Designing a Call Center with Impatient Customers

O. Garnett • A. Mandelbaum • M. Reiman
Davidson Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel
Davidson Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel
Bell Laboratories, Murray Hill, New Jersey 07974
oferpbg@zahav.net.il • avim@tx.technion.ac.il • marty@research.bell-labs.com

The most common model to support workforce management of telephone call centers is the $M/M/N/B$ model, in particular its special cases $M/M/N$ (Erlang C, which models out busy signals) and $M/M/N/N$ (Erlang B, disallowing waiting). All of these models lack a central prevalent feature, namely, that impatient customers might decide to leave (abandon) before their service begins.

In this paper, we analyze the simplest abandonment model, in which customers' patience is exponentially distributed and the system's waiting capacity is unlimited ($M/M/N + M$). Such a model is both rich and analyzable enough to provide information that is practically important for call-center managers. We first outline a method for exact analysis of the $M/M/N + M$ model, that while numerically tractable is not very insightful. We then proceed with an asymptotic analysis of the $M/M/N + M$ model, in a regime that is appropriate for large call centers (many agents, high efficiency, high service level). Guided by the asymptotic behavior, we derive approximations for performance measures and propose ''rules of thumb'' for the design of large call centers. We thus add support to the growing acknowledgment that insights from diffusion approximations are directly applicable to management practice.

(*Tele-Queues; Erlang C; Erlang A; Telephone Call and Contact Centers; Multiserver Exponential Queues; Workforce Management or Staffing; Queues with Abandonment; Diffusion Approximation*)

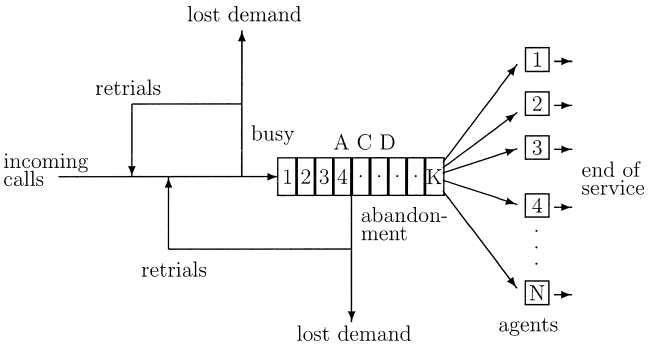
1. Introduction and Summary

During recent decades there has been explosive growth in the number of companies that provide services via the telephone, and also in the variety of telephone services provided. A central challenge in designing and managing any service operation is to achieve a balance between operational efficiency and service quality, and in telephone services this challenge is often pushed to the extreme: A large call center serves thousands of calls per day, each of which demands a response within seconds. Analytical models provide guidance regarding the sought-after balance.

Modeling a Call Center. A simplified representation of call-center flows is given in Figure 1. Incoming calls form a single queue to wait for service from one of N statistically identical agents. $K + N$ telephone trunks are connected to an Automatic Call Distributor (ACD) that manages the queue, connects customers to available agents, and archives operational data. Customers arriving when all trunks are occupied encounter a busy signal. Such customers might try again later (''retrial'') or give up (''lost demand''). Customers who succeed in getting through at a time when all agents are busy (that is, when there are at least N but fewer than $K + N$ customers within the call center), are placed in the queue. If waiting customers run out of patience before their service begins,

1523-4614/02/0403/0208$05.00
1526-5498 electronic ISSN

Figure 1 Schematic Representation of a Telephone Call Center



they hang up (''abandon''). After abandoning, customers might try calling again later.

In basic models of call centers it is commonly assumed that the only parameters under the system manager's control are the number of trunks available ($K + N$) and the number of agents (N). In most contexts the cost of trunk lines is trivial compared to personnel costs, so in this paper we focus on staffing decisions (N), assuming that $K = \infty$ for modeling purposes. Thus busy signals are absent in the models to be considered, though, we would like to emphasize that in no way do we advocate here the practice of unconditional ''no busy signal''. Indeed, a trade-off between busy signals and abandonment, in the spirit of Borst et al. (2000), is a worthwhile direction for future research.

The classical $M/M/N$ queueing model, also called the Erlang-C model, is obtained by further assuming Poisson arrivals, exponentially distributed service times, and no abandonment. It is the model most often used in call-center analysis, but it has one glaring defect: Call abandonment is not a negligible or minor aspect of call-center operations. This issue is broadly addressed in §2.

The Square Root Rule for Safety Staffing. In this paper, we consider the $M/M/N$ model with abandonment added. To summarize the central findings, it will be useful to first review an important principle regarding capacity choice *in the absence of abandonment.* Let $R = \lambda/\mu$ denote the (average) *offered load,* where λ is the average call arrival rate and $1/\mu$ is the mean call duration. (R is measured in units of service duration per unit of time.) The principle is as follows: For moderate to large values of R (or equivalently moderate to large λ, since we are assuming that μ is a fixed parameter of the call center), the appropriate staffing level is

$$N = R + \beta\sqrt{R}, \tag{1}$$

where β is a positive constant that depends on the desired level of service. Of course, in practice the value of N derived from this formula must be rounded to an integer.

The second term on the right side of (1) may be described as the *excess capacity* needed, beyond nominal requirements (the first term), to achieve the target service level in the face of stochastic variability. Equation (1) shows that the required excess capacity grows less than proportionately with the load of calls to be handled. This phenomenon is aptly described as *statistical economies of scale.*

The square root formula (1) was derived and discussed by Ward Whitt (1992), but as he explained, similar design rules had been advanced by a number of other authors during the 1970s and 1980s. (We refer the reader to Borst et al. (2000) for a historical perspective.) Whitt's treatment of this subject was based primarily, but not exclusively, on his pioneering work with Shlomo Halfin (Halfin and Whitt 1981) regarding diffusion approximations for many-server queues. To be more specific, the foundation of Whitt's argument is the following: If one considers a variety of systems with different moderate to large values of R, and if the number of agents N is chosen according to (1) in each case, then the quality of service will be approximately the same in each system. The measure of service quality underlying this statement is the steady-state probability that a caller must wait in the queue before service. This probability is commonly referred to as the Erlang-C formula, hereafter abbreviated as $P\{W > 0\}$. Halfin and Whitt (1981) provided a formula for computing β in terms of the target value for $P\{W > 0\}$.

The Square Root Rule with Abandonment. Enriching the $M/M/N$ model to include abandonment, we assume the following: There is associated with each arriving caller an exponentially distributed ran-

dom variable that quantifies the individual's *patience*; if a caller's waiting time in the queue grows to equal his or her patience, then the call is abandoned. (In the interest of tractability, we assume that customers who abandon do not retry.) The patience variables characterizing different callers are independent and identically distributed with mean θ^{-1}, and they are independent of all other model elements as well. The positive quantity θ will be referred to as either the *abandonment rate* or the *impatience parameter,* depending on context. For this model, we use the notation $M/M/N + M$, as introduced by Baccelli and Hebuterne (1981), and we propose to refer to it as Erlang-A (A for Abandonment, and for the fact that it interpolates between Erlang-C and Erlang-B, the latter being $M/M/N/N$).

It will be shown in this paper that the square root rule (1) remains valid in the model with abandonment for moderate to large R. Of course, the formula for β is different in our context: β now depends on both the abandonment rate θ and the target value for $P\{W > 0\}$, and in our setting β may be negative, even when a small probability of waiting is specified. That is, to achieve a given target value for $P\{W > 0\}$, it may be sufficient to take N *smaller* than the offered load R.

Furthermore, the appropriate value for β in formula (1) is monotonically *decreasing* in θ for fixed $P\{W > 0\}$. That is, a higher abandonment rate *reduces* the amount of capacity one needs to achieve a given ''service level''. This fact may be surprising initially, but the following makes it obvious: Temporarily denoting by Q the steady-state number of callers in the system, we have $P\{W > 0\} = P\{Q \geq N\}$, and Q decreases stochastically as one increases θ. Of course, $P\{W > 0\}$ is not the only measure one could use to quantify the notion of ''service level'' or ''service quality'', but the discussion immediately below shows that performance according to other obvious measures is uniformly excellent when R is not small and the square root rule is used for staffing.

Three Staffing Regimes. Another fundamental measure of service quality is the steady-state probability that an arrival will abandon before getting service, hereafter abbreviated as $P\{Ab\}$. Table 1 summarizes the behavior of $P\{W > 0\}$ and $P\{Ab\}$ in three limiting regimes studied later in the paper, each of which represents a different philosophy with regard to the design of a call center. (The limit referred to here is $R \to \infty$.)

Table 1 System Performance in Three Staffing Regimes

Regime	Staffing Level	Performance Characteristics
Rationalized	$N = R + \beta\sqrt{R}$	$P\{W > 0\} \to \alpha(\beta)$ and $P\{Ab\} \to 0$
Quality-Driven	$N = R + \epsilon R, \quad \epsilon > 0$	$P\{W > 0\} \to 0$ and $P\{Ab\} \to 0$
Efficiency-Driven	$N = R - \epsilon R, \quad \epsilon > 0$	$P\{W > 0\} \to 1$ and $P\{Ab\} \to \epsilon$

The ''rationalized regime'' is that where the square root rule (1) is used to determine system capacity: A formula will be derived for the limiting $P\{W > 0\}$, denoted by $\alpha(\beta)$ in Table 1 (the formula appears below Table 4 in §5), and it will be shown that $P\{Ab\} \to 0$ as $R \to \infty$, regardless of how β is chosen. The ''quality-driven'' regime is where capacity exceeds nominal requirements by a fixed percentage: It will be shown that both $P\{W > 0\}$ and $P\{Ab\}$ vanish as $R \to \infty$. Finally, in the ''efficiency-driven'' regime, capacity *falls short* of nominal requirements by a fixed percentage: It will be shown that virtually all arrivals wait in this case, but $P\{Ab\}$ is simply equal to the capacity shortfall, expressed as a fraction of the offered load.

Another interesting performance measure is the steady-state average waiting time before either service begins or the caller abandons, hereafter abbreviated $E[W]$. One has the useful identify $\theta \cdot E[W] = P\{Ab\}$, so results cited for $P\{Ab\}$ in Table 1 translate immediately into results for $E[W]$. Even in the ''efficiency-driven'' staffing regime where virtually all callers wait, both $P\{Ab\}$ and $E[W]$ remain small if the capacity shortfall is small.

Actually, the results proved later about the three staffing regimes are both more refined and more extensive than the summary provided in Table 1, but this summary communicates the most important findings for purposes of system design. Based on this analysis, we conclude that the rationalized regime is appropriate in most settings: By choosing the con-

stant β appropriately, a system manager using the square root rule (1) can achieve a rational balance between efficiency and service quality, unless one of those two concerns utterly dominates the other. Determination of β, based on economic considerations via (asymptotic) optimization, is an important topic under current research. (See Borst et al. 2000 for $M/M/N$ dimensioning analysis.)

The results reported in Table 1 support the following important, if unsurprising, conclusion: With impatient customers and a moderate to large call volume, system performance is relatively robust; any staffing level close to the nominal requirement R produces ''good'' service. Callers who refuse to wait indefinitely impose smaller externalities on later arrivals than do patient callers, and generally speaking, they make the system designer's job easier.

Contributions to System Modeling. From a mathematical standpoint, the classical $M/M/N$ model is fundamentally changed when one incorporates the phenomenon of abandonment. For example, the model with abandonment is stable for all parameter combinations, whereas the classical model achieves statistical equilibrium if and only if $R < N$. Also, as we show by numerical examples, the models with and without abandonment tend to give very different performance estimates in the parameter regime of primary interest, even when the abandonment rate θ is small. Because the $M/M/N$ model is so commonly used for quick-and-dirty performance analysis, these effects of adding abandonment are thoroughly discussed and illustrated in this paper (see §2).

Denoting by $Q(t)$ the number of callers present in the system at time t, either waiting or being served, we focus on the stochastic process $Q = \{Q(t), t \geq 0\}$. This process is central in call centers, as will now be explained. First, it is a visible cue for management and agents and its realtime value is often displayed for everyone to see. Moreover, in call centers that provide toll free services, Q is proportional to the cost of incoming calls. However, the customer's point of view is represented by the *waiting time*—either *potential* (i.e. the time he would wait in queue for his service to commence if his patience was infinite) or *actual.* Indeed, in §3 we introduce a method for calculating a wide variety of performance measures that involve the potential and actual waiting times. An important outcome of our analysis, as shown in Theorem 3 below, is that potential waiting time and queue length are deterministically related in heavy traffic.

Given the various assumptions laid out earlier, Q is a birth-and-death process, and so its steady-state distribution can be written out in an ''explicit'' formula. However, that formula is complicated enough to cause difficulties both in numerical evaluation and in qualitative understanding. After some brief remarks about numerical evaluation of ''exact'' formulas for steady-state performance measures, most of this paper deals with approximations in the ''heavy traffic'' regime, where R is large and R/N is near 1. That is, we develop approximations for call centers having a moderate to large number of agents and high agent utilization.

Rather than developing approximations only for steady-state quantities, we show that a properly scaled version of the stochastic process Q is well approximated by a certain diffusion process in the heavy traffic regime. This analysis parallels the diffusion approximation developed by Halfin and Whitt (1981) for many-server models without abandonment. It helps to explain the steady-state approximations that spawn the square root formula (1), and gives a more complete understanding of system behavior in the heavy traffic regime.

Related Research. As attributed in the sequel, some of our results are motivated or based on previous work by Palm (1937, 1943, 1953), Riordan (1962), Baccelli and Hebuterne (1981), and especially Halfin and Whitt (1981), and Fleming et al. (1994). Indeed, both Halfin and Whitt (discussed earlier in this section) and Fleming et al. focus on heavy traffic analysis of $M/M/N$ and $M/M/N + M$ systems respectively. In Fleming et al. a diffusion approximation of the queue process is derived leading to approximations for the fraction of customers abandoning and other performance measures.

A general overview of models with abandonment appears in Boxma and de Waal (1994), with a review of relevant literature. There have been attempts to analyze more complex models, of which we mention a

few: Ancker and Gafarian (1963) analyze queues with abandonment, multiple heterogeneous agents, and finite capacity through their steady-state equations and derive the waiting time density. Sze (1984) compares different approximations for an $M/PH/N + PH$ model (*PH* stands for Phase Type distribution) with retrials, priorities, and nonstationary arrivals. The results are verified by simulation.

In Harris et al. (1987) and Hoffman and Harris (1986) the basic model is $M/M/N$, with the addition of abandonment, retrials, and a variety of service disciplines. Assuming a heavily-loaded call center and using some approximations, they arrive at a system of steady-state equations that can be solved numerically. In two recent papers by Brandt and Brandt (1998, 1999), $M/M/N + G$ models with state-dependent arrivals are analyzed. Applications include systems with an integrated voice-mail-server and cases in which idle agents initiate outbound calls. Finally, fluid and diffusion approximations for time-dependent models with abandonment and retrials, are described in Mandelbaum et al. (1999, 2000), both of which are based on Mandelbaum et al. (1998).

Overall Contribution and Contents. The contributions of the present paper, in our opinion, are both theoretical and practical, but even more so the bridging of the two. Specifically:

- Extending the fundamental findings of Halfin and Whitt (1981) to accommodate abandonment (for example, Theorems 4 and 2, Table 4) and waiting times (Theorem 3).
- Revisiting the classical Erlang (1909, 1917) and Palm (1943) results, and adapting them to the environment of the modern call center (§§3, 5.1, and 5.2).
- Adding support to the growing acknowledgment that insights from diffusion approximations are directly applicable to management practice (§§5.2, 5.3, Appendix A).
- Prepare the necessary ground for an economic analysis of abandonment, following Borst et al. (2000).

The rest of the paper is organized as follows: In §2 we provide further motivation as to the relevance of our results and the $M/M/N + M$ model in general, to the management of modern call centers. In §3 we outline a method for exact calculations of a wide variety of performance measures. Then in §4 we focus on heavy traffic limit theorems, which lead to some implementations discussed in §5. The paper also has three appendices: Appendix A displays graphs showing the (excellent) quality of a number of approximations derived in §5; Appendix B includes computational details of the method outlined in §3; and Appendix C contains proofs of Theorems 1–4 with an extended version of Theorem 2.

2. The Significance of Abandonment in Practice and Modeling

A major drawback of models that ignore abandonment is that they either distort or fail to provide information that is important to call-center managers. When trying to manage a large call center in heavy traffic, one must consider the effect of abandoning customers on service level. It is not enough to consider waiting times or busy signals, especially since abandonment statistics constitute the only ACD data that unveils customers' perception of service quality.

According to the Help Desk and Customer Support Practices Report (1997), more than 40% of call centers set a target for fraction of abandonment, but in most cases this target is not achieved. Moreover, the lack of understanding of the abandonment phenomenon and the scarcity of models that acknowledge it, has led practitioners to ignore it altogether. (For example, this lead Cleveland and Mayben (1997) to conclude that abandonment is ''not a good indicator of call-center performance''.) This can cause either under- or over-staffing: On the one hand, if service level is measured only for those customers who reach service, the result is unjustly optimistic—the immediate effect of an abandonment is less delay for those further back in line, as well as for future arrivals. This would lead to under-staffing. On the other hand, using workforce management tools that ignore abandonment would result in over-staffing as actually fewer agents are needed in order to meet most abandonment-ignorant service goals.

Figure 2 **Fraction Queueing—$M(48)/M(1)/N$ (a) vs. $M(48)/M(1)/N + M(0.5)$ (b)**

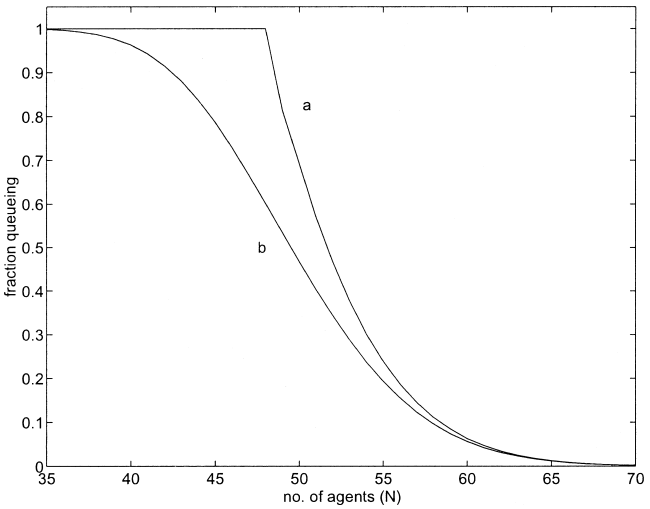


Table 2 **Comparing Results for Models With/Without Abandonment**

	$M/M/N$	$M/M/N + M$
Fraction abandoning	—	3.1%
Average speed of answer	20.8 sec.	3.6 sec.
Waiting time's 90th percentile	58.1 sec.	12.5 sec.
Average queue percentile	17	3
Agents' utilization	96%	93%

Note. 50 agents, 48 calls per minute, 1 minute average service time, 2 minute average patience. The values were calculated, and can be verified, using the "iProfiler" tool available at ⟨www.4callcenters.com⟩. This analysis tool is based on parts of the present work—in particular §3 and Appendix B.

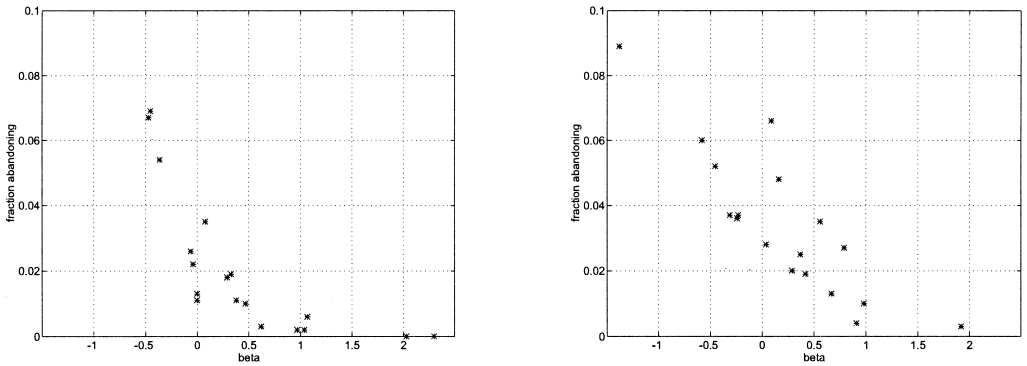
The significance of abandonment can be seen in simple numerical examples. Figure 2 shows graphs of the fraction of customers queueing according to the $M/M/N$ and $M/M/N + M$ models. It is clear that these graphs convey a rather different picture of what is happening in the system they depict, in particular, in the range of 40 to 50 agents.

Table 2 displays some results from an $M/M/N$ model and a corresponding $M/M/N + M$ model with only 3% abandonment. There is a significant difference in the distributions of waiting time and queue length—in particular, the average wait and queue length are both strikingly shorter when abandonment is taken into account. It is important to realize that such good performance is *not* achieved if the arrival rate to the $M/M/N$ system decreases by 3% (for example, the "average speed of answer" in such case is approximately nine seconds). We note, however, that the performance of systems in such heavy traffic is very sensitive to the staffing level—adding three or four agents to the model without abandonment would result in performance similar to that displayed for the model with abandonment (the horizontal distance between the graphs in Figure 2 shows this "margin"). Nonetheless, since personnel costs are the major expense of call centers (prevalent estimates run at about 60–70% of total cost), even a 6–8% reduction in personnel is significant.

Both Figure 2 and Table 2 clearly indicate that it is possible, while operating in heavy traffic, to simultaneously achieve high efficiency (agent utilization near 100%) and good service (low, but nonnegligible abandonment rate and waiting time). This is a direct outcome of economies of scale, as demonstrated in the next example.

The following is a possible scenario in which our staffing rules can be used (this scenario is revisited later in §5.3)—A given call center with N agents, service rate μ, and arrival rate λ (the offered load $R = \lambda/\mu$) , has "service grade" β (high values of β correspond to high service levels). There is a forecast of a higher arrival rate $\hat{\lambda}(\hat{R} = \hat{\lambda}/\mu)$ during a forthcoming holiday. The call center's manager wishes to maintain the present service level at the call center during the holiday, and hence needs to decide on an increase in the number of agents $\hat{N} = \hat{N}(\beta)$ for the holiday shifts. To this end, the manager must first determine the operational regime of the call center, representing the desired balance between quality and efficiency. Our rules then provide $\hat{N}(\beta)$. For example, in the rationalized regime, our recommended staffing level is $\hat{N}(\beta) = \lceil \hat{R} + \beta\sqrt{\hat{R}} \rceil$, where $\beta = (N - R)/\sqrt{R}$. Moreover, our analysis actually yields explicit approximations for a wide range of performance measures (see §5.2). Specifically, the fraction of customers delayed in queue is expected to remain unchanged, while the fraction of customers abandoning, as well as the average waiting time, will decrease by a factor of $\sqrt{N/\hat{N}}$, thus exhibiting economies of scale.

Figure 3 **β as a Service Grade for Large Call Centers—Correlation with Abandonment**



Note. Each dot plotted represents data of a half-hour interval. The x-coordinate was calculated via $\beta = (N - R)/\sqrt{R}$, where R and N are averages over the half-hours, and the y-coordinate is the fraction abandoning during that half-hour.

One might question the practical value of the above staffing rules since they are based on limits for *large* systems operating in *heavy traffic*. In this regard, Figures 4–8 in Appendix A indicate that our results can be safely applied to call centers with as few as 30 or even 20 highly utilized agents. Moreover, analysis of data from moderate to large call centers shows that indeed both high efficiency and service quality are achieved as the centers operate in the ''rationalized'' regime with $-0.5 \leq \beta \leq 1$. This is demonstrated by the scatter plots in Figure 3. The plots display real data from two call centers, collected in half-hour intervals during a single working day (discarding the opening and closing hours in which the traffic is not ''heavy''). Both traffic volumes and staffing levels varied at these call centers throughout the day (left plot: from 350 to 900 calls per hour, 35 to 90 agents; right plot: 1,400–3,100 calls per hour, 125–250 agents).

3. Calculating Performance Measures

Here we present a useful format for expressing performance measures for an $M/M/N + M$ model in steady state. Details about the calculations of these expressions appear in Appendix B, also covering models with finite capacity ($M/M/N/B + M$). Due to the underlying birth-death structure, such calculations are almost (but not quite, due to numerical issues) trivial. Nevertheless, it is natural and important to present them for practical completeness, as well as a lead to our approximations.

Our calculation of performance measures is based on the assumption that the system has reached its steady state. Although the arrival rate to many call centers is time varying (according to the time of day, day of the week, holidays, seasonal effects, etc.), and other parameters such as the number of agents on the shift may be subject to change, it is assumed that throughout short time intervals (e.g., an hour) such changes are small enough to disregard, and are ''slow'' relative to the speed at which the system reaches its new (e.g., hourly) steady state. This latter assumption can be safely applied to call centers at which the service rate is significantly higher than the rate of such time variations (e.g., hourly variations vs. average service time of a few minutes).

We are interested in a ''typical'' customer, arriving at the system in steady state (for a discussion on the rigorous meaning of ''typical,'' see Appendix B). Let V be the customer's potential waiting time, X be his patience, and W be the actual waiting time ($W = V \wedge X$).

Many performance measures that are of interest to call-center managers can be expressed as expectations of simple functions of V and X. A representative list appears in Table 3. Here we make use of indicator functions of the form $1_{(a,b]}(t)$ that are defined by

$$1_{(a,b]}(t) = \begin{cases} 1, & a < t \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Table 3 **Performance Measures of the Form $E[f(V,X)]$**

$f(v,x)$	$E[f(V,X)]$
$1_{(x,\infty)}(v)$	$P\{Ab\}$
$1_{(t,\infty)}(v \wedge x)$	$P\{W > t\}$
$1_{(t,\infty)}(v \wedge x)1_{(x,\infty)}(v)$	$P\{W > t;\ Ab\}$
$(v \wedge x)1_{(x,\infty)}(v)$	$E[W;\ Ab\}$
$(v \wedge x)1_{(t,\infty)}(v \wedge x)1_{(x,\infty)}(v)$	$E[W;\ W > t;\ Ab]$
$g(v \wedge x)$	$E[g(W)]$

Some important performance measures cannot be expressed directly by the method proposed, but only as quotients of performance measures of the type $E[f(V,X)]$. For example, the fraction of customers abandoning out of those having to wait in queue is an important measure, yet some experienced managers of call centers tend to discard customers who were not willing to wait for even a short period of time t. In such a case one uses

$$P\{Ab \mid W > t\} = \frac{P\{V \wedge X > t;\ V > X\}}{P\{V \wedge X > t\}} = \frac{E[1_{(t,\infty)}(V \wedge X)1_{(X,\infty)}(V)]}{E[1_{(t,\infty)}(V \wedge X)]}.$$

4. Operational Regimes and Diffusion Approximations

A central outcome of the present paper is approximations of performance measures that yield insight as to their dependence on the model's parameters. From our theoretical results, which are also supported by prevailing practice, it follows that moderate to large telephone call centers are capable of delivering high service level while also operating under high utilization. This justifies our focus on approximations through heavy traffic limits, as $N \to \infty$ (this is equivalent to $R \to \infty$, since we will be assuming $N/R \to 1$). To this end, a subscript N will now be added to our notation to indicate the processes and parameters of the Nth system (i.e., associated with an $M/M/N + M$ model).

Motivated by the work of Halfin and Whitt (1981) we use two performance measures—the fraction of customers abandoning ($P_N\{Ab\}$), and the fraction delayed in queue ($P_N\{W > 0\}$)—as guidelines for choosing appropriate operational regimes. Most telephone call centers try to avoid a high percentage of abandonment, without over-staffing. This usually translates into operating with a nonnegligible fraction of customers having to queue, and a small fraction of abandonment.

We now characterize the dependence of the model's parameters on N. Primarily, we are interested in a sequence in which $\lambda_N \to \infty$ as $N \to \infty$ and $\mu_N \equiv \mu$, which corresponds to scaling up the staffing level (N) to accommodate the increasing load (λ_N) while maintaining a service rate (μ) that does not vary with staffing level or load. Furthermore, we restrict our current discussion to the case $\lim_{N\to\infty} \theta_N = \theta$, $0 < \theta < \infty$, although our results, as reported in Appendix C, include the regimes $\theta = 0$ (extreme patience) and $\theta = \infty$ (extreme impatience) as well.

We first introduce the notion of *traffic intensity* defined by $\rho_N = \lambda_N/N\mu$. The results of Theorems 1 and 2 below, together with the guidelines stated above, lead us to focus on the following regime:

$$\lim_{N\to\infty} \sqrt{N}(1 - \rho_N) = \beta, \quad -\infty < \beta < \infty.$$

The discussion later in §5.3, formalized by Theorem 4, strongly supports our contention that this is indeed the regime of interest for moderate to large call centers.

Following are Theorems 1 and 2, with a discussion about the derivation of the stated regimes from their results. Proofs of these theorems, and other selected results quoted throughout the paper, appear in Appendix C.

THEOREM 1. *Assume that* $\lim_{N\to\infty} \rho_N = \rho_\infty$, *for some* $0 \le \rho_\infty \le \infty$. *Then the limiting behavior of the fraction of abandoning customers is given by*

$$\lim_{N\to\infty} P_N\{Ab\} = \begin{cases} 0, & 0 \le \rho_\infty \le 1, \\ 1 - \dfrac{1}{\rho_\infty}, & \rho_\infty > 1. \end{cases}$$

Based on this result, it seems clear that from the point of view of abandonment there is no reason to operate with $\rho_\infty < 1$: $\rho_\infty = 1$ already yields a vanishing abandonment probability. On the other hand, when $\rho_\infty \gg 1$, the limiting abandonment probability is high-

er than usually desired. Moreover, from the point of view of the agents' *utilization* (i.e., the fraction of time they spend answering calls, given by $[\lambda_N(1 - P_N\{Ab\})]/N\mu$), the maximum limiting utilization is already achieved with $\rho_\infty = 1$. Thus, $\rho_\infty = 1$ arises as a special balance point between the call center's efficiency and service quality.

The restriction to $\rho_\infty = 1$ is consistent with the work of Halfin and Whitt (1981) who analyze the $M/M/N$ model, and find that interesting limiting behavior occurs when $\rho_N \sim 1 - \beta/\sqrt{N}$, $0 < \beta < \infty$ (''interesting'' in the sense that only then is the limiting behavior of the fraction of customers having to wait in queue nondegenerate). Because an $M/M/N + M$ model with very patient customers is ''close'' to an $M/M/N$ model (supported by Theorem 2* appearing in Appendix C—an extension of Theorem 2 below), we also restrict ourselves to the case of $\lim_{N\to\infty} \sqrt{N}(1 - \rho_N) = \beta$, but here $-\infty < \beta < \infty$. (As already mentioned, Halfin and Whitt (1981) covers only $\beta > 0$, because otherwise there is no steady state.)

Theorem 2 below states diffusion approximations for the process $Q = \{Q(t), t \geq 0\}$. We consider the sequence of stochastic processes $\{q_N\}$ that is obtained from $\{Q_N\}$ through centering and rescaling, namely

$$q_N(t) = \frac{Q_N(t) - N}{\sqrt{N}}.$$

Centering around N gives rise to a process whose absolute value is either the queue length ($q_N \geq 0$) or the number of idle servers ($q_N \leq 0$). The rescaling factor $\sqrt{N}$ emerges as the appropriate order of magnitude, that gives rise to a nontrivial *continuous* limiting process q. The latter will be used to approximate our original birth-death processes $\{Q_N\}$, via $Q_N \overset{d}{\approx} N + q\sqrt{N}$, thus offering approximations for both transient ($Q_N(t)$, $t > 0$) and steady-state ($Q_N(\infty)$) behavior.

The mathematical details of the theorem are not a prerequisite for following its consequences, which are explained immediately after the theorem and its corresponding remarks.

THEOREM 2. *Assume that*

$$\lim_{N\to\infty} \sqrt{N}(1 - \rho_N) = \beta, \qquad -\infty < \beta < \infty,$$

$$\lim_{N\to\infty} \theta_N = \theta, \qquad 0 < \theta < \infty.$$

If $q_N(0) \overset{d}{\to} q(0)$, *then* $q_N \overset{d}{\to} q$, *where* q *is the unique solution of the following stochastic differential equation*

$$\begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}db(t), \\ f(x) = \begin{cases} -\mu(\beta + x), & x \leq 0, \\ -(\mu\beta + \theta x), & x > 0. \end{cases} \end{cases}$$

(Here b denotes a standard Brownian Motion, and $X_N \overset{d}{\to} X$ denotes the ''weak convergence'' or ''convergence in distribution'' of a sequence $\{X_N\}$ to X.)

REMARKS.

(1) This limit was conjectured (with a slightly different centering) by Fleming et al. (1994), and a proof was given for the weak limit of the stationary distributions (i.e., $q(\infty)$). An extended version of this theorem, including diffusion limits for the cases $\theta = 0$ and $\theta = \infty$ appears in Appendix C.

(2) The limiting process stated in this theorem is qualitatively characterized by a combination of two Ornstein-Uhlenbeck processes ($q \leq 0$, $q > 0$) with different restraining forces.

So far we have dealt with diffusion limits of the queue-length process Q. However, as displayed in Table 3, many performance measures involve the *potential waiting time* V. Now, the distribution of V coincides (see Appendix B) with the stationary distribution of the process $v(t)$—the *virtual waiting time* at time t (i.e., the time spent waiting in queue by a hypothetical infinitely patient customer arriving at time t). We shall thus focus on approximating the stationary distribution of $v(t)$, denoted $v(\infty)$.

A simple relationship between the diffusion limits of the queue-length process and the virtual waiting-time process can be motivated heuristically as follows: If there are idle agents, the virtual waiting time is zero; otherwise the number of waiting customers is $\approx q\sqrt{N}$ (in view of Theorem 2). How long does it take for a customer to pass through this queue? Customers will be leaving at a rate of $N\mu$ (through service) + $o(N)$ (abandonment; indeed, the abandonment rate of customers in front of our tagged customer is no greater

than $\theta q\sqrt{N}$). Dividing the queue length by the rate that customers are leaving it yields the virtual waiting time, which is therefore $\approx[q/(\sqrt{N}\mu)]^+$.

Formally, such an approximation is derived through Theorem 3 that follows. Here we use the common notation for the standard normal density and distribution functions (ϕ and Φ respectively)

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \qquad \Phi(x) = \int_{-\infty}^{x} \phi(y)\, dy,$$

as well as the hazard rate defined by

$$h(x) = \phi(x)/[1 - \Phi(x)] = \phi(x)/\Phi(-x).$$

Proof of this theorem is based on a useful result by Puhalskii (1994) that links diffusion approximations for ν and Q.

THEOREM 3. *Let $v = [q/\mu]^+$, where q solves the stochastic differential equation stated in Theorem 2. Then*

(1) $\sqrt{N}\nu_N(t) \xrightarrow{d} v(t)$, $0 \le t \le \infty$, where both $\nu_N(\infty)$ and $v(\infty)$ are limits in distribution, as $t \to \infty$, of $\nu_N(t)$ and $v(t)$, respectively.

(2) $v(\infty)$ has the distribution function F_v given by

$$1 - F_v(x) = \begin{cases} w(-\beta, \sqrt{\mu/\theta}), & x = 0, \\ w(-\beta, \sqrt{\mu/\theta}) \cdot \dfrac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{\mu\theta}\,x)}, & x > 0. \end{cases}$$

Here

$$w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1},$$

$$\Psi(x, y) = \frac{\phi(x)}{1 - \Phi(x + y)}.$$

Because $\sqrt{N}\nu_N(\infty) \xrightarrow{d} v(\infty)$, the approximation we use is

$$V \stackrel{d}{=} \nu_N(\infty) \stackrel{d}{\approx} v(\infty)/\sqrt{N},$$

which translates into $F_V(x) \approx F_v(\sqrt{N}x)$.

5. Implementation

In §§3 and 4 we reported a variety of results concerning the $M/M/N + M$ model. Because we advocate this model as a substitute for the $M/M/N$ model commonly used in call-center analysis, it makes sense to shed some light on how to apply and interpret our results. The context of the discussion will be that of managing a moderate to large call center in heavy traffic (''heavy'' such that the abandonment phenomenon is not negligible). First, we briefly address the issue of estimating the values of the model's parameters. Then we suggest which performance measures should be used by call-center managers to define the service level. Finally, we discuss the use of our approximation results and derive ''staffing rules.''

5.1. Estimating the Parameters

To use the model and the results introduced, it is necessary to determine the values of the various parameters. The number of agents on shift is fully controlled by the call center's manager. Arrival and service rates are usually estimated from historical ACD data. As discussed in §3 for time varying arrival rates, small time intervals are selected, in which the arrival rate is approximately constant.

The main difficulty is to estimate the abandonment rate (θ) or equivalently, the average patience ($1/\theta$). The difficulty arises from the fact that the direct data we can collect is censored—we can only measure the patience of customers who abandon the system before their service began. For the customers receiving service we only have a lower bound for their patience—the amount of time they spent waiting in queue. There are statistical methods to deal with such censored samples. While we shall not discuss these methods here, interested readers are referred to the Appendix in Zohar et al. (2000) for a survey and further statistical references. Another, more basic problem for estimating θ, is that in most cases the ACD data only contains averages, as opposed to call-by-call measurements (see Mandelbaum et al. (2000) for call-by-call data analysis). To this end we suggest a method for estimating the average patience that is based on the following balance equation:

$$\theta \cdot E[\#\text{ waiting in queue}] = \lambda P\{Ab\}. \qquad (2)$$

This equation describes the steady-state balance between the rate that customers abandon the queue

(left-hand side) and the rate that abandoning customers (i.e., customers who will eventually abandon) enter the system.

Through Little's theorem ($\lambda \cdot E[W] = E[\#\text{ waiting in queue}]$), we obtain an alternative equation

$$\theta \cdot E[W] = P\{Ab\}. \tag{3}$$

The average wait in queue and fraction of customers abandoning are fairly standard ACD data outputs, thus providing the means for estimating θ. We note, however, that (2) and (3) are known to hold exactly only under *exponentially* distributed patience. (See Zohar et al. (2000) for a discussion of (3) and generalizations.)

5.2. Approximations

As stated earlier, the abandonment phenomenon is extremely important to a call center's manager. Nevertheless, this does not imply that the only performance measure of interest is the fraction of customers abandoning the queue. There are many additional important performance measures, and it is necessary to select the few that best reflect the service level at the call center, and can serve as service goals and service grades.

Approximations can be used to overcome computational difficulties arising when attempting exact evaluation of performance measures, but they can also reveal how performance measures depend on the model's parameters. Such an understanding is necessary when trying to derive simple rules of thumb (see §5.3 below).

Combining the approximation for the stationary virtual waiting time (Theorem 3) with the general representation of performance measures in §3 enables us to derive approximations for many performance measures. These approximations should be most accurate in the case of a large call center operating in heavy traffic, with negligible blocking. (The accuracy of some of these approximations is demonstrated in Appendix A.)

Example. Assume a performance measure that can be expressed as $E[g(W)]$ for some function g. Recall that $W \equiv X \wedge V$, and that X and V are independent. Therefore,

$$\begin{aligned}
&E[g(W)] \\
&= \int_0^\infty \int_0^\infty g(x \wedge v)\theta e^{-\theta x}\, dF_V(v)\, dx \\
&\approx E[g(0)](1 - w(-\beta, \sqrt{\mu/\theta})) \\
&\quad + \int_0^\infty \int_0^\infty g(x \wedge v)\theta e^{-\theta x}\sqrt{N\mu\theta}\cdot w(-\beta, \sqrt{\mu/\theta}) \\
&\quad \times \Psi(\sqrt{N\mu\theta}v + \beta\sqrt{\mu/\theta}, -\sqrt{N\mu\theta}v)\, dv\, dx.
\end{aligned}$$

The resulting approximations for several performance measures are

$$P\{W > 0\} \approx w(-\beta, \sqrt{\mu/\theta}),$$

$$P\{Ab \mid W > 0\} \approx 1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})},$$

$$P\{Ab\} \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right]\cdot w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right),$$

$$E[W] \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right]\cdot w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right)\cdot\frac{1}{\theta},$$

$E[\#\text{ busy agents}]$

$$\approx \frac{\lambda}{\mu} - \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right]\cdot w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right)\cdot\frac{\lambda}{\mu},$$

$E[\#\text{ waiting in queue}]$

$$\approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right]\cdot w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right)\cdot\frac{\lambda}{\theta},$$

$$P\{W > t\} \approx w(-\beta, \sqrt{\mu/\theta})\cdot\frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t)}\cdot e^{-\theta t}, \quad t \geq 0,$$

$$P\{Ab \mid W > t\} \approx 1 - \frac{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t)}{\Psi(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta}t)}\cdot e^{\theta t}, \quad t \geq 0.$$

Remarks.

(1) Along the same lines, we have developed further useful approximations, notably for $E[W \mid W > t]$ and $E[W \mid \text{Served}]$. These have been omitted due to excessive "bulk."

(2) Some of the approximations above can also be derived via the diffusion limit of the queue-length

process (Theorem 2). Note, however, that the approximating expressions thus arrived at are not identical, but they coincide as $N \to \infty$. For example, considering the fraction of customers abandoning, we have on the one hand

$$P\{Ab\} = E[1_{(X,\infty)}(V)] = \int_0^\infty \int_0^t \theta e^{-\theta x}\, dx\, dF_V(t)$$

$$\approx \int_0^\infty (1 - e^{-\theta t/\sqrt{N}})\, dF_v(t),$$

and on the other hand (using (2))

$$P\{Ab\} = \frac{\theta}{\lambda} E[\#\text{ waiting in queue}]$$

$$\approx \int_0^\infty \frac{\theta t\sqrt{N}}{\lambda}\, dF_q(t) \left(\approx \int_0^\infty \frac{\theta t\sqrt{N}}{\lambda}\, dF_v(t) \right).$$

5.3. Staffing Rules

It is important for a call center's manager to be able to anticipate the impact of changes on the service level. Examples of such a change are an increase in the call arrival rate due to a marketing campaign, or a change in the number of agents on shift.

Most expressions for performance measures derived using the $M/M/N + M$ model are quite complex. Even the approximations in §5.2 tend to be too complex to enable an understanding of how the values of the parameters affect the performance measure. It is desirable, therefore, to derive simple "rules of thumb" to support decision making.

We have the following result, analogous to the result by Halfin and Whitt (1981) that concerns $M/M/N$ queues.

THEOREM 4. *Assume that* $\theta_N \equiv \theta$, $0 < \theta < \infty$. *Then*

$$\lim_{N\to\infty} \sqrt{N}(1 - \rho_N) = \beta, \qquad -\infty < \beta < \infty,$$

if and only if

$$\lim_{N\to\infty} P_N\{W > 0\} = \alpha, \qquad 0 < \alpha < 1,$$

if and only if

$$\lim_{N\to\infty} \sqrt{N} P_N\{Ab\} = \Delta, \qquad 0 < \Delta < \infty,$$

in which case

$$\alpha = w(-\beta, \sqrt{\mu/\theta}),$$

$$\Delta = [\sqrt{\theta/\mu} \cdot h(\beta\sqrt{\mu/\theta}) - \beta] \cdot \alpha.$$

(Here w and h are as in Theorem 3 above).

REMARK. This result holds at the "extremes" as well, namely

$$\beta = -\infty \quad \text{iff } \alpha = 1 \quad \text{iff } \Delta = \infty \quad \text{and}$$

$$\beta = \infty \quad \text{iff } \alpha = 0 \quad \text{iff } \Delta = 0.$$

We deduce from the above results that also for $M/M/N + M$ queues the "interesting" (as explained in §4) limiting behavior is when $\rho_N \sim 1 - \beta/\sqrt{N}$, but here (in contrast to Halfin and Whitt 1981) β is *not* restricted to be positive.

In light of Theorem 4 and Theorem 1, we introduce in Table 4 three regimes of operation, with three matching staffing rules (this is a slightly extended version of Table 1 from §1).

Table 4 Staffing Rules for Three Operational Regimes

Regime	Staffing Level		Guidelines
Rationalized	$N = [R + \beta\sqrt{R}]$		$P\{W > 0\} \to \alpha(\beta)$ and $P\{Ab\} \sim \Delta(\beta)/\sqrt{N}$
Quality-Driven	$N = [R + \epsilon R]$,	$\epsilon > 0$	$P\{W > 0\} \to 0$ and $P\{Ab\} = o(1/\sqrt{N})$
Efficiency-Driven	$N = [R - \epsilon R]$,	$\epsilon > 0$	$P\{W > 0\} \to 1$ and $P\{Ab\} \to \epsilon$

REMARKS.

(1) The staffing level for the rationalized regime is derived directly from $\rho_N \sim 1 - \beta/\sqrt{N}$, $-\infty < \beta < \infty$ (see §§1 and 2 in Whitt (1992) for a detailed discussion). The "extremes" of Theorem 4 only set bounds for the staffing levels. Any staffing level such as $N = \lceil \lambda \pm \epsilon \cdot \lambda^a \rceil$ with $\epsilon > 0$, $1 \geq a > 0.5$, is adequate ($+\epsilon$ for quality-driven, $-\epsilon$ for efficiency-driven). However, we suggest taking $a = 1$, with which there is a clear differentiation between slightly underloaded call centers (quality-driven), slightly overloaded call centers (efficiency-driven), and the "critically" loaded call centers (rationalized).

(2) The "guidelines" above follow directly from Theorem 4, except for the fraction abandoning ($P\{Ab\}$) in the efficiency-driven regime that involves Theorem 1.

Following the result of Theorem 4, and continuing

in the spirit of Whitt (1992), we suggest β (or ϵ) as a ''service grade.'' The main significance of this grade is for comparing two systems, in particular, in the case of a single system before and after an expected change. Once a manager has decided which of the three regimes of operation is suitable for her call center, she can determine the service grade and use the appropriate staffing rule. Moreover, analysis of empirical data (Figure 3 is representative) shows that, in practice, the value of β lies in the range of $-0.5 < \beta < 1$.

We conclude by *revisiting the scenario* from §2: Suppose a given call center operates in the ''rationalized'' regime with N agents, service rate μ, and arrival rate λ. The service level is quantified by a service grade β. There is a forecast of a higher arrival rate $\hat{\lambda}$ during a holiday. The call center's manager wishes to maintain the service level at the call center, and needs to decide how many agents to have on shift ($\hat{N}$). Based on the appropriate staffing rule $\beta \approx \sqrt{\mu/\lambda}N(1 - \lambda/N\mu)$, we get $\hat{N} = \lceil \hat{\lambda}/\mu + \beta\sqrt{\hat{\lambda}/\mu} \rceil$. Moreover, the anticipated holiday performance is:

(1) Fraction waiting: $P\{W > 0\} \approx \alpha(\beta)$ (as in the original system).

(2) Fraction abandoning: $P\{Ab\} \approx \Delta(\beta)/\sqrt{\hat{N}}$ (decrease by a factor of $\sqrt{\hat{N}/N}$).

Acknowledgments

The authors thank the Editor-in-Chief, the Senior Editor, and the referees for an exceptionally thorough review process. Their helpful comments turned the very different originally submitted manuscript into the present, much more readable version. A. Mandelbaum thanks Professor I. Meilijson of Tel-Aviv University and Professor O. Kella of the Hebrew University in Jerusalem. These colleagues have contributed greatly to his understanding of call centers, and to the challenge and joy of analyzing them. Mandelbaum's research was supported by the fund for the promotion of research at the Technion, by the Technion V.P.R. funds—Smoler Research Fund and B. and G. Greenberg Research Fund (Ottawa)—and by the Israel Science Foundation (grant no. 388/99).

Appendix A. Accuracy of the Approximations

The approximations derived in §5.3 are based on heavy traffic limit theorems in which the number of agents and call volume are taken to infinity. It is therefore of practical interest to see how accurate these approximations are when applied to call centers that are not extraordinarily large.

In Figures 4–8 we plot approximations ('+' signs) for a number of performance measures vs. exact values (solid line) based on the analysis of an $M/M/N + M$ model. All cases assume a call volume of 50 calls per minute, each requiring an average handling time of one minute. Staffing levels range from 20 to 80 agents (implying traffic intensities from 0.625 up to 2.5!). Three different values of average patience are considered:

- Graph a: 10 minutes (very patient).
- Graph b: 1 minute (moderately patient).
- Graph c: 6 seconds (very impatient).

The main conclusion from this display is that in most cases these approximations are excellent (for any practical use) even in the case

Figure 4 **Approximating** $P\{\text{Wait} > 0\}$

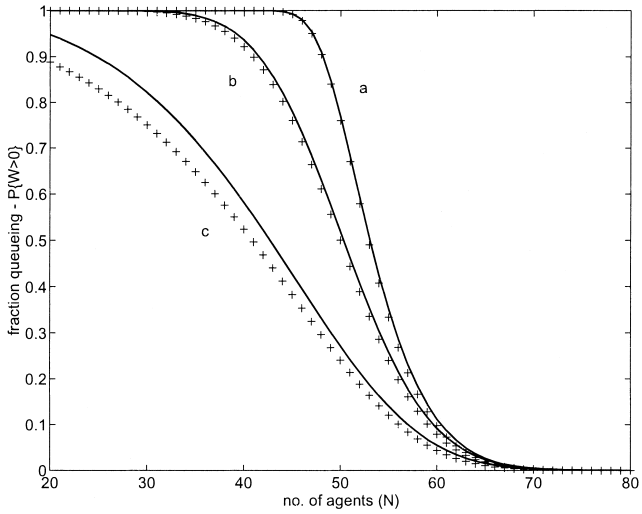


Figure 5 **Approximating** $P\{\text{Abandon}\}$

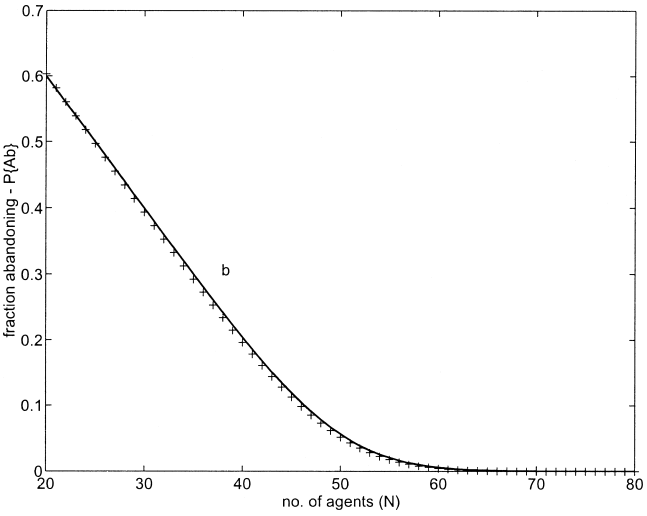


Figure 6 **Approximating** $P\{\text{Wait} > 10 \text{ secs.}\}$

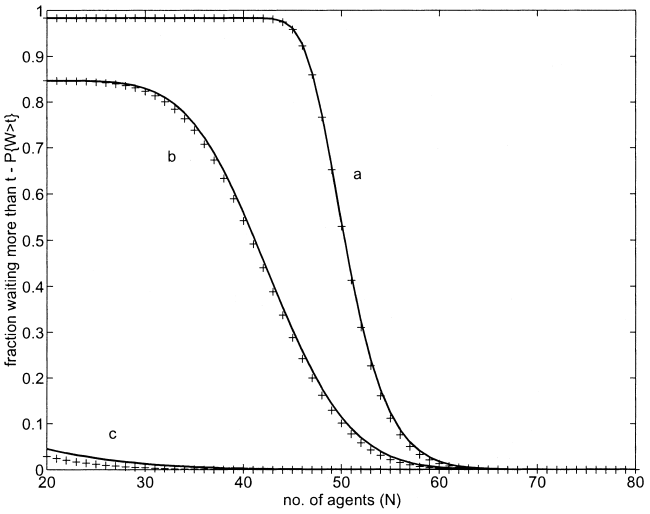
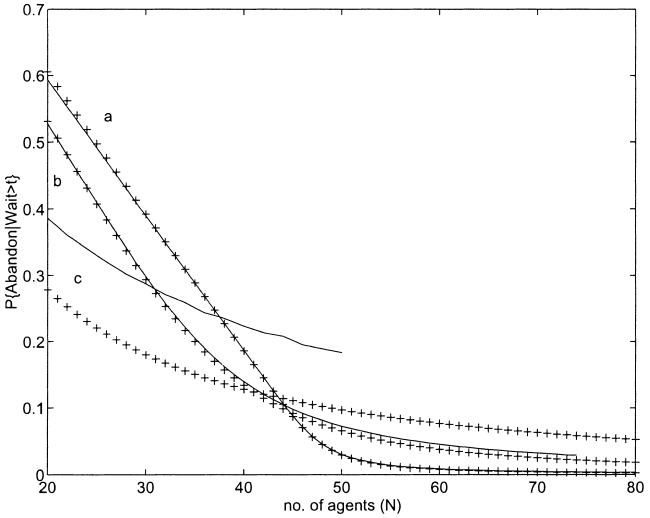


Figure 7 **Approximating** $P\{\text{Abandon/Wait} > 10 \text{ secs.}\}$

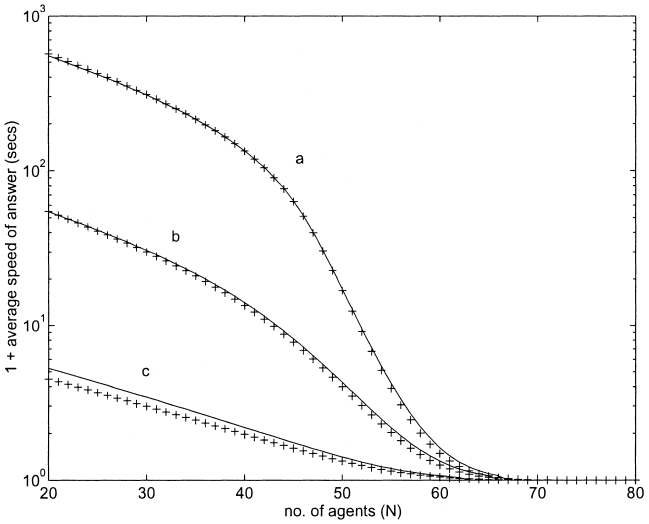


of a medium-sized call center handling moderate traffic intensities. Some additional, specific remarks are stated below.

REMARKS.

(1) Figure 5 shows only a single graph (b) because the graphs (both exact and approximate) of the other two cases (a and c) practically coincide with it. This is easy to explain from our theoretical results. By Theorem 1, to leading order the abandonment probability with $\rho > 1$ does not depend on θ. Although the abandonment probability with $\rho \leq 1$ is more sensitive to θ (the leading order term is zero), the differences do not show up due to the scale of the graph.

Figure 8 **Approximating** $E[\text{Wait} | \text{Served}]$



(2) In Figure 7 the exact-value graphs for cases b and c were not calculated for all values up to 80 agents, due to numerical difficulties in obtaining these values. Specifically, because we use $P\{Ab | W > t\} = (P\{Ab; W > t\})/(P\{W > t\})$, when $P\{W > t\}$ becomes extremely small (see Figure 6), one encounters precision difficulties. To overcome such problems, the exact-value graph for case c was produced using simulations. This graph is limited to 50 agents because as the number of agents increases, the event in which a call waits in queue for more than 10 seconds becomes extremely rare, thus requiring very long simulations. The difficulties we encountered here provide a good example of the benefits of having approximations for such performance measures. Indeed, the approximation is not as accurate as the others, but the values it provides are useful and capture exact behavior.

(3) Note that the scale in Figure 8 is of log-type ($\log_{10}$), for benefit of the clarity of display. The upper range of the graphs (a, b, and c) in this case is 550, 55, and 5 seconds, respectively.

Appendix B. Calculating $E[f(V, X)]$ in an $M/M/N/B + M$ Model

To calculate $E[f(V, X)]$, we start with the following decomposition:

$$E[f(V, X)] = E[f(V, X)\cdot 1_{(0,\infty)}(V)] + E[f(V, X\cdot 1_{\{0\}}(V)]$$

$$= E[f(V, X)\cdot 1_{(0,\infty)}(V)] + E[f(0, X)]\cdot\left(\pi_B + \sum_{k=0}^{N-1} \pi_k\right). \quad (4)$$

Here we use π to denote the stationary distribution of the queue-length process $Q(t)$, namely

$$\lim_{t\to\infty} P\{Q(t) = n\} = \pi_n, \qquad n = 0, 1, 2, \ldots, B.$$

A general expression for these probabilities is given by

$$\pi_k = \begin{cases} \dfrac{(\lambda/\mu)^k}{k!}\pi_0, & 0 \le k \le N, \\ \displaystyle\prod_{j=N+1}^{k}\left(\frac{\lambda}{N\mu + (j-N)\theta}\right)\frac{(\lambda/\mu)^N}{N!}\pi_0, & N < k \le B, \end{cases}$$

where

$$\pi_0 = \left[\sum_{k=0}^{N}\frac{(\lambda/\mu)^k}{k!} + \sum_{k=N+1}^{B}\prod_{j=N+1}^{k}\left(\frac{\lambda}{N\mu + (j-N)\theta}\right)\frac{(\lambda/\mu)^N}{N!}\right]^{-1}.$$

REMARK. For a blocked customer (i.e. the queue was full upon his arrival) the convention $V = 0$ is introduced.

For all functions f which seem of interest in our case, $E[f(0, X)]$ evaluates to zero or one. Therefore, we proceed to calculate the first expression. We present three different methods for performing this calculation, each with its own virtues and drawbacks.

Our calculations require the distribution function of V. Recall that V is the *potential waiting time* of a typical customer. What is meant by a ''typical'' customer? Consider the sequence $\{w_n, n \in \mathbb{N}\}$, where w_n is the potential waiting time of the nth customer. Let F_w be the stationary distribution of this sequence. Quoting from Baccelli and Hebuterne (1981), F_w is also the stationary distribution of the process $v(t)$—the *virtual waiting time* at time t (i.e., the time spent waiting in queue of a hypothetical infinitely patient customer arriving at time t). Therefore, a typical customer's potential waiting time, V, has distribution function F_w.

Similarly, we are interested in V_n, which is a random variable whose distribution is that of V given n customers in queue upon arrival, and all agents busy, $n = 0, 1, \ldots$; V_n has distribution function F_n.

The distribution of V is not given beforehand, and is derived through analysis of the model. On the other hand, V_n can be expressed as the sum of $n + 1$ independent exponential random variables with parameters $N\mu$, $N\mu + \theta$, $\ldots$, $N\mu + n\theta$, the ith of these representing the period of time the customer spent in the ith place in queue, before advancing to the $(i - 1)$th (due to end of service or abandonment from the queue in front of him).

Method A. Conditioning on the number of customers in the queue upon arrival, and substituting the explicit expression given by Riordan (1962) (Equation (83) on p. 111) for $\bar{F}_n(t) = 1 - F_n(t)$, we have

$$E[f(V, X)1_{(0,\infty)}(V)] = c\pi_N \sum_{k=0}^{B-N-1}(-1)^k\frac{(\lambda/\theta)^k}{k!}I(k)\sum_{n=k}^{B-N-1}\frac{(\lambda/\mu)^{n-k}}{(n-k)!}, \quad (5)$$

where

$$I(k) = \theta^2 c\int_0^\infty\int_0^\infty f(t, x)e^{-(c+k)\theta t}e^{-\theta x}\,dt\,dx \quad \text{and} \quad c = N\mu/\theta.$$

Calculating the values of $I(k)$ is usually a simple task. The main drawback of this method is the alternating signs in the first sum, which cause it to be numerically unstable. Therefore, we present the next method, which avoids this problem.

Method B. Starting similarly to Method A, and using the relation

$$\sum_{k=0}^{n}\binom{n}{k}(-e^{-\theta t})^k = (1 - e^{-\theta t})^n$$

to eliminate one sum, we arrive at

$$E[f(V, X)1_{(0,\infty)}(V)] = \theta^2 c\pi_N \sum_{n=0}^{B-N-1}\frac{(\lambda/\theta)^n}{n!}J(n), \quad (6)$$

where

$$J(n) = \int_0^\infty\int_0^\infty f(t, x)e^{-(x+ct)\theta}(1 - e^{-\theta t})^n\,dx\,dt. \quad (7)$$

Here, calculating the values of $J(n)$ tends to be more costly because the integrals must usually be solved numerically.

These methods lose some of their attractiveness when dealing with infinite buffers ($B = \infty$). Then, sums appearing in both methods become infinite, and must be truncated at some point for implementation (the alternating signs in Method A can be problematic in the aspect of truncation). Because this case forces us to consider the issue of precision tolerance, we present the third method, which is a straightforward numerical integration.

Method C. Following Riordan (1962), and solving the more general case of any buffer size B, we arrive at the function f_V^+, where $f_V^+/P\{V > 0\}$ is a density function, given by

$$f_V^+(t) = N\mu\pi_N\left[1 - \frac{\gamma\left(B - N, \dfrac{\lambda}{\theta}(1 - e^{-\theta t})\right)}{\Gamma(B - N)}\right] \times \exp\left\{\frac{\lambda}{\theta}(1 - e^{-\theta t}) - N\mu t\right\}, \quad t > 0. \quad (8)$$

Here Γ and γ denote the gamma and incomplete gamma functions, respectively, defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)\,dt \quad \text{and}$$

$$\gamma(x, y) = \int_0^y t^{x-1}\exp(-t)\,dt, \qquad y > 0.$$

Now we are left with the evaluation of the double integral

$$E[f(V, X)\cdot 1_{(0,\infty)}(V)] = \int_0^\infty\int_0^\infty f(t, x)\theta e^{-x\theta}f_V^+(t)\,dx\,dt. \quad (9)$$

The integral with respect to x is usually solved analytically and rather easily (depending on f), leaving us to perform one numerical integration (with respect to t).

Some additional remarks concerning the infinite buffer case follow.

REMARKS.

(1) When the system's buffer is unlimited, solving the stationary distribution equations involves an infinite sum. A solution is given by Palm (1943), expressing the stationary distribution as a function of the easily calculated blocking probability in an $M/M/N/N$ system (denoted here by $P\{Bl\}$), with the same arrival and service rates:

$$\pi_n = \begin{cases} \dfrac{P\{Bl\}}{1 + \left(A\left(\dfrac{\lambda}{N\mu}, \dfrac{N\mu}{\theta}\right) - 1\right)P\{Bl\}} \cdot \dfrac{N!}{n!\left(\dfrac{\lambda}{\mu}\right)^{N-n}}, & n < N, \\ \dfrac{P\{Bl\}}{1 + \left(A\left(\dfrac{\lambda}{N\mu}, \dfrac{N\mu}{\theta}\right) - 1\right)P\{Bl\}} \times \dfrac{\left(\dfrac{\lambda}{\theta}\right)^{n-N}}{\left(\dfrac{N\mu}{\theta} + 1\right)\cdots\left(\dfrac{N\mu}{\theta} + (n - N)\right)}, & n \geq N, \end{cases}$$

where

$$A(x, y) = \frac{ye^{xy}}{(xy)^y} \cdot \gamma(y, xy).$$

(2) For $B = \infty$ the density function f_V^+ given here becomes a special case of the result by Baccelli and Hebuterne (1981) for an $M/M/N + G$ model with patience distribution F, namely

$$f_V^+(t) = N\mu\pi_N \exp\left\{\lambda \int_0^t (1 - F(u))\, du - N\mu t\right\}, \qquad t > 0.$$

Appendix C. Outlines of Proofs

Outlines for the proofs of Theorems 1–4 follow.

Proof of Theorem 1. We first point out an intuitive approach for the overloaded case ($\rho_\infty > 1$), based on the fact that in systems with many agents it is possible to achieve very high utilization. Indeed, in an $M/M/N + M$ model the utilization is given by the rate at which ''work'' reaches the agents ($\lambda_N(1 - P_N\{Ab\})$) divided by the maximum rate at which it can be processed ($N\mu$). Thus when $N \to \infty$, assuming that the utilization is ≈ 1, we obtain the result.

We now proceed with the rigorous proof, based on bounding the sequence $\{P_N\{Ab\}\}$ from above and below, with the two bounds converging to the desired limit.

We begin with the lower bound, which is more intuitive. The utilization of agents in an $M/M/N + M$ queue in steady state must be less than one. Therefore,

$$\lambda_N(1 - P_N\{Ab\}) < N\mu,$$

from which we obtain

$$\liminf_{N\to\infty} P_N\{Ab\} \geq 1 - \frac{1}{\rho_\infty}.$$

Before turning to the upper bound, we note two monotonicity properties of $P_N\{Ab\}$ that are proved in Bhattacharya and Ephremides (1991):

(i) With N, θ, and μ fixed, $P_N\{Ab\}$ is increasing in λ (or ρ).

(ii) With N, μ, and λ fixed, $P_N\{Ab\}$ is increasing in θ.

Now we deal with the upper bound. Note that $P_N\{Bl\}$, the probability of blocking in an $M(\lambda_N)/M(\mu)/N/N$ queue, is the limit of $P_N\{Ab\}$ as $\theta \to \infty$ in the $M(\lambda_N)/M(\mu)/N + M(\theta_N)$ queue. Thus, by (ii) above, $P_N\{Ab\} \leq P_N\{Bl\}$ for any θ_N with $0 < \theta_N < \infty$.

If $\lambda_N = N\mu\rho_\infty$, with $1 < \rho_\infty < \infty$, it was shown by Jagerman (1974, p. 538), that

$$P_N\{Bl\} \sim \left[\frac{\rho_\infty}{\rho_\infty - 1} - \frac{\rho_\infty}{(\rho_\infty - 1)^3}\frac{1}{N} + \frac{2\rho_\infty^2 + \rho_\infty}{(\rho_\infty - 1)^5}\frac{1}{N^2}\right]^{-1}. \qquad (10)$$

We deal with $\lambda_N = N\mu\cdot\rho_\infty + o(N)$ as follows. Choose $0 < \epsilon < \rho_\infty - 1$, and define $\lambda_N^+ = N\mu\cdot(\rho_\infty + \epsilon)$, $\lambda_N^- = N\mu\cdot(\rho_\infty - \epsilon)$. Hence we have (through Jagerman's result and monotonicity)

$$\frac{\rho_\infty - \varepsilon - 1}{\rho_\infty - \varepsilon} \leq \liminf_{N\to\infty} P_N\{Bl\} \leq \limsup_{N\to\infty} P_N\{Bl\} \leq \frac{\rho_\infty + \varepsilon - 1}{\rho_\infty + \varepsilon},$$

and taking $\epsilon \downarrow 0$ yields

$$\limsup_{N\to\infty} P_N\{Ab\} \leq \lim_{N\to\infty} P_N\{Bl\} = 1 - \frac{1}{\rho_\infty}$$

for $\rho_\infty > 1$.

We complete the proof for $\rho_\infty \leq 1$ using (i) above: Because $P_N\{Ab\}$ with $\rho_\infty \leq 1$ must be smaller than with $\rho_\infty = 1 + \epsilon$ for any $\epsilon > 0$, we have that for $\rho_\infty \leq 1$

$$\limsup P_N\{Ab\} \leq \lim_{\varepsilon\downarrow 0}\left(1 - \frac{1}{1 + \varepsilon}\right) = 0.$$

An extended version of Theorem 2 that includes diffusion limits for the cases $\theta = 0$ and $\theta = \infty$ follows.

Theorem 2*. *Assume that*

$$\lim_{N\to\infty} \sqrt{N}(1 - \rho_N) = \beta, \qquad -\infty < \beta < \infty,$$

$$\lim_{N\to\infty} \theta_N = \theta, \qquad 0 \leq \theta \leq \infty.$$

If $q_N(0) \xrightarrow{d} q(0)$, *then*

(1) *Weak convergence:* $q_N \xrightarrow{d} q$, *where* q *is the unique solution of a stochastic differential equation, according to the following regimes:*

$\theta = 0$:
$$\begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}db(t), \\ f(x) = \begin{cases} -\mu(\beta + x), & x \leq 0, \\ -\mu\beta, & x > 0. \end{cases} \end{cases}$$

$0 < \theta < \infty$:
$$\begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}db(t), \\ f(x) = \begin{cases} -\mu(\beta + x), & x \leq 0, \\ -(\mu\beta + \theta x), & x > 0. \end{cases} \end{cases}$$

$\theta = \infty$:
$$\begin{cases} dq(t) = -\mu(\beta + q(t))dt + \sqrt{2\mu}db(t) - dY(t), \\ q \leq 0; \quad Y(0) = 0, \quad Y \text{ nondecreasing}, \\ \displaystyle\int_0^\infty q\, dY = 0. \end{cases}$$

(2) *Interchangeable limits:* $\lim_{N\to\infty} P\{q_N(\infty) \leq x\} = \lim_{t\to\infty} P\{qt \leq x\}$.

Proof of Theorem 2*, Part 1. We will deal with each of the three cases (corresponding to the value of θ) separately. When $\theta = 0$ and $0 < \theta < \infty$, Stone's criteria (1963) hold, hence the limiting process

is easily found through the convergence of the infinitesimal expectation and variance. The $\theta = \infty$ case is more difficult because the state space "shrinks."

We omit a discussion of uniqueness and refer readers to Dupuis and Ishii (1993) and Mandelbaum and Pats (1995).

$\theta = 0$: Here the abandonment rate converges to 0. As N grows, the abandonment becomes less significant, and indeed the limiting process is identical to the heavy traffic limit of a sequence of $M/M/N$ queues (Halfin and Whitt 1981). The proof in this case is almost identical to that in Halfin and Whitt (see the proof of Theorem 2) by using Stone's criteria: The state space of the rescaled process q_N becomes dense in $\mathbb{R}$ as $N \to \infty$; the infinitesimal expectation (μ_N) and variance (σ_N^2) are given by

$$\mu_N(x) = \begin{cases} -\dfrac{\lfloor N + \sqrt{N}x \rfloor \mu}{\sqrt{N}} + \dfrac{\lambda_N}{\sqrt{N}}, & x \le 0, \\ -\dfrac{N\mu + \lfloor \sqrt{N}x \rfloor \theta_N}{\sqrt{N}} + \dfrac{\lambda_N}{\sqrt{N}}, & x > 0, \end{cases}$$

$$\sigma_N^2(x) = \begin{cases} \dfrac{\lfloor N + \sqrt{N}x \rfloor \mu}{N} + \dfrac{\lambda_N}{N}, & x \le 0, \\ \dfrac{N\mu + \lfloor \sqrt{N}x \rfloor \theta_N}{N} + \dfrac{\lambda_N}{N}, & x > 0, \end{cases}$$

converging, as $N \to \infty$

$$\lim_{N\to\infty} \mu_N(x) = \begin{cases} -\mu(\beta + x), & x \le 0, \\ -\mu\beta, & x > 0, \end{cases}$$

$$\lim_{N\to\infty} \sigma_N^2(x) = 2\mu.$$

$0 < \theta < \infty$: This case appears in Fleming et al. (1994) as a conjecture without a proof, with slightly different centering. It can be proved either as in the case $\theta = 0$ or using Fleming et al.

$\theta = \infty$: Here the proof is more complex. Stone's criteria does not hold because the state space of the limiting process shrinks to $(-\infty, 0]$, exhibiting reflection at the origin. We circumvent this difficulty as follows: Let $X_N = Q_N - N$, and define two complementary and disjoint subsets of $\mathbb{R}_+$, corresponding to the times X_N spent in $(-\infty, 0]$ or in $(0, \infty)$. Thus via time changes we obtain (from X_N) two processes, each "existing" in a different part of $\mathbb{R}$. We then show that the process "existing" in $(-\infty, 0]$ converges to the proposed limit. This is achieved using the procedure introduced in Mandelbaum and Pats (1995). Because the process X_N makes alternating excursions to $(-\infty, 0]$ ("negative" excursions) and $(0, \infty)$ ("positive" excursions), by showing that the duration of the "negative" excursions is of order $\Omega(1/\sqrt{N})$ and that of the "positive" excursions is $o(1/\sqrt{N})$, we conclude that the time spent by X_N in $(0, \infty)$ becomes "negligible" as $N \to \infty$. The proof is then completed using an Inverse Random Time Change Theorem (see the Appendix in Nguyen 1993).

Proof of Theorem 2*, Part 2. The proof of the interchangeable limits is done through specific calculation of both cases, namely the stationary distribution of the diffusion limits (right-hand side, "Rhs" below) and the weak limit of the stationary distributions (left-hand side, "Lhs" below). Here we also deal separately with the three cases corresponding to the value of θ.

Rhs: First, we find the stationary distribution of the diffusion limits. This is accomplished by using the results of Browne and Whitt (1995, §18.3). They provide a simple procedure for calculating the density function ($f(x)$) of the stationary distribution for diffusion processes that have piecewise continuous parameters; reflecting boundary points, if finite, or inaccessible, if infinite. Following their procedure we obtain

$\theta = 0$:

$$f(x) = \begin{cases} \alpha(\beta)\cdot\beta\cdot\dfrac{\phi(x+\beta)}{\phi(\beta)}, & x \le 0, \\ \alpha(\beta)\cdot\beta\,\exp(-x\beta), & x > 0, \end{cases}$$

$0 < \theta < \infty$:

$$f(x) = \begin{cases} \sqrt{\theta/\mu}\cdot h(\beta\sqrt{\mu/\theta})\cdot w(-\beta\sqrt{\mu/\theta})\cdot\dfrac{\phi(x+\beta)}{\phi(\beta)}, \\ \qquad x \le 0, \\ \sqrt{\theta/\mu}\cdot h(\beta\sqrt{\mu/\theta})\cdot w(-\beta, \sqrt{\mu/\theta})\cdot\dfrac{\phi(x\sqrt{\theta/\mu} + \beta\sqrt{\mu/\theta})}{\phi(\beta\sqrt{\mu/\theta})}, \\ \qquad x > 0, \end{cases}$$

$\theta = \infty$:

$$f(x) = \begin{cases} \dfrac{\phi(x+\beta)}{\Phi(\beta)}, & x \le 0, \\ 0, & x > 0. \end{cases}$$

Remark. When $\theta = 0$, a stationary distribution exists only for positive values of β.

Lhs: Now we find the weak limit (if it exists) of the sequence of stationary distributions $\{q_N(\infty), N = 1, 2, \ldots\}$. Note that these distributions always exist because $\theta_N > 0$. Our discussion is in terms of the sequence of cumulative distribution functions, denoted by $\{F_N\}$ converging to F.

We deal separately with the intervals $x \le 0$ (corresponding to $Q_N(\infty) \le N$ in the original system) and $x > 0$. Given $x \le 0$ there is no queue and therefore no abandonment. Hence the conditional distribution (denoted F^+) is identical to that emerging from a sequence of $M/M/N/N$ queues, namely

$$F^+(x) = \begin{cases} \dfrac{\Phi(x+\beta)}{\Phi(\beta)}, & x < 0, \\ 1, & x \ge 0. \end{cases}$$

This leaves us with determining F on $x > 0$. For the $0 < \theta < \infty$ case we quote the result by Fleming et al. (1994), with a slight adjustment since their rescaling is

$$\bar{q}_N = \frac{Q_N - \lambda_N}{\sqrt{\lambda_N}}.$$

This difference only amounts to a "shift" of the distribution

$$q_N(\infty) = \frac{Q_N(\infty) - N}{\sqrt{N}} = \sqrt{\frac{\lambda_N}{N}}\left[\frac{Q_N(\infty) - \lambda_N}{\sqrt{\lambda_N}} + \frac{\lambda_N - N}{\sqrt{\lambda_N}}\right] \xrightarrow{d} \bar{q}(\infty) - \beta.$$

Therefore, the density of $q(\infty)$ is obtained by "shifting" the density of $\bar{q}(\infty)$ by β, which yields

$$f(x) = \begin{cases} \sqrt{\theta/\mu}\cdot h(\beta\sqrt{\mu/\theta})\cdot w(-\beta, \sqrt{\mu/\theta})\cdot \dfrac{\phi(x+\beta)}{\phi(\beta)}, \\ \qquad x \le 0, \\ \sqrt{\theta/\mu}\cdot h(\beta\sqrt{\mu/\theta})\cdot w(-\beta, \sqrt{\mu/\theta})\cdot \dfrac{\phi(x\sqrt{\theta/\mu} + \beta\sqrt{\mu/\theta})}{\phi(\beta\sqrt{\mu/\theta})}, \\ \qquad x > 0. \end{cases}$$

Now we use this result as an upper and lower bound for the $\theta = 0$ and $\theta = \infty$ cases, respectively.

When $\theta = 0$ we must assume $\beta > 0$, otherwise the sequence is not tight. Denoting $\hat{F}_N = P\{q_N(\infty) \le x \mid q_N(\infty) > 0\}$, we find the limit of this sequence, by "sandwiching" it between two converging sequences with a common limit. The "lower" sequence (bounding from below) is of conditional stationary distributions corresponding to a sequence of $M/M/N$ queues, denoted by $\{F_N^-\}$. According to Halfin and Whitt (1981), this sequence has a limit

$$F^-(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\beta x}, & x \ge 0. \end{cases}$$

As stated above, the "upper" sequence corresponds to a sequence of $M/M/N + M$ queues with $0 < \theta < \infty$, and is denoted by $\{\bar{F}_N\}$. Here we have that

$$\bar{F}(x) = \frac{\Phi(x\sqrt{\theta/\mu} + \beta\sqrt{\mu/\theta}) - \Phi(\beta\sqrt{\mu/\theta})}{1 - \Phi(\beta\sqrt{\mu/\theta})}$$

$$= 1 - \frac{h(\beta\sqrt{\mu/\theta})\phi(x\sqrt{\theta/\mu} + \beta\sqrt{\mu/\theta})}{h(x\sqrt{\theta/\mu} + \beta\sqrt{\mu/\theta})\phi(\beta\sqrt{\mu/\theta})}.$$

By taking $\theta \downarrow 0$ and relying on the asymptotic behavior of $h(t)$ as $t \to \infty$, we get

$$\lim_{\theta\to 0} \bar{F}(x) = 1 - \lim_{\theta\to 0} \frac{x\sqrt{\theta/\mu} + \beta\sqrt{\mu/\theta}}{\beta\sqrt{\mu/\theta}} \exp\left[-\left(x^2\frac{\theta}{\mu} + 2x\beta\right)\Big/ 2\right]$$

$$= 1 - e^{-\beta x}.$$

This has completed the "sandwich." These results, put together, yield the density function for this case:

$$f(x) = \begin{cases} \alpha(\beta)\cdot\beta\cdot\dfrac{\phi(x+\beta)}{\phi(\beta)}, & x \le 0, \\ \alpha(\beta)\cdot\beta\,\exp(-x\beta), & x > 0. \end{cases}$$

Finally, by taking $\theta \to \infty$ in the $0 < \theta < \infty$ case we get that for $\theta = \infty$ all the mass of the distribution is concentrated in $x \le 0$, and $F \equiv F^+$. Therefore, for this case we have

$$f(x) = \begin{cases} \dfrac{\phi(x+\beta)}{\Phi(\beta)}, & x \le 0, \\ 0, & x > 0. \end{cases}$$

Proof of Theorem 3. We start out by showing that $\sqrt{N}\nu_N \xrightarrow{d} [q/\mu]^+$, a result that relies on a corollary by Puhalskii (1994) dealing with first passage times. Most of the notation we use here follows the example in Puhalskii (pp. 951–954), replacing the superscript n with subscript N for the parameters and processes corresponding to a model with N agents. Hence we have

$$Q_N = \{Q_N(t), t \ge 0\}, \qquad A_N = \{A_N(t), t \ge 0\},$$

$$D_N = \{D_N(t), t \ge 0\},$$

as the queue, arrival, and departure processes, respectively.

Let $w_N(t)$ be the virtual waiting time at t:

$$w_N(t) = \inf\{s \ge 0 : D_N(s+t) \ge Q_N(0) + A_N(t) - (N-1)\}.$$

We define rescaled processes

$$X_N(t) = \frac{1}{N}D_N(t), \qquad Y_N(t) = \frac{1}{N}A_N(t), \qquad K_N(t) = \frac{1}{N}Q_N(t),$$

and an additional process Z_N^3 characterized via $w_N(t) = (Z_N^3(t) - t)^+$, or equivalently

$$Z_N^3(t) = \inf\{s \ge 0 : X_N(s) \ge Y_N(t) + K_N(0) - (1 - 1/N)\}.$$

Now introduce

$$X(t) = \mu t \qquad (X'(t) = \mu), \qquad Y(t) = \mu t, \qquad K(0) = 1,$$

and a first passage time

$$Z^3(t) = \inf\{s \ge 0 : X(s) \ge Y(t)\},$$

noting that $Z^3(t) \equiv t$.

Finally, let

$$U^3(t) = q(0) - \mu\beta t + \sqrt{\mu}b(t) - q(t),$$

$$V^3(t) = -\mu\beta t + \sqrt{\mu}b(t) + q(0).$$

From here, applying Puhalskii (1994) and the result of Theorem 2* for the $0 < \theta < \infty$ case, we get

$$\sqrt{N}(Z_N^3 - t) \xrightarrow{d} \frac{q(t)}{\mu},$$

which yields, through continuous mapping

$$\sqrt{N}w_N(t) = \sqrt{N}(Z_N^3(t) - t)^+ \xrightarrow{d} \left[\frac{q(t)}{\mu}\right]^+,$$

completing the proof.

Part 1: Follows immediately from Theorem 2* ($0 < \theta < \infty$ case) where the parameters of the diffusion process q are provided, and the density of $q(\infty)$ is given (in the proof above). Part 2: Note that q_N has a stationary distribution and let $q_N(0)$ have this distribution. Hence, for all $0 \le t \le \infty$, $q_N(t)$ has this distribution. Therefore, $\nu_N(t)$ also has the same distribution, for all $0 \le t \le \infty$. Using the opening result completes the proof.

Proof of Theorem 4. The directions going from the center ($-\infty < \beta < \infty$) outward are by-products of Lemmas 1 and 2 below, as are the explicit expressions for α and Δ. The remaining directions are

dealt with by taking β up to ∞ and down to $-\infty$, using the known asymptotic behavior of $h(t) : h(t) \sim t,\ t \to \infty$.

$\beta \to \infty$: An increase in β represents a decrease in congestion, and therefore α (and Δ) also decreases. Δ is found by upper bounding the fraction abandoning with the fraction blocked in an $M/M/N/N$ queue. Hence, as $\beta \to \infty$,

$$\alpha \le \lim_{\beta\to\infty} w(-\beta, \sqrt{\mu/\theta}) = \lim_{\beta\to\infty} \frac{\sqrt{\mu/\theta}h(-\beta)}{h(\beta\sqrt{\mu/\theta}) + \sqrt{\mu/\theta}h(-\beta)} = 0,$$

$$\Delta \le \lim_{\beta\to\infty}\lim_{N\to\infty} \sqrt{N}P_N\{Ab\} \le \lim_{\beta\to\infty}\lim_{N\to\infty} \sqrt{N}P_N\{Bl\} = \lim_{\beta\to\infty} h(-\beta) = 0.$$

$\beta \to -\infty$: Here we use the reverse argument, bounding from below:

$$\alpha \ge \lim_{\beta\to-\infty} w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) = \lim_{\beta\to-\infty} \frac{\sqrt{\mu/\theta}h(-\beta)}{h(\beta\sqrt{\mu/\theta}) + \sqrt{\mu/\theta}h(-\beta)} = 1,$$

$$\Delta \ge \lim_{\beta\to-\infty} [\sqrt{\theta/\mu}\cdot h(\beta\sqrt{\mu/\theta}) - \beta]\cdot\alpha = \infty.$$

LEMMA 1.

$$\lim_{N\to\infty} P_N\{W > 0\} = \begin{cases} \alpha(\beta), & \theta = 0, \\ w(-\beta, \sqrt{\mu/\theta}), & 0 < \theta < \infty, \\ 0, & \theta = \infty. \end{cases}$$

PROOF. Calculating directly, we have

$$\lim_{N\to\infty} P_N\{W > 0\} = \lim_{N\to\infty} P\{Q_N(\infty) > N\} = P\{q(\infty) > 0\}.$$

Hence, this result is arrived at through simple integration of the densities found in part 2 of the proof of Theorem 2*.

LEMMA 2. *Assume* $0 < \theta < \infty$. *Then*

$$\lim_{N\to\infty} \sqrt{N}P_N\{Ab\} = [\sqrt{\theta/\mu}\cdot h(\beta\sqrt{\mu/\theta}) - \beta]\cdot w(-\beta, \sqrt{\mu/\theta}).$$

PROOF. First, we express $P_N\{Ab\}$ as a function of $P_N\{W > 0\}$ and $P_N\{Bl\}$, whose asymptotic behavior is known (see Jagerman 1974 and Lemma 1). In §5.1 we give the balance equation $P_N\{Ab\} = \theta\cdot E[W]$, which can be rewritten as

$$P_N\{Ab\} = \theta\cdot E[W \mid W > 0]P_N\{W > 0\}.$$

Inserting Riordan's (1962) expression for the conditional expectation, we get

$$P_N\{Ab\} = \left(1 - \frac{1}{\rho_N} + \frac{(\lambda_N\theta)^{N\mu/\theta-1}e^{-\lambda_N/\theta}}{\gamma(N\mu/\theta, \lambda_N/\theta)}\right)P_N\{W > 0\}.$$

Through Palm's (1943) representation we obtain after a few simple manipulations

$$P_N\{Ab\} = \left(1 - \frac{1}{\rho_N} + \frac{P_N\{Bl\}/\rho_N}{P_N\{Bl\}/\pi_N - 1 + P_N\{Bl\}}\right)\cdot P_N\{W > 0\}.$$

Finally, using the connection between π_N and $P_N\{Bl\}$, we get

$$P_N\{Ab\} = \left(1 - \frac{1}{\rho_N} + \frac{P_N\{Bl\}/\rho_N}{(1 - P_N\{Bl\})/(1 - P_N\{W > 0\}) - 1 + P_N\{Bl\}}\right) \times P_N\{W > 0\}.$$

Now multiplying by $\sqrt{N}$ and taking $N \to \infty$

$$\lim_{N\to\infty} \sqrt{N}P_N\{Ab\} = \left(-\beta + \frac{h(-\beta)(1 - w(-\beta, \sqrt{\mu/\theta}))}{w(-\beta, \sqrt{\mu/\theta})}\right)w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right),$$

which completes the proof because $h(x)(1 - w(x, y)) = (1/y)h(-xy)w(x, y)$.

References

Ancker, C.J. Jr., A.V. Gafarian. 1963. Queueing with reneging and multiple heterogeneous servers. *Naval Res. Logist. Quart.* **10** 125–149.

Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. F.J. Kylstra, ed. *Performance '81.* North-Holland, Amsterdam, The Netherlands, 159–179.

Bhattacharya, P.P., A. Ephremides. 1991. Stochastic monotonicity properties of multiserver queues with impatient customers. *J. Appl. Probab.* **28** 673–682.

Borst, S., A. Mandelbaum, M. Reiman. 2000. Dimensioning of large call centers. Submitted for publication ⟨ie.technion.ac.il/serveng⟩.

Boxma, O.J., P.R. de Waal. 1994. Multiserver queues with impatient customers. *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks.* J. Labetoulle and J.W. Roberts, eds. Elsevier, Amsterdam, The Netherlands, 743–756.

Brandt, A., M. Brandt. 1998. On a two-queue priority system with impatience and its application to a call center. Preprint.

——, ——. 1999. On the $M(n)/M(m)/s$ queue with impatient calls. *Performance Evaluation* **35** 1–18.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J.H. Dshalalow, ed. *Probability and Stochastic Series: Advances in Queueing. Theory, Methods, and Open Problems.* CRC Press, Boca Raton, FL, 463–480.

Cleveland, B., J. Mayben. 1997. *Call Center Management on Fast Forward.* Call Center Press, Annapolis, MD.

Dupuis, P., H. Ishii. 1993. SDE's with oblique reflection on nonsmooth domains. *Ann. Probab.* **21** 554–580.

Erlang, A.K. 1909. The theory of probabilities and telephone conversations. *Nyt Tidsskrift Mat.* **B 20** 33–39.

——. 1917. Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Electrotekniker* (Danish) **13** 5–13. English translation. 1917–1918. *P.O. Electr. Engrg. J.* **10** 189–197.

Fleming, P.J., A. Stolyar, B. Simon. 1994. Heavy traffic limit for a mobile phone system loss model. *Proc. 2nd Internat. Conf. Telecommunication Systems, Modeling, and Anal.,* Nashville, TN, 158–176.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–587.

Harris, C.M., K.L. Hoffman, P.B. Saunders. 1987. Modeling the IRS telephone taxpayer information system. *Oper. Res.* **35** 504–523.

Help Desk and Customer Support Practices Report. 1997. Survey results, The Help Desk Institute, SOFTBANK Forums (May).

Hoffman, K.L., C.M. Harris. 1986. Estimation of a caller retrial rate for a telephone information system. *Eur. J. Oper. Res.* **27** 207–214.

Jagerman, D.L. 1974. Some properties of the Erlang loss function. *Bell System Tech. J.* **53** (3) 525–551.

Mandelbaum, A., G. Pats. 1995. State-dependent queues: Approximations and applications. F.P. Kelly and R.J. Williams, eds. *Stochastic Networks.* Springer-Verlag, New York, 239–282.

——, W.A. Massey, M. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems: Theory and Applications (QUESTA)* **30** 149–201.

——, A. Sakov, S. Zeltyn. 2000. Empirical analysis of a call center. Technical report ⟨ie.technion.ac.il/serveng/course/096324⟩.

——, W.A. Massey, M. Reiman, B. Rider. 1999. Time varying multiserver queues with abandonment and retrials. *Teletraffic Engineering in a Competitive World.* P. Key and D. Smith, eds. Elsevier, Amsterdam, The Netherlands.

——, ——, ——, ——, A. Stolyar. 2000. Queue lengths and waiting times for multiserver queues with abandonment and retrials. Submitted to *Selected Proc. 5th INFORMS Telecomm. Conf.*

Nguyen, V. 1993. Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits. *Ann. Appl. Probab.* **3** 28–55.

Palm, C. 1937. Etude des delais d'attente. *Ericsson Technics* **5** 37–56.

——. 1943. Intensitatsschwankungen im fernsprechverkehr. *Ericsson Technics* **44**(1) 1–189.

——. 1953. Methods of judging the annoyance caused by congestion. *Tele* **4** 189–208.

Puhalskii, A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* **19** (4) 946–954.

Riordan, J. 1962. *Stochastic Service Systems.* Wiley, New York.

Stone, C. 1963. Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois J. Math.* **7** 638–660.

Sze, D.Y. 1984. A queueing model for telephone operator staffing. *Oper. Res.* **32** 229–249.

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

Zohar, E., A. Mandelbaum, N. Shimkin. 2000. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. Technical report ⟨ie.technion.ac.il/serveng/course/096324⟩.

The consulting Senior Editor for this manuscript was Michael Harrison. This manuscript was received on October 27, 1999, and was with the authors 684 days for 3 revisions. The average review cycle time was 77 days.