(t_2 - t_1)} 1_{\{Q_+(s) \leq M\}}\, ds \to (t_2 - t_1) P(Q_+(s) \leq M) \tag{29}$$

where the limit holds w.p.1 as $n \to \infty$ due to the ergodic theory of CTMC's. We thus have

$$\lim_{M \to \infty} \limsup_{n \to \infty} P\left(\int_{t_1}^{t_2} 1_{\{D^n(s) \leq M\}}\, ds \geq t_2 - t_1 \right) = 1.$$

On the other hand, the inequality in the other direction holds trivially w.p.1, i.e.,

$$\int_{t_1}^{t_2} 1_{\{D^n(s) \leq M\}}\, ds \leq t_2 - t_1 \quad \text{w.p.1 for all } n \geq 1.$$

By taking $M \to \infty$, the above inequality, together with (29), implies the claim that $P(\mathcal{A}^n) \to 0$ as $n \to \infty$.

In addition, by Proposition 5.5 in Robert (2003), the time until Q_+ hits level 0 after hitting any level that is $O(\log n)$ is proportional to $(\log n)/n$ so that, after the time acceleration by n, $Q_+(nt)$ will hit level 0 within $O_P((\log n)/n)$ time units, which implies the same for D^n. □

Remark 2. Note that the statement in Theorem 2 *does not* imply that the idleness process is always positive with probability converging to 1. Specifically, the theorem does not imply that $\inf_{t \in [s,u]} I^n(t) \Rightarrow 0$. As the proof of the theorem shows, we can potentially have many time points $t \geq T$, possibly infinitely many, for which there is no idleness, asymptotically. Nevertheless, the probability measure of all those points converges to 0 as $n \to \infty$, implying the statement of the theorem. Moreover, when a queue forms, it empties instantly as $n \to \infty$. Thus, in the limit, a queue can be positive on a set of points with probability zero over any finite subinterval of $[T, \infty)$.

7.3. Proofs of the Results in §3.3

The proof of Theorem 3 relies on showing that all fluid limits of the tight sequence $\{\bar{X}_3^n : n \geq 1\}$ have a unique stationary point, and that each of these fluid limits converge to that point as $t \to \infty$. Hence, we start by rigorously defining the stationary state.

DEFINITION 3. (fluid stationarity) We say that the fluid limit $x := \{x(t) : t \geq 0\}$ is stationary if $x(0) = x^*$ implies $x(t) = x^*$ for all $t \geq 0$. If such a point x^* exists, then it is called a stationary state for the fluid limit.

We will show below that x^* in (4) is the unique stationary state of all fluid limits of $\bar{X}_3^n$. The CTMC X_3^n is clearly irreducible and positive recurrent, and thus ergodic, i.e., it possesses a unique stationary distribution. Recall that $X^n(\infty)$ denotes a random variable having the stationary distribution (and limiting distribution) of the process X_3^n, for each $n \geq 1$.

LEMMA 6. (*stationary state*) x^* *in* (4) *is the unique stationary state of any fluid limit* $\bar{X}_3$ *of a converging subsequence* $\bar{X}_3^{n'}$. *Moreover,* $\bar{X}_3(t) \to x^*$ *w.p.1 as* $t \to \infty$ *for any initial condition* $\bar{X}_3(0)$.

The proof of Lemma 6 appears in §A, as the proofs of the other supporting lemmas in this section.

Proof of Theorem 3 To prove that $\bar{X}_3^n(\infty) \Rightarrow x^*$, note that Lemma 3 implies that $\{\bar{X}_3^n(\infty) : n \geq 1\}$ is stochastically bounded in $\mathbb{R}_3$. Recall that stochastic boundedness in $\mathbb{R}_d$ is equivalent to tightness in $\mathbb{R}_d$, $d \geq 1$; see, e.g., §5.2.1 in Pang et al. (2007). Consider a sequence $\{\bar{X}_3^n : n \geq 1\}$ initialized with $\bar{X}_3^n(0) \stackrel{\mathrm{d}}{=} \bar{X}_3^n(\infty)$ for each $n \geq 1$. With this initial condition, for every $n \geq 1$, the process $\bar{X}_3^n$ is strictly stationary. Moreover, $\{\bar{X}^n(0) : n \geq 1\}$ is tight in $\mathbb{R}_3$, because it is distributed the same as $\bar{X}^n(\infty)$ for each $n \geq 1$. Applying Lemma 2, we can consider a limit $\bar{X}$ of a converging subsequence of the stationary sequence $\{\bar{X}^n : n \geq 1\}$ which must itself be stationary. However, by Lemma 6, the unique stationary point is x^*, so that necessarily, $\bar{X}^n(0) \Rightarrow x^*$ as $n \to \infty$, implying the statement. □

Proof of Theorem 4 From Lemma 6 we know that x^* is globally asymptotically stable. In particular, if we initialize close enough to x^*, so that $\alpha(x^C) < \beta(x^C)$ for $\alpha(\cdot), \beta(\cdot)$ in (20), then any fluid limit $\bar{X}_3$ of $\bar{X}_3^n$ is ensured to remain in the ϵ-neighborhood of x^*. The only complication is due to the fact that $\bar{Q}_1^n$ may be positive (we only required that it is $o_P(n)$). However, the arguments in the proof of Lemma 5 show that, if the fluid limit of $\bar{X}_3^n$ is in an appropriate ϵ-neighborhood of x^*, then the time until D^n hits 0 is $o_P(1)$. (In the statement of the theorem we take a "C-neighborhood" because ϵ has the connotation of being a small number, which is not necessarily the case here.) Then, with T^n denote the hitting time of 0 by D^n (i.e., the time at which the number of busy agents equals K^n), for any $\delta > 0$ we can find n_δ large enough (which is random) so that for all $n > n_\delta$, $T^n < \delta$ w.p.1. We then apply the arguments of Lemma 6 with the deterministic initial condition x^C to conclude the statement of the theorem. □

Proof of Theorem 5 The proof is implied by the previous proofs. In particular, since $\bar{X}^n(0) \Rightarrow x^*$ in $\mathbb{R}_3$ as $n \to \infty$, and since x^* is the stationary state of the fluid limit in Theorem 4, we have that $\bar{X}_3^n \Rightarrow x^* e$ in $\mathcal{D}_3$ as $n \to \infty$. The proof of Theorem 2 showed that, in that case, $Q_1^n(t)$ is positive for at most a finite number of points on any finite interval, and that at those points, its size is $O_P(\log n)$. Hence, $\hat{Q}_1^n \Rightarrow 0e$ in $\mathcal{D}$ as $n \to \infty$. By the fact that $I^n(t) = 0$ for at most a finite number of points on any finite interval, the representation of Z_1^n in (10) is asymptotically equivalent to

$$\tilde{Z}_1^n(t) = Z_1^n(0) + N_2^a(\lambda_1^n t) - N_3^s\left(\mu_1 \int_0^t \tilde{Z}_1^n(s)ds\right), \quad t \geq 0,$$

which is known to converge to the stated diffusion limit, given the conditions of the theorem. See, e.g., Theorem 1.1 and Representation (12) in Pang et al. (2007). □

7.4. Estimating time T in Theorem 1

The proof of Theorem 1, together with the result stated in Theorem 4 can be used together to estimate time T in the following manner. For $\gamma := (q_1, z_1, z_2)$ let $\beta(\gamma) := \theta_1 q_1 + \mu_1 z_1 + \mu_2 z_2$, as in the proof of Theorem 1, and denote by $Q_\gamma := \{Q_\gamma(t) : t \geq 0\}$ the queue process of an $M/M/1$ with arrival rate λ_1 and service rate $\beta(\gamma)$. Define $\mathbb{A} := \{\gamma \in \mathbb{R}_3 : \beta(\gamma) > \lambda_1\}$. Then, if $\gamma \in \mathbb{A}$, Q_γ is ergodic.

Next, we treat $\bar{X}_3^n$ as a converging sequence with a fluid limit $x := (q_1, z_1, 1 - z_1)$ for which z_1 satisfies (5). Let $t_{\mathbb{A}} := \inf\{t \geq 0 : \bar{X}(t) \in \mathbb{A}\}$. We use (5) to estimate the time that the fluid hits the set $\mathbb{A}$. (Note that $t_{\mathbb{A}} = 0$ is possible.) It follows from the continuity of the arrival and service rates of Q_γ as functions of γ, as well as the continuity of x, that once x hits the set $\mathbb{A}$ it must remain there for some time interval. While in $\mathbb{A}$, the queue process q_1 drains quickly, because it has a strong drift toward zero. Specifically, q_1 satisfies the ODE

$$\dot{q}_1(t) = \lambda_1 - \mu_1 z_1(t) - \mu_2 z_2(t) - \theta_1 q_1(t), \quad x(t) \in \mathbb{A}$$

and it follows from the definition of $\mathbb{A}$, that $\dot{q}_1(t) < 0$, so that q_1 is decreasing.

Given (5) we can numerically solve and find the time t_0 at which q_1 hits zero. It then follows from the last paragraph of the proof of Theorem 1 that, almost immediately after that hitting time, the unscaled stochastic queue will hit zero as well (if n is large). More precisely, for any $\delta > 0$, $P(Q_1^n(u) = 0 \text{ for some } u \in (t, t + \delta)) \to 1$ as $n \to \infty$. Hence, T in Theorem 1 is approximately equal to t_0 just described. Since in practice, $\theta_1 q_1(t)$ is typically negligible relative to $\tilde{\beta}(x(t)) := \mu_1 z_1(t) + \mu_2 z_2(t)$, it can be ignored for a (not so rough) estimate of a the hitting time of 0 t_0, and thus of T.

It is easy to compute the first time t for which $\tilde{\beta}(x(t)) \geq \lambda_1$ using (5). In the numerical example in §5.2 this time is computed to be approximately 0.5. Our simulation experiments show that at about this time, the queue reaches its maximum value, and starts to deplete; see Figures 9 and

10. As we wrote above, we can use the ODE to compute numerically the time at which $q_1(t)$ hits zero to estimate time T. However, since the increase period of q_1 is so short, we can estimate that the additional time, after $t = 0.5$, until the queue depletes is short as well, without any numerical calculations. Indeed, in the two simulation examples in §5.2, T was smaller than 1.5 time units.

Appendix

A. Proofs of Supporting Lemmas

We now give the proofs of the supporting lemmas in §7.

Proof of Lemma 3 To see why (15) holds, recall the initial condition $Q_1^n(0) \le Q_{bd}^n(0)$, and the construction of Q_{bd}^n, which employs the same Poisson processes as in the construction of Q_1^n, making both processes be defined on the same probability space. Hence, we can ensure that whenever the two processes are equal, if a departure occurs in Q_{bd}^n, then an abandonment is generated in Q_1^n. Specifically, if $Q_1^n(t-) = Q_{bd}^n(t-)$ at a time $t > 0$ and a departure from Q_{bd}^n occurs at time t, then we can generate an abandonment from Q_1^n at the same time t. This implies $Q_1^n \le Q_{bd}^n$, as in (15).

It is well-known that if $\bar{Q}_{bd}^n(0) \Rightarrow a$ in $\mathbb{R}$, then $\bar{Q}_{bd}^n \Rightarrow q_{bd}$ in $\mathcal{D}$ as $n \to \infty$, for Q_{bd}^n in (14) and q_{bd} in (16); see, e.g., Theorem 3.6 in Pang et al. (2007). Since stochastic order is maintained in the limit, it follows that $\bar{Q}_1^n(t) \le q_{bd}(t)$, for q_{bd} in (16) for all $t \ge 0$ and for all n large enough. □

Proof of Lemma 4 We consider a converging subsequence $\bar{X}_{1,2}^{n'}$, which exists by Lemma 2: $\bar{X}_{1,2}^{n'} \Rightarrow \bar{X}_{1,2} := (\bar{Q}_1, \bar{Z}_1)$ in $\mathcal{D}_2$ as $n \to \infty$. Recall that $\bar{Y}^n = \bar{Q}_1^n + \bar{Z}_1^n$. By the continuous mapping theorem, $\bar{Y}^{n'} \Rightarrow \bar{Y}$ in $\mathcal{D}$ as $n \to \infty$, where

$$\bar{Y}(t) := \bar{Y}(0) + \lambda_1 t - \mu_1 \int_0^t \bar{Z}_1(s)ds - \theta_1 \int_0^t \bar{Q}_1(s)ds, \quad t \ge 0. \tag{30}$$

Moreover, by Lemma 2, $\bar{Y}$ is differentiable almost everywhere. Taking derivatives, we have

$$\bar{Y}'(t) = \lambda_1 - \mu_1 \bar{Z}_1(t) - \theta_1 \bar{Q}_1(t), \quad t \ge 0. \tag{31}$$

Define the shifted process $U(t) := \bar{X}_{1,2}(t) - x_{1,2}^*$, which is a shift (by the constant vector $x_{1,2}^*$) of the process $\bar{X}_{1,2}$. Clearly, if $U(t) \to 0e$, then $\bar{X}_{1,2}(t) \to x_{1,2}^*$ as $t \to \infty$. Writing $\bar{X}_{1,2}'(t) = f(\bar{X}_{1,2}(t))$ and $U'(t) = g(U(t))$, and noting that $U'(t) = \bar{X}_{1,2}'(t)$, we see that

$$\bar{X}_{1,2}'(t) = f(\bar{X}_{1,2}(t)) = f(U(t) + x_{1,2}^*) = g(U(t)) = U'(t).$$

For $y = (y_1, y_2) \in \mathbb{R}_2$, let $V(y) := y_1 + y_2$, and note that, for $\dot{V}(\bar{X}_{1,2}) := \nabla V \cdot \bar{X}_{1,2}'$,

$$\dot{V}(\bar{X}_{1,2}) := \nabla V \cdot \bar{X}_{1,2}' = \nabla V \cdot f(\bar{X}_{1,2}) = \bar{Y}',$$

where ∇V denotes the gradient of V, and $\nabla V \cdot \bar{X}_{1,2}$ is the usual inner product of vectors in $\mathbb{R}_2$. It follows that

$$\dot{V}(U(t)) := \nabla V \cdot g(U(t)) = \nabla V \cdot f(\bar{X}_{1,2}(t) + x_{1,2}^*) = -\mu_1 \bar{Z}_1(t) - \theta_1 \bar{Q}_1(t), \quad t \ge 0, \tag{32}$$

so that $\dot{V}(U) < 0$ whenever $U \neq 0e$. By the Barbashin-Krasovskii theorem (see, e.g., Theorem 4.2 in Khalil (2002)), $U(t) \to 0$ w.p.1 as $t \to \infty$, for any initial condition $U(0)$, which proves that $\bar{X}_{1,2}(t) \to x^*_{1,2}$ w.p.1 as $t \to \infty$, regardless of the initial condition $x_{1,2}(0)$, for $x_{1,2}(0)$ in (17). In particular, for a given limit $\bar{X}_{1,2}$ of $\bar{X}^n_{1,2}$ and for a given $\epsilon > 0$, we can find $T(\bar{X}_{1,2}, \epsilon)$, such that $\|\bar{X}_{1,2}(t) - x^*_{1,2}\| < \epsilon$ w.p.1, for all $t \geq T(\bar{X}_{1,2}, \epsilon)$ (where we are free to choose the norm in $\mathbb{R}_2$).

To finish the proof, we need to show that $T(\bar{X}_{1,2}, \epsilon)$ can be taken independently of $\bar{X}_{1,2}$. To that end, we consider the L_1 norm: For $y \in \mathbb{R}_2$, $\|y\|_{L_1} := |y_1| + |y_2|$. Since the state space of $\bar{X}_2$ is $\mathcal{S}_{1,2} = [0, \infty) \times [0, 1]$, it is easy to see that

$$V(y) = \|y\|_{L_1} \quad \text{and} \quad \dot{V}(y) \leq -(\theta_1 \wedge \mu_1)\|y\|_{L_1}, \quad y \in \mathcal{S}_{1,2},$$

where, for $a, b \in \mathbb{R}$, $a \wedge b := \min\{a, b\}$. It then follows from Theorem 3.4 on pp. 82 of Marquez (2003), that for every limit point $\bar{X}_2$, $\|\bar{X}_{1,2}(t)\|_{L_1} \leq \|\bar{X}_{1,2}(0) - x^*_{1,2}\|_{L_1} e^{-(\theta_1 \wedge \mu_1)t/2}$. Since $\bar{X}_{1,2}(0) = x_{1,2}(0)$ for $x_{1,2}(0)$ in (17), for all limits $\bar{X}_{1,2}$ by Assumption 2, the bound on the rate of convergence above is independent of the specific limit point $\bar{X}_{1,2}$, which implies the result. □

Proof of Lemma 5 By definition, we clearly have that the rate at which D^n jumps up from any state is the same as the rate at which Q^n_+ jumps up from all states, namely λ^n_1. Since the rates of both processes are determined by X^n_3, conditional on X^n_3, the rate at which D^n jumps down from each state is no smaller than the constant death rate of Q^n_+. The statement of the lemma then follows from the typical arguments, as in the proof of Lemma 3. □

Proof of Lemma 6 Consider a fluid limit $\bar{X}_3$ of a converging subsequence $\bar{X}^{n'}_3$. To prove that x^* is the unique stationary state of $\bar{X}_3$, we need to show that if $\bar{X}_3(0) = x^*$, then $\bar{X}_3(t) = x^*$ for all $t > 0$. It follows from Theorem 1 that if there exists a stationary state for $\bar{X}_3$, then it must be of the form $(0, z^*_1, 1 - z^*_1)$, so it is sufficient to show that $z_1(t) = \lambda_1/\mu_1$ whenever $z_1(0) = \lambda_1/\mu_1$.

To that end, assume that $\bar{X}^n_3(0) \Rightarrow (0, \lambda_1/\mu_1, 1 - \lambda_1/\mu_1)$ in $\mathbb{R}_3$ as $n \to \infty$. Consider the process $\bar{Y}^n = \bar{Q}^n_1 + \bar{Z}^n_1$ and note that, by Theorem 1, $d_{J_1}(\bar{Y}^n, \bar{Z}^n_1) \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$. The argument leading to $\bar{Y}^n \Rightarrow \bar{Y}$ in (30) implies that $\bar{Z}^n_1 \Rightarrow z_1$ in $\mathcal{D}$ as $n \to \infty$, where (recalling that $\bar{Q}^n_1 \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$)

$$z_1(t) = z_1(0) + \lambda_1 t - \mu_1 \int_0^t z_1(s)ds.$$

Note that when $z_1(0)$ is deterministic, the integral equation above has a deterministic solution, which is easily seen to be unique and of the form (5). Clearly, if $z_1(0) = \lambda_1/\mu_1$, then $z_1(t) = \lambda_1/\mu_1$ for all t. This proves that x^* is a stationary state of $\bar{X}_3$. Uniqueness follows from Lemma 4. In particular, if there was another stationary state $\tilde{x}^*$, then $\bar{X}_3(t) = \tilde{x}^*$ whenever $\bar{X}_3(0) = \tilde{x}^*$, by definition. However, by Lemma 4, $\bar{X}_3(t) \to x^*$ w.p.1 for any initial condition, so that $\bar{X}_3$ cannot be fixed at any state other than x^*. □

Remark 3. A careful read of the proofs shows that Assumption 2 can be weakened to assuming that $\bar{Q}^n_1$, and thus $\bar{X}^n_3$, is *stochastically bounded* in $\mathbb{R}$ (i.e., that it is $O_P(1)$), at the expense of complicating the stochastic-order bound arguments for $\bar{Q}^n_1$ in Lemma 3 and for $\bar{X}^n_{1,2}$ in Lemma 4.

B. Systems with Outbound Delay Times

In this appendix we consider a model which accounts for the time it takes an outbound customer to reply. We refer to this time as "outbound delay time", and show that the logarithmic safety staffing ensures that the number of dropped calls is negligible in large systems. Specifically, we assume that the outbound delay time is exponentially distributed with mean $1/M$, and that one of the waiting agents (see §2.1) will begin helping an inbound arrival if there are no idle agents left when an inbound customer arrives. We consider the threshold policy in Definition 1, in which an outbound call is initiated when an agent becomes idle, and there are K^n additional idle agents. Whenever an outbound call is initiated, one of the idle agents becomes a waiting agent, so that the idleness in the system is never larger than K^n.

Let $\widetilde{W}^n(t)$ denote the number of waiting agents at time t in system n, and let $\widetilde{Q}_1^n$ and $\widetilde{Z}_i^n$, $i=1,2$, denote the processes of the number of inbound customers in queue and number of class-i customers in service, respectively. Then

$$\widetilde{X}^n(t) = (\widetilde{Q}_1^n(t), \widetilde{Z}_1^n(t), \widetilde{Z}_2^n(t), \widetilde{W}^n(t)), \quad t \geq 0,$$

is a CTMC for each $n \geq 1$. We let

$$\widetilde{I}^n(t) := N^n - \widetilde{Z}_1^n(t) - \widetilde{Z}_2^n(t) - \widetilde{W}^n(t), \quad t \geq 0, \tag{33}$$

denote the idleness process, and note that an outbound call is dropped if it is replied at a time s at when there are no idle agents and no waiting agents, namely, if $\widetilde{I}^n(s) + \widetilde{W}^n(s) = 0$.

Let $\mu := \max\{\mu_1, \mu_2\}$. As was mentioned in §2.2, $\mu << M$ because the average service time $1/\mu$ is in the order of minutes, whereas $1/M$ is in the order of seconds. In a large system, the total number of events (arrivals, departures and abandonment) that occur during an average service time $1/\mu_i$, $i=1$ or $i=2$, is substantially larger than the total number of events occurring during an average outbound delay time $1/M$ For instance, if $1/\mu_2 = 5$ minutes in the example in §5 with 200 agents (so that $1/\mu_1 = 2.5$ minutes), then there is an order of 350 events during any time interval of length $1/\mu_1$, and an order of 700 events during intervals of length $1/\mu_2$, but an order of 20 events during time intervals of length $1/M = 9$ seconds (0.15 of a minute).

It is therefore not surprising that if we model $\widetilde{W}^n$ as being of the same order of size as the other three component processes in $\widetilde{X}^n$, so that

$$P(\widetilde{W}^n(t) > \xi) \to 1 \quad \text{as } n \to \infty \quad \text{for some } \xi > 0 \text{ and for all } t \geq T, \tag{34}$$

(i.e., $\widetilde{W}^n$ is strictly positive in fluid scale after some finite time $T \geq 0$), leads to inconsistencies when the asymptotic results are applied to a finite stochastic system.

Before elaborating on that point, we first observe that if (34) holds, then there are asymptotically no dropped calls after time T, because $P(\widetilde{Z}_1^n(t) + \widetilde{Z}_2^n(t) < N^n) \to 1$ as $n \to \infty$ for all $t \geq T$. On the other hand, if $\widetilde{W}^n = o_P(n)$, then we have an asymptotic SSC, namely $d_{J_1}(\widetilde{X}^n/n, (\bar{X}_3^n, 0)) \Rightarrow 0$ as $n \to \infty$, for X_3^n in (1), so that the proof of Theorem 1 (see (21)) implies that

$$\lim_{n\to\infty} P(\|\widetilde{X}^n/n - \tilde{x}^*\| \leq \epsilon) = 1 \quad \text{for all } t \geq T_\epsilon,$$

for any $\epsilon > 0$ and for T_ϵ in Lemma 4, where $\tilde{x}^* := (x^*, 0)$ and x^* is defined in (4). We can therefore apply the arguments in the proof of Theorem 1, and the stochastic-order bound on the fluctuations of the number-in-system process in Lemma 5, to conclude that Theorem 2 holds for $\widetilde{X}^n$ as well.

B.1. Fluid Analysis of $\widetilde{W}^n$

As in (10), we use random time changed Poisson processes to represent $\widetilde{X}^n$. The representation of $\widetilde{W}^n$ is then

$$\widetilde{W}^n(t) = \widetilde{W}^n(0) + N_1^w\left(\mu_1\int_0^t \mathbf{1}_{\{\widetilde{I}^n(s-)=K^n\}}\widetilde{Z}_1^n(s)\ ds\right) + N_2^w\left(\mu_2\int_0^t \mathbf{1}_{\{\widetilde{I}^n(s-)=K^n\}}\widetilde{Z}_2^n(s)\ ds\right)$$
$$- N_3^w\left(M\int_0^t \widetilde{W}^n(s)\ ds\right) - N_1^a\left(\lambda_1^n\int_0^t \mathbf{1}_{\left\{\{\widetilde{W}^n(s-)>0\}\cap\{\widetilde{I}^n(s-)=0\}\right\}}\ ds\right),$$

where N_1^a and N_i^w, $i=1,2,3$, are independent unit-rate Poisson processes. Just as in Lemma 2, $\widetilde{X}^n/N^n$ can be shown to be $\mathcal{C}$-tight in $\mathcal{D}_3$, with each limit being differentiable almost everywhere.

Observe that, conditional on $\widetilde{X}^n(t)$, the instantaneous increase rate of $\widetilde{W}^n(t)$ at time t is

$$\mu_1\mathbf{1}_{\{\widetilde{I}^n(t)=K^n\}}\widetilde{Z}_1^n(t) + \mu_2\mathbf{1}_{\{\widetilde{I}^n(t)=K^n\}}\widetilde{Z}_2^n(t).$$

In particular, the instantaneous increase rate of $\widetilde{W}^n(t)$ is strictly smaller than μN^n at any time $t\geq 0$.

The instantaneous decrease rate of $\widetilde{W}^n(t)$ at time t is, conditional on $\widetilde{X}^n(t)$,

$$M\widetilde{W}^n(t) + \lambda_1^n\mathbf{1}_{\left\{\{\widetilde{W}^n(t)>0\}\cap\{\widetilde{I}^n(t)=0\}\right\}},$$

implying that any limit of $\widetilde{W}^n/N^n$ will be a solution to an ODE of the form

$$\dot{w}(t) = \mu_1\pi_1(t)\tilde{z}_1(t) + \mu_2\pi_1(t)\tilde{z}_2(t) - Mw(t) - \lambda_1\pi_2(t), \quad t\geq 0, \tag{35}$$

where $\pi_i(\cdot)$ is some function satisfying $0\leq\pi_i(t)\leq 1$, for all $t\geq 0$, $i=1,2$. (We use $\tilde{z}_i$ to denote the fluid limit of $\widetilde{Z}_i^n/N^n$, and similarly for all other processes in $\widetilde{X}^n$.) Therefore all fluid limits of $\widetilde{W}^n/N^n$ satisfy

$$\dot{w}(t) < \mu - Mw(t),$$

where, since we scaled the prelimit sequence by N^n, the pool size in the fluid model equals 1, so that $\tilde{z}_1(t)+\tilde{z}_2(t)\leq 1$ for all $t\geq 0$. It follows that any fluid limit $\{w(t): t\geq 0\}$ of $\widetilde{W}^n/N^n$ with an initial condition $w(0)\in[0,1]$ is strictly bounded from above by the exponential function

$$y(t) = \mu/M + (w(0)-\mu/M)e^{-Mt} \to \mu/M \quad \text{as } t\to\infty.$$

The convergence rate to μ/M is especially fast in this case due to M being large, *regardless of the values of the other processes in the fluid limit* $\tilde{x}$ under consideration.

In the example considered in §2.1 with $1/\mu = 5$ minutes and $1/M = 9$ seconds, we get $w^* < \mu/M = 0.03$, where w^* is the stationary point $w(t)$. Since our computations employed rough upper bounds, the mean number of waiting agents in stationarity will be *strictly smaller* than 3% of the total number of agents. (It can be shown that w^* in this example is 0.02.) In practice, w^* will be even smaller than $K^n = O(\log n)$, which is inconsistent with $\widetilde{W}^n$ being $O(n)$.

Acknowledgements

The authors are grateful to the review team for their careful review and for many helpful comments which helped improve the paper.

References

[1] Aksin, Z., M. Armony, and V. Mehrotra. (2007) The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management* **16** (6), 665–688.

[2] Anderson, C.W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability* **7**, 99–113.

[3] Armony, M. and Gurvich, I. (2010) When promotions meet operations: cross selling and its effect on call-center performance. *Manufacturing and Service Operations Management* **12** (3), 470–488.

[4] Armony, M. and C. Maglaras. (2004a) On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *Operations Research* **52** (2), 271–292.

[5] Armony, M. and C. Maglaras. (2004b) Call centers with a call-back option and real-time delay information. *Operations Research* **52** (4), 527–545.

[6] Bassamboo, A., Randhawa, R.S. and Van Mieghem, J.A. (2009) A little flexibility is all you need: on the asymptotic value of flexible capacity in parallel queueing systems. To appear in *Operations Research*.

[7] Bassamboo, A., Randhawa, R.S. and Zeevi, A. (2010) Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited. *Operations Research* **56** (10), 1668–1686.

[8] Ben-Chanoch, E. (2004) Outbound calling system in a contact center. *US Patent 6707906*.

[9] Bhulai, S. and G. Koole. (2003) A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control* **48** (8), 1434–1438.

[10] Deslauriers, A., P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A. N. Avramidis. (2007) Markov chain models of a telephone call center with call blending. *Computers & Operations Research* **34**, 1616–1645.

[11] Dumas, G., Perkins, M. M., White, C. M. (1996) Call Sharing for inbound and outbound call center agents. *US patent 5,519,773*.

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

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

[14] Gans, N. and Y-P. Zhou. (2003) A call-routing problem with service-level constraints. *Operations Research* **51** (2), 255–271.

[15] Garnet, O., Mandelbaum, A. and Reiman, M. (2002) Designing a call center with impatient customers. *Manuf. Serv. Oper. Mgmnt.* **4** (3), 208–227.

[16] Gurvich I., Armony, M. and Maglaras, C. (2009) Cross-Selling in a Call Center with a Heterogeneous Customer Population. *Operations Research* **57** (2), 299–313.

[17] Gurvich I., Junfei H. and Mandelbaum A. (2012) Excursion-based universal approximations for the Erlang-A queue in steady-state. *working paper*.

[18] Gurvich, I and Perry, O. (2012) Overflow networks: approximations and implications to call-center outsourcing. *Operations Research* **60** (4), 996-1009.

[19] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** (3), 567–588.

[20] Iglehart, D.L. (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability* **2** (2), 429–441.

[21] Khalil, H. K. (2002). *Nonlinear Systems.* Prentice Hall, New Jersey.

[22] Mandelbaum, A. and Zeltyn, S. (2009). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research* **57** (5), 1189–1205.

[23] Marquez, H. J. (2003). *Nonlinear Control Systems.* Wiley, New York.

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

[25] Perry, O. and Whitt, W. (2011). A Fluid Approximation for Service Systems Responding to Unexpected Overloads. *Operations Research* **59** (5), 1159–1170.

[26] Perry, O. and Whitt, W. (2013). A fluid limit for an overloaded X model via a stochastic averaging principle. *Mathematics of Operations Research* 38 (2), 294–349.

[27] Perry, O. and Whitt, W. (2014) Diffusion approximation for an overloaded X model via a stochastic averaging principle. To appear in *Queueing Systems.*

[28] Robert, P. (2003). *Stochastic networks and queues.* Springer-Verlag, Berlin.

[29] Reynolds, P. (2010). Basics of staffing for outbound calling. In *www.thecallcenterschool.com*

[30] Samuelson, D.A. (1999). Predictive dialing for outbound telephone call centers. *Interfaces* **29** (5), 66–81.

[31] Tsitsiklis, J.N. and Xu, K. (2012) On the power of (even a little) resource pooling. *Stochastic Systems* **2**, 1–66.

[32] 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

[33] Whitt, W. (1992) Understanding the efficiency of multi-server service systems. *Management Science* **38** (5), 708–723.

[34] Whitt, W. (1999) Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* **24**, 205–212.

[35] Whitt, W. (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50** (10), 1449–1461.