Designing a Call Center with Impatient Customers

O. Garnett* A. Mandelbaum*† M. Reiman ‡

October 8, 1999

ABSTRACT. 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 ($M/M/N/B+M$). Such a model is both rich and analyzable enough to provide information that is practically important for call center managers. We first provide an exact analysis for the $M/M/N/B+M$ 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.

*Davidson Faculty of Industrial Engineering and Management, Technion, Haifa 32000, ISRAEL.

†Research supported by the fund for the promotion of research at the Technion, and by the Technion V.P.R. funds - Smoler Research Fund, and B. and G. Greenberg Research Fund (Ottawa).

‡Bell Laboratories, Murray Hill, NJ 07974, USA.

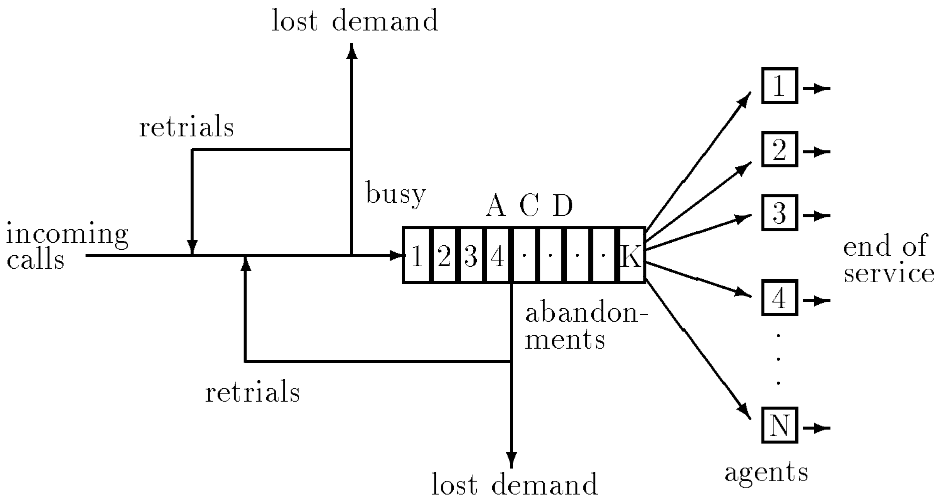
1 Introduction and Motivation

During recent decades there has been an explosive growth in the number of companies providing services via telephone, and in the variety of telephone services provided. The overall challenge in designing and managing a service center is to achieve a balance between operational efficiency and service quality. Service via the telephone requires a relatively short response time (seconds), while the number of daily calls to a large call center can reach tens of thousands. Under such circumstances, one must rely on analytical models to achieve the sought-after balance, and in the present paper we advance the state-of-the-art of such models.

1.1 A simple schematic operational model of a call center appears in Figure 1. In this model a single queue leads to N statistically identical agents. There are incoming calls by customers seeking service at the call center; $K+N$ trunks are connected to an ACD (Automatic Call Distributor), which handles the queue, connects customers to free agents and archives operational data. A customer "arriving" when all the trunks are occupied receives a busy signal. Such a customer might try again later ("retrial") or give up ("lost demand"). A customer who succeeds in getting through at a time when all agents are busy (when there are less than $K+N$ customers within the call center), is required to wait in queue ("music"). If a waiting customer runs out of patience before his service begins, he hangs up and "abandons". After abandoning a customer might try calling again later. In this simple model it is usually assumed that the only parameters over which the call center's manager has control are the number of trunks available $(K+N)$ and the number of agents (N).

A commonly used model for the analysis of call centers with many agents is the $M/M/N/B$ model ($B \leq \infty$, frequently $B = \infty$ is chosen): It is assumed that retrials occur long enough after the original call so as not to affect the Poisson nature of arrivals, and therefore retrials

Figure 1: Schematic operational model of a telephone call center



are ignored; for simplicity abandonments are typically ignored as well. However, as we argue throughout this section, models which ignore abandonments have serious drawbacks. We thus strongly suggest the $M/M/N/B$ model with the addition of exponential patience as a good approximation for call center analysis: it is a simple model, easy to implement, and far superior, as demonstrated below.

1.2 In service centers in general, and service via telephone in particular, customers tend to be impatient, and usually at least some of the customers waiting in queue decide to hang up (abandon) before their service begins. A model that accounts for abandonments is therefore a more accurate portrait of a call center. The effect of adding abandonments is a decrease in congestion, since not all arriving traffic will require service. Consequently, queue lengths and waiting times will be reduced, hence using merely the $M/M/N/B$ model leads to overstaffing.

For models with abandonments we use the notation introduced in Baccelli and Hebuterne [1], where a $G/G/N/B$ model with general abandonments is denoted $G/G/N/B+G$. Table 1 displays some results from an $M/M/N$ model and a corresponding $M/M/N+M$

Table 1: Comparing results for models with/without abandonments
(50 agents, 48 calls per min., 1 min. average service time, 2 min. average patience)

	$M/M/N$	$M/M/N+M$
Fraction abandoning	–	3.1%
Average waiting time	20.8 sec	3.7 sec
Waiting time's 90-th percentile	58.1 sec	12.5 sec
Average queue length	17	3
Agents' utilization	96%	93%

with only 3% abandonments. There is a significant difference in the distributions of waiting time and queue length - in particular, the average wait and queue length are both much shorter when abandonments are taken into account. It should be noted, however, that the performance of systems in such heavy traffic is very sensitive to the staffing level - adding 3 or 4 agents to the model without abandonments would result in performance similar to that displayed for the model with abandonments. 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. Table 1 clearly indicates that in the heavy traffic regime that we propose it is possible to simultaneously achieve high efficiency (agent utilization near 100%) and good service (low, but not negligible abandonment and waiting time).

Remark 1 With abandonments, the "average waiting time" includes both abandoning and served customers. (See Table 2 and §5.2 for a discussion of performance measures.) Therefore, the average waiting time of those served is also of interest, being a service measure of the persistent ("loyal") customers. However, in our example the values of these measures are very close (3.6 sec for those served).

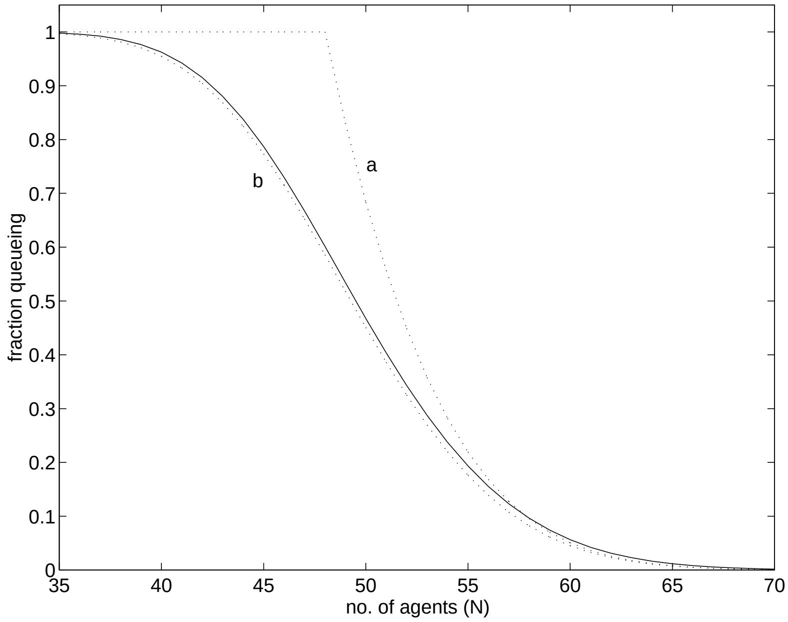
An important advantage of adding exponential abandonments to the $M/M/N$ model is that the new model is more "robust" - there is always a steady state, regardless of the number of agents (N), arrival (λ) and service (μ) rates (for a general criterion see Remark 2.1 in §2 below): this is contrary to the $M/M/N$ case which requires $\lambda < N\mu$ for stability. The robustness becomes crucial when analyzing systems in heavy traffic, which may be (at least temporarily) "overloaded" (i.e. $\lambda \geq N\mu$).

A perhaps less obvious drawback of the $M/M/N/B$ model is that it cannot 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 the waiting times and fraction of customers receiving busy signals, especially since abandonment statistics constitute the only ACD data that unveils customers' perception of service quality. The service level is "three dimensional" - there are three separate aspects to consider - waiting, blocking and abandoning, and the $M/M/N/B$ model provides for only two of them. Indeed, according to the "Help Desk and Customer Support Practices Report, 1997" [18] more than 40% of such call centers set a target for fraction of abandonments, but in most cases this target is not achieved.

1.3 Diffusion limits constitute an important tool for the analysis of queueing models. In this paper we derive diffusion limits for systems operating in heavy traffic. The importance of this tool is three-fold: the diffusion processes are used to derive approximations, they provide necessary insight, and enable us to formulate rules of thumb.

A central result is Theorem 5 that deals with a sequence of $M/M/N+M$ queues in heavy traffic. It is analogous to a result by Halfin and Whitt [19] (Proposition 1, pp. 574) for $M/M/N$ queues. Beyond theoretical interests, both results have practical significance in that they parameterize the regimes that are most suitable, in our opinion, to the operations of large call centers (See Table 3 in §5.4). Indeed, both results give rise to approximations of the fraction of customers having to queue, which are compared here in Figure 2. Note

Figure 2: Approximate fraction queueing in $M(48)/M(1)/N$ (denoted by a) and $M(48)/M(1)/N+M(0.5)$ (b) models vs. $M(48)/M(1)/N+M(0.5)$ exact value (solid)



that for all values of $N \leq \frac{\lambda}{\mu}$ (N, λ and μ being the number of agents, arrival rate and service rate, respectively), the approximation based on the model without abandonments is 1, since this is the overloaded regime in which there is no steady state. For large values of N the systems are underloaded, in which case the approximations coincide because of negligible abandonments. The expression for the fraction of customers queueing can be derived through Method C (see §3), and relying on Remark 4.1 which follows it. It is given by

$$1 - \frac{1 - P\{Bl\}}{1 + P\{Bl\}\left[\frac{50\times 1}{0.5}\left(\frac{0.5}{48}\right)^{50\times 1/0.5} e^{48/0.5}\gamma\left(\frac{50\times 1}{0.5}, \frac{48}{0.5}\right) - 1\right]},$$

$P\{Bl\}$ being the fraction of customers blocked in an $M(48)/M(1)/50/50$ model.
The formula for our approximation appears in (1) below.

Our heavy traffic results also give rise to rules of thumb (see §5.4). The following is a possible scenario in which such rules can be used: A given call center with N agents, service rate μ and arrival rate λ, has "service grade" ζ (high values of ζ correspond to high service levels). There is a forecast of a higher arrival rate $\hat{\lambda}$ 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 the number of agents $N = \hat{N}(\zeta)$ for the holiday shifts. As elaborated on in §5.4, we propose three operational regimes for a call center, determined by the relative importance of service-quality and operational-efficiency. The regimes are: *quality-driven*, where the focus is on service quality, which is manifested by rare waitings and abandonments; *efficiency-driven*, where one emphasizes agents' efficiency, thus the majority of customers wait and a significant fraction abandons; and *rationalized*, where quality and efficiency are balanced to yield busy agents but only a controlled fraction of customers that wait and few that abandon.

Once the manager has defined the operational regime of the call center, representing the desired balance between quality and efficiency, our rules provide $\hat{N}(\zeta)$ (See Table 3). For example, in the rationalized regime, our recommended staffing level is $\hat{N}(\zeta) = \lceil \frac{\hat{\lambda}}{\mu} + \zeta\sqrt{\frac{\hat{\lambda}}{\mu}} \rceil$, where $\zeta = \sqrt{\frac{\mu}{\lambda}}\left(N - \frac{\lambda}{\mu}\right)$. Moreover, at this level, the fraction of delayed and abandoning customers are anticipated to be

$$P\{Wait > 0\} \approx \left[1 + \frac{h(\zeta\sqrt{\mu/\theta})}{\sqrt{\mu/\theta}h(-\zeta)}\right]^{-1} , \tag{1}$$

$$P\{Abandon\} \approx \frac{1}{\sqrt{\hat{N}}}[\sqrt{\theta/\mu} \cdot h(\zeta\sqrt{\mu/\theta}) - \zeta] \cdot \left[1 + \frac{h(\zeta\sqrt{\mu/\theta})}{\sqrt{\mu/\theta}h(-\zeta)}\right]^{-1} .$$

Here $h(x) = \phi(x)/[1-\Phi(x)]$ is the hazard rate of the standard normal distribution, namely

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

Note that the above example manifests economies of scale in the following sense: while increasing N and maintaining the fraction of delayed customers unaltered, server utilization increases and the fraction of customers abandoning as well as the average wait (see (11)) both decrease.

Our analysis actually yields explicit approximations for a wide range of performance measures, for example the fraction of customers waiting beyond a given threshold, the fraction abandoning among those waiting beyond a given threshold, and more (see §5.3 and §5.2).

1.4 As attributed in the sequel, some of our results are motivated or based on previous work by Palm [28][29][30], Riordan [32], Baccelli and Hebuterne [1], Halfin and Whitt [19], and Fleming, Stolyar and Simon [14]. A general overview of models with abandonments appears in Boxma and de Waal [6], with a review of relevant literature. There have been attempts to analyze more complex models, of which we mention a few: Sze [34] compares different approximations for an $M/PH/N + PH$ model (PH stands for Phase Type distribution) with retrials, priorities, and non-stationary arrivals. The results are verified by simulation. In Harris, Hoffman and Saunders [20] and Hoffman and Harris [21] the basic model is $M/M/N$, with the addition of abandonments, 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. In two recent papers by Brandt [7][8], $M/M/N + G$ 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 abandonments and retrials, are described in Mandelbaum, Massey, Reiman and Rider [24], which is based on Mandelbaum Massey and Reiman [23].

1.5 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 [19] to accommodate abandonments (for example, Theorems 5 and 2, Table 3) and waiting times (Theorem 3).

- Revisiting the classical Erlang [12][13] and Palm [29] results, and adapting them to the environment of the modern call center (§3 and §5.1, §5.2).
- Adding support to the growing acknowledgment that insights from diffusion approximations are directly applicable to management practice (§5.3, §5.4).

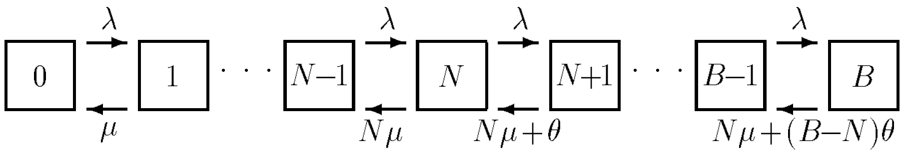
The rest of the paper is organized as follows: In §2 we set up the analytical model and introduce notation. In §3 we develop methods for exact calculations of a wide variety of performance measures. Approximations are derived in §4 through heavy traffic limit theorems. Implementation of the results is discussed in §5.

2 Formulation and Notation

For convenience we review the assumptions and underlying processes of the $M/M/N/B+M$ model. The system has a single queue feeding N independent and statistically identical agents. Customers arrive at rate λ according to a Poisson process, and are served in order of arrival (FCFS). Service times are $\exp(\mu)$ random variables. The patience of each customer (the period of time he is willing to wait in queue before abandoning) is an $\exp(\theta)$ random variable, independent of everything else. The system's capacity is B customers (i.e. at the most $K = B - N$ customers in queue). Customers arriving to a full system are "blocked", and leave the system.

The state of the system at time t is defined as the number of customers in the system (being served or waiting in queue), and is denoted $Q(t)$. Since the service times, customers' patience and interarrival times are all independent exponentials, $Q = \{Q(t), t \geq 0\}$ is a

Figure 3: $\{Q(t), t \geq 0\}$ - Transition diagram



birth-death process with birth and death rates (λ_k and μ_k respectively) given by:

$$\lambda_k = \begin{cases} \lambda \ , & 0 \leq k \leq B-1 \\ 0 \ , & otherwise \end{cases} \quad ; \quad \mu_k = \begin{cases} (N \wedge k)\mu + [k-N]^+\theta \ , & 1 \leq k \leq B \\ 0 \ , & otherwise \end{cases} \tag{2}$$

The transition diagram of this process appears in Figure 3.

Remark 2 1. The process Q always has a steady state, most significant being the case of $B = \infty$ where the process has an infinite state space. This is contrary to the $M/M/N$ model, in which steady state performance measures cannot be calculated in the case of an overloaded system ($\lambda \geq N\mu$).

For customer patience with a general distribution F, and a system with infinite capacity, the criterion for the existence of a steady state (see [1]) is:

$$\lambda(1 - F(\infty)) < N\mu \ ,$$

since the agents must be able to overcome the traffic consisting of customers with infinite patience.

2. Our model assumes that a customer's patience is independent of his place in the queue. This assumption is not unreasonable for service via the telephone in which queues are "invisible" (see Mandelbaum and Shimkin [26]) - usually customers have no information about the queue.

Notations Throughout the paper we use the gamma function, defined by

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

and the incomplete gamma function, denoted $\gamma(x, y)$, which is given by

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

Weak convergence is the standard notion that formalizes approximations of probability distributions ([4] is a standard reference). In our paper, it arises for either convergence of stationary distributions or convergence of stochastic processes. In both cases, we denote weak convergence of a sequence $\{X_N\}$ to X by $X_N \xrightarrow{d} X$, the "d" standing for the terminology "convergence in <u>d</u>istribution".

3 Exact Calculation for the $M/M/N/B + M$ Model

Here we introduce a method for calculating a wide variety of performance measures for an $M/M/N/B + M$ model in steady state. Due to the underlying birth-death structure, such calculations are almost (but not quite, due to numerical issues) trivial. These results are later (see §5) used to derive 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 steady state.

We are interested in a "typical" customer, arriving to the system in steady state (the exact meaning of "typical" will be given momentarily). Let the random variable V be a typical customer's *potential waiting time* (i.e the time he would wait in queue for his

Table 2: Performance measures of the form $E[f(V,X)]$

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

service to commence if his patience was infinite). Let X be this customer's patience (note that X and V are independent), and let W be his *actual* waiting time. It is clear that $W = V \wedge X$. Finally, let $\{Bl\}$ be the event that the customer was blocked, and $\{Ab\}$ the event that he abandoned (i.e $\{Ab\} = \{V > X\}$).

What is meant by a "typical" customer? Consider the sequence $\{w_n, n \in I\!N\}$, where w_n is the potential waiting time of the n-th customer. Let F_w be the stationary distribution of this sequence. Quoting from Baccelli and Hebuterne [1], F_w is also the stationary distribution of the process $\nu(t)$ - 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.

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 2.

Remark 3 1. In this section we use π to denote the stationary distribution of the process $Q(t)$, namely

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

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

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

2. 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 i-th of these representing the period of time the customer spent in the i-th place in queue, before advancing to the $(i-1)$-th (due to end of service or abandonment from the queue in front of him).

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

4. Some important performance measures cannot be calculated 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 even a short period of time t. In such a case one uses

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

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

$$\begin{aligned} E[f(V,X)] &= E[f(V,X)\cdot 1_{\{V>0\}}] + E[f(V,X)\cdot 1_{\{V=0\}}] \\ &= E[f(V,X)\cdot 1_{\{V>0\}}] + E[f(0,X)]\cdot(\pi_B + \sum_{k=0}^{N-1}\pi_k)\ . \end{aligned} \tag{3}$$

For all functions f which seem of interest in our case, $E[f(0,X)]$ evaluates to 0 or 1. Therefore we proceed to calculate the first expression. We present three different methods for performing this calculation, each with its own virtues and drawbacks. These methods are the ones to use in analyzing small call centers. Sections 4 and subsequent sections focus on large systems. Readers can now safely skip to Section 4 without interrupting the flow of reading.

Method A: Conditioning on the number of customers in the queue upon arrival, and substituting the explicit expression given by Riordan [32] (equation (83) on page 111) for $\bar{F}_n(t) = 1 - F_n(t)$, we have

$$E[f(V,X)\cdot 1_{\{V>0\}}] = c\pi_N \sum_{k=0}^{B-N-1}(-1)^k \frac{(\lambda/\theta)^k}{k!}\ I(k) \sum_{n=k}^{B-N-1}\frac{(\lambda/\theta)^{n-k}}{(n-k)!}\ , \tag{4}$$

where

$$\begin{aligned} I(k) &= \frac{1}{c+k}\int_0^\infty\int_0^\infty f(t,x)(c+k)\theta e^{-(c+k)\theta t}\theta e^{-\theta x}\,dt\,dx \\ &= \frac{1}{c+k}\ E[f(X_k,X)]\ , \end{aligned} \tag{5}$$

$c = N\mu/\theta$, and X_k is a random variable independent of X, with an $\exp(N\mu + k\theta)$ distribution.

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

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

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

to eliminate one sum, we arrive at

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

where

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

Here calculating the values of $J(n)$ tends to be more costly since 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 too). Since 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 through Riordan [32], and solving the more general case of any buffer size B, we arrive at the function f_V^+, where $\frac{f_V^+}{P\{V>0\}}$ is a density function, given by

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

Now we are left with the evaluation of the double integral

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

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 are necessary for the infinite buffer case:

Remark 4 1. Solving the steady state equations also involves an infinite sum. A solution is given by Palm [29], expressing the stationary distribution as a function of the easily calculated blocking probability in an $M/M/N/N$ system (denoted here $P\{Block\}$), with the same arrival and service rates:

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

where

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

2. For $B = \infty$ the density function f_V^+ given here becomes a special case of the result by Baccelli and Hebuterne [1] for an $M/M/N+G$ model, with patience distribution F, namely:

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

4 Diffusion Approximations and Operational Regimes

The main theme of the present paper is approximating performance measures, thus gaining insight as to their dependence on the model's parameters. The first step is to approximate the process Q and it's stationary distribution π, which will be denoted $Q(\infty)$. From our theoretical results, which are also supported by prevailing practice, it follows that large telephone call centers are capable of delivering high service-level while still operating under high utilization. This justifies our focus on approximations through heavy traffic limits, as $N \uparrow \infty$. Furthermore, most large telephone call centers have enough trunks to essentially eliminate customer blocking (busy-signal). We therefore assume, for the remainder of this section, that the buffer is infinite ($B = \infty$).

Remark 5 We would like to emphasize that in no way do we advocate here the practice of unconditional "no-busy-signal". Indeed, a busy-signal is the simplest way to convey congestion. For toll-free services it is, moreover, the cheapest with respect to operational costs. Thus a tradeoff between blocking and abandonments, in the spirit of [5], suggests itself. We leave it as a worthwhile direction for future research.

Consider the sequence of processes $\{Q_N, N = 1, 2, \ldots\}$, where $Q_N = \{Q_N(t), t \geq 0\}$ is the queue length process associated with an $M/M/N + M$ model (N agents). The subscript N is added to our notation to indicate the parameters of the N-th system.

We now characterize the dependence of the model's parameters on N. Specifically, we are interested in a sequence in which $\lambda_N \uparrow \infty$ as $N \uparrow \infty$ and $\mu_N \equiv \mu$, which corresponds to scaling up the call center ($N \uparrow$) to fit its load ($\lambda_N \uparrow$) while assuring that service rate (μ) does not vary with size (N).

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. (One should note that the average wait in queue is linearly related to $P_N\{Ab\}$ through $P_N\{Ab\} = \theta_N \cdot E[W]$; see (11) below).

Most telephone call centers try to avoid a high percentage of abandonments, without overstaffing. This usually translates to operating with a non negligible fraction of customers having to queue, and a small fraction of abandonments. Following is an analytic result in which we introduce the notion of *traffic intensity* defined by $\rho_N = \frac{\lambda_N}{N\mu}$.

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

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

The proof of this theorem and other selected results quoted throughout the paper appear (in full or outlined) in the Appendix.

Based on Theorem 1 it seems clear that, from the point of view of abandonments, 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 higher than usually desired. One may conclude from Theorem 1 that $\rho_\infty = 1 + \epsilon$ would be appropriate, since this yields a limiting abandonment probability of $\epsilon/(1+\epsilon)$. However, from the point of view of the agents' *utilization* (i.e. the fraction of time they spend answering calls, given by $\frac{\lambda_N(1-P_N\{Ab\})}{N\mu}$) the maximum limiting utilization is already achieved with $\rho_\infty = 1$. Thus, $\rho_\infty = 1$ is a special balance point between the call center's efficiency and quality. Slightly underloaded ($\rho = 1 - \epsilon$) or overloaded ($\rho = 1 + \epsilon$) call centers can be regarded as perturbations of this regime. Therefore, focusing on a single regime, we restrict ourselves to $\rho_\infty = 1$. The high accuracy of the approximations that this regime yields is apparent from Figure 2. More research is called for to explore the $\rho_\infty = 1 + \epsilon$ regime.

The restriction to $\rho_\infty = 1$ is consistent with the work of Halfin and Whitt [19] 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 non degenerate). Since an $M/M/N+M$ model with very patient customers ($\theta_N \downarrow 0$) is "close" to an $M/M/N$ model (this will be formally supported momentarily by Theorem 2), we also restrict ourselves to the case of $\lim_{N\to\infty} \sqrt{N}(1-\rho_N) = \beta$, $-\infty < \beta < \infty$. The discussion later in §5.4, formalized by Theorem 5, strongly supports our contention that this is indeed the regime of interest. (Note that in [19] the result covers only $\beta > 0$, since otherwise there is no steady state).

As for the patience parameter, we will be assuming $lim_{N\to\infty}\theta_N = \theta$, where $0 \le \theta \le \infty$. This naturally motivates three regimes:

- $\theta = 0$: "Patient" customers.
- $\theta = \infty$: "Impatient" customers.
- $0 < \theta < \infty$: "Balanced" abandoning.

The "balanced" case has been analyzed by Fleming, Stolyar and Simon [14]. This seems to be an appropriate regime - it is reasonable to assume the customers' patience is independent of the number of agents manning the call center, especially since this number is usually unknown to the caller. It could, however, be argued that when calling large call centers, customers tend to expect prompt service, and although they do not know the exact number of agents on shift, they have a qualitative notion of their number. Nevertheless, our following theorem supports the choice of the "balanced" regime, as will be explained shortly.

Consider the sequence of stochastic processes $\{q_N\}$ where

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

jointly with their stationary distributions $q_N(\infty)$.

Following is a theorem on the convergence of the sequence $\{q_N\}$, which is obtained from $\{Q_N\}$ through centering and rescaling. Specifically, 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 non-trivial *continuous* limiting process q. The latter will be used to approximate our original birth-death processes $\{Q_N\}$, via $Q_N \overset{d}{\approx} N + \sqrt{N}q$. The limit theorem supports three types of limits that correspond to the three types of customers' behavior. The mathematical details of the theorem are not a prerequisite for following its consequences, which are explained immediately after the theorem.

Theorem 2 *Assume that* $\lim_{N\to\infty} \sqrt{N}(1-\rho_N) = \beta \ , \ -\infty < \beta < \infty$*. 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*

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

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

$$\theta = \infty \quad : \quad \begin{cases} dq(t) = -\mu(\beta + q(t))dt + \sqrt{2\mu}\, db(t) - dY(t) \\ Y(0) = 0 \ , \ Y \uparrow 0 \ , \ q\, dY = 0 \end{cases}$$

(Here b denotes a standard Brownian Motion)

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

Remark 6 An equivalent representation, which specifically displays the limits being "interchanged" is: $\lim_{N\to\infty} \lim_{t\to\infty} P\{q_N(t) \leq x\} = \lim_{t\to\infty} \lim_{N\to\infty} P\{q_N(t) \leq x\}$.

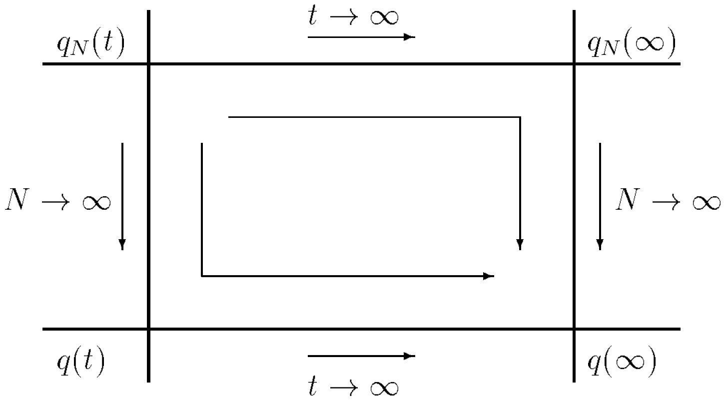
From the first part of Theorem 2, we see that there are three different limiting diffusion processes, according to the regimes defined by the value of θ. In the "patient" customers case ($\theta = 0$) the limit is the same as in [19], emerging from the heavy traffic limit of a sequence of $M/M/N$ queues. This means that the abandonment phenomena has been "lost" during the limit process. Similarly, for the "impatient" customers case ($\theta = \infty$), through the limiting process the queue has been lost - arriving at the same limit as that for a sequence of $M/M/N/N$ loss systems (based on [25]). This leads to the conclusion, also supported by computations (see [15]), that:

- With extremely patient customers, it is reasonable to use the $M/M/N$ model.
- With very impatient customers, the $M/M/N/N$ model should be used.

The limit of the "balanced" case ($0 < \theta < \infty$) was conjectured (with a slightly different centering) by Fleming, Stolyar and Simon [14], and proof was given for the weak limit of the stationary distributions (i.e. $q(\infty)$). Overall, this case seems the most fitting as an approximation to the original $M/M/N + M$ system, and will be used as such from this point on.

The second part of Theorem 2 is important since such an interchange of limits is not automatic. For most applications, one is ultimately interested in approximations for the stationary distribution, therefore it is important to know that both "paths" leading to this distribution, as depicted in Figure 4, coincide. Theorem 4 in the sequel uses the approximation for the "balanced" case.

Figure 4: Arriving at the Stationary Distribution $q(\infty)$



The usefulness of Theorem 2 to call center managers, as presently stated, is limited, since most of the information it provides is about the system and not the service. To get more information about the service being offered, it is necessary to examine the potential waiting times or, equivalently, the virtual waiting time process (as stated in §3, the limits of these processes coincide, as time increases indefinitely). An "Invariance Principle" result by Puhalskii [31] enables us to establish a simple relationship between the diffusion limits of the queue process and the virtual waiting time process (for $0 < \theta < \infty$). To this end, let ν_N denote the virtual waiting time process of the N-th system.

Theorem 3 *Assume that* $\lim_{N\to\infty}\sqrt{N}(1-\rho_N)=\beta$, *for some* $-\infty<\beta<\infty$, *and that* $\theta_N\equiv\theta$. *If* $q_N(0)\xrightarrow{d}q(0)$ *then*

$$\sqrt{N}\nu_N\xrightarrow{d}\left[\frac{q}{\mu}\right]^+ .$$

This central result can be motivated heuristically as follows. If there are idle agents, the virtual waiting time is 0, otherwise the queue length 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) + $o(N)$ (abandonments; 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 $\approx\left[\frac{q}{\sqrt{N}\mu}\right]^+$.

5 Implementation

In §3 and §4 we have reported a variety of results concerning the $M/M/N/B+M$ model. Since we believe that this model should replace the $M/M/N/B$ and $M/M/N$ models, commonly used in call center analysis, we must shed some light on how to use and interpret these results. The context of the discussion will be that of managing a large call center in heavy traffic ("heavy" such that abandonments are not negligible). First we discuss the issue of estimating the values of the model's parameters. Later we suggest which performance measures should be used by call center managers to define the service level. Finally we discuss the use of approximations and derive "rules of thumb".

5.1 Estimating the Parameters

In order to use the model and the results introduced, it is necessary to set the values of the different parameters. Some of the parameters are fully controlled by the call center's manager - the number of agents on shift, and the number of trunks leading into the ACD. Arrival and service rates are usually estimated from historical ACD data. As discussed in §3, in the case of a time varying arrival rate, 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 - the amount of time they spent waiting in queue. There are statistical methods to deal with such censored samples which will not be discussed here. 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. To this end we suggest here two methods for estimating the average patience. The first is based on the following balance equation:

$$\theta \cdot E[\text{no. of customers in queue}] = \lambda P\{Ab\} \ . \tag{10}$$

This equation describes the steady state balance between the rate customers abandon the queue (left hand side) and the rate abandoning customers (i.e - customers who will eventually abandon) enter the system.
Through Little's theorem ($\lambda \cdot E[W] = E[\text{no. of customers in queue}]$), we obtain an alternative equation

$$\theta \cdot E[W] = P\{Ab\} \ . \tag{11}$$

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 (10) and (11) hold under *exponential* patience.

A second more general approach is to calculate any performance measure (using methods from §3) and compare the result to the value derived from ACD data. The goal is to calibrate the patience parameter until these estimates closely match. One advantage of this method is the flexibility in choosing the performance measure being matched, which might depend on the given ACD data. Furthermore, this calibration represents a form of validation of the model's assumptions, and can compensate for discrepancies. We note however, that choosing two performance measures is likely to give rise to two estimates. Further reaseach is needed to establish how such estimates can be combined to achieve a third, "better", estimate.

5.2 The Service Level

As we have stated earlier, the abandonment phenomena is extremely important to a call center's manager. This does not imply that the only performance measure of interest is the fraction of customers abandoning the queue. There are many 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.

We suggest three basic performance measures, representing the three major aspects of service at a call center, as described in §1 :

- $P\{Bl\}$ - Fraction of customers blocked.
- $P\{W > t_1 \ ; \ \neg Ab \ ; \ \neg Bl\}$ - Fraction of customers served, with "long" waiting periods (more than t_1). These are "loyal" customers - they did not abandon, who "suffered" a long waiting time.
- $P\{Ab \ ; \ W > t_2\}$ - Fraction of customers abandoning, discounting "unworthy" abandoning customers who were not willing to wait even a short period (t_2 small).

These measures represent the customers who were willing to make some minimal effort to reach the call center's agents, but received poor service or none at all. A call center's manager should aim to keep their numbers low - these are customers he wants, but is in danger of losing.

It is desirable to have a single measure with which to evaluate the service given at a call center. Such a performance measure can then be used to set a service goal for the call center, usually used to set the number of agents needed on shift according to the expected traffic. A single measure can be constructed as a weighted sum of measures similar to those presented above (for a resulting "grade" between 0 and 1, lower grades indicating better service):

$$\neg SL = a_1 P\{Bl\} + a_2 P\{W > t_1\ ;\ \neg Ab\} + a_3 P\{Ab\ ;\ W > t_2\}\ .$$

Here $a_1 + a_2 + a_3 = 1$.

Even in the case of the relatively simple model presented in this paper, exact evaluation of such a measure is "difficult", in the sense that producing a table of values, especially if inverse calculations are needed (i.e. finding the values of one or more parameters for a given value of the performance measure), cannot always be accomplished in real time. In such cases approximations can be beneficial, as discussed below. Specifically, our results below lead to the following heavy traffic approximation (assuming $P\{Bl\} \approx 0$):

$$\neg SL =$$

$$a_2 w(-\beta, \sqrt{\mu/\theta}) \cdot \frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta t_1})} \cdot e^{-\theta t_1} +$$

$$a_3 w(-\beta, \sqrt{\mu/\theta}) \cdot \frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta t_2})} \cdot \left(\frac{\Psi(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta t_2})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta t_2})} - 1 \right) \cdot e^{-\theta t_2}\ .$$

Here $a_2 + a_3 = 1$. (For the definitions of Ψ, w, and h see (12) below).

5.3 Approximations

Approximations can be used to overcome computational difficulties, 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 §5.4 bellow). Combining the diffusion approximation for the virtual waiting time process (see 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 performance measures we are calculating assume the system is in steady state, therefore we are interested in approximating the stationary distribution of the virtual waiting time. Such an approximation is derived through the following theorem:

Theorem 4 *Let $v = \left[\frac{q}{\mu}\right]^+$, where q solves the stochastic differential equation that corresponds to $0 < \theta < \infty$ in Theorem 2 ("balanced abandoning"), and assume that $\theta_N \equiv \theta$. Then:*

1. *$v(\infty) = \lim_{t\to\infty} v(t)$ has the distribution function F_v given by:*

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

2. *$\sqrt{N}\nu_N(\infty) \xrightarrow{d} v(\infty)$.*

Note that in this theorem we are again justifying interchangeable limits, now concerning the sequence of virtual waiting time processes $\{\nu_N\}$. Since $\sqrt{N}\nu_N(\infty) \xrightarrow{d} v(\infty)$, the approximation we use is $V \stackrel{d}{=} \nu_N(\infty) \stackrel{d}{\approx} v(\infty)/\sqrt{N}$, which translates into $F_V(x) \approx F_v(\sqrt{N}x)$.

Example: Assume a performance measure which 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,

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

Following are the resulting approximations for several performance measures:

$$\begin{aligned} P\{W > 0\} &\approx w(-\beta, \sqrt{\mu/\theta}) \\ P\{Ab|W > 0\} &\approx 1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})} \\ P\{Ab\} &\approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right] \cdot w(-\beta, \sqrt{\mu/\theta}) \\ E[W] &\approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right] \cdot \frac{w(-\beta, \sqrt{\mu/\theta})}{\theta} \end{aligned}$$

$$
\begin{aligned}
P\{W > t\} &\approx w(-\beta, \sqrt{\mu/\theta}) \cdot \frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta t})} \cdot e^{-\theta t} \\
P\{Ab|W > t\} &\approx 1 - \frac{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta t})}{\Psi(\beta\sqrt{\mu/\theta}+\sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta t})}
\end{aligned}
$$

Here

$$
w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1} \ , \ h(x) = \frac{\phi(x)}{1 - \Phi(x)} \ , \ \Psi(x, y) = \frac{\phi(x)}{1 - \Phi(x + y)} \tag{12}
$$

Remark 7 Along the same lines, we have developed further useful approximations, notably for $E[W|W > t]$. These have been omitted due to excessive "bulk".

5.4 Rules of Thumb

It is important for a call center's manager to be able to anticipate the impact of changes on the service level. Such a change could be 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 §5.3 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.

Continuing our discussion for the "balanced abandoning" case, we have the following result, analogous to the result by Halfin and Whitt [19] that concerns $M/M/N$ queues.

Theorem 5 *Assume that $\theta_N \equiv \theta$. Then*

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

if and only if

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

if and only if

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

in which case

$$\begin{aligned}\alpha &= w(-\beta, \sqrt{\mu/\theta}) \\ \Delta &= [\sqrt{\theta/\mu} \cdot h(\beta\sqrt{\mu/\theta}) - \beta] \cdot \alpha \ .\end{aligned}$$

(Here w and h are as in (12) above).

Remark 8 This result holds at the "extremes" as well, namely

$$\beta = -\infty \quad \textit{iff} \quad \alpha = 1 \quad \textit{iff} \quad \Delta = \infty \quad \text{and} \quad \beta = \infty \quad \textit{iff} \quad \alpha = 0 \quad \textit{iff} \quad \Delta = 0 \ .$$

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 β is *not* restricted to be positive.

In light of Theorem 5 and Theorem 1 we now introduce three regimes of operation, with three matching rules of thumb (see Table 3), which tie the staffing level to the *offered load* defined by $R \equiv \frac{\lambda}{\mu}$. (The entity R, measured in "Erlangs" in telecommunications, is a dimensionless quantity that represents the amount of work, measured in units of time, that is added to the system per unit of time):

- "Quality-driven": In such call centers the staffing level is greater than the offered load - in large call centers this translates to negligible abandonments, and negligible waiting. For the analysis of such a call center it is reasonable to use the $M/M/N$ model.

- "Efficiency-driven": Here the staffing level is less than the offered load, there will be a significant fraction of abandonments, and high fraction of waiting.

- "Rationalized": We believe that most large call centers doing telemarketing, customer support, business, or providing information aim to operate in this regime in which abandonments are few, and the fraction of customers having to wait in queue is not high. (Emergency call centers, on the other hand, would strive to be quality-driven.)

A quantitative relation between staffing levels and the offered load follows from Theorem 5. The theorem naturally defines three regimes - that of the non degenerate, "interesting" limiting behavior, and the two "extremes" referred to in Remark 8. These respectively correspond to the rationalized, efficiency-driven ($\beta = -\infty$ case) and quality-driven ($\beta = \infty$ case) regimes.

The staffing level for the rationalized regime is derived directly from $\rho_N \sim 1 - \beta/\sqrt{N}$, $-\infty < \beta < \infty$ (see sections 1 and 2 in Whitt [35] for a detailed discussion). The "extremes" of Theorem 5 only set bounds for the staffing levels. Any staffing level such as $N = \lceil R \pm \epsilon \cdot R^a \rceil$ 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).

The "guidelines" in Table 3 follow directly from Theorem 5, except for the fraction abandoning ($P\{Ab\}$) in the efficiency-driven regime which involves Theorem 1.

Table 3: Rules of thumb

Operational regime	Staffing level	Guidelines
Quality-driven	$N = \lceil R \cdot (1+\epsilon) \rceil$, $\epsilon > 0$	$P\{W > 0\} \to 0$, $P\{Ab\} = o(1/\sqrt{N})$
Efficiency-driven	$N = \lceil R \cdot (1-\epsilon) \rceil$, $\epsilon > 0$	$P\{W > 0\} \to 1$, $P\{Ab\} \to \epsilon$
Rationalized	$N = \lceil R + \zeta\sqrt{R} \rceil$, $\infty > \zeta > -\infty$	$P\{W > 0\} \to \alpha(\zeta)$, $P\{Ab\} \sim \frac{\Delta(\zeta)}{\sqrt{N}}$

Note that

$$\alpha(\zeta) = w(-\zeta, \sqrt{\mu/\theta}) \quad , \quad \Delta(\zeta) = [\sqrt{\theta/\mu} \cdot h(\zeta\sqrt{\mu/\theta}) - \zeta] \cdot \alpha(\zeta) \ .$$

Following the result of Theorem 5, and continuing in the spirit of Whitt [35], we suggest ζ (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 his call center, he can determine the service grade and use the appropriate rule of thumb.

We conclude by *revisiting the scenario from* §1: 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 ζ. 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 rule of thumb $\zeta \approx \sqrt{\frac{\mu}{\lambda}} N \left(1 - \frac{\lambda}{N\mu}\right)$, we get $\hat{N} = \lceil \frac{\hat{\lambda}}{\mu} + \zeta\sqrt{\frac{\hat{\lambda}}{\mu}} \rceil$. Moreover, the anticipated holiday performance is:

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

2. Fraction abandoning: $P\{Ab\} \approx \Delta(\zeta)/\hat{N}$.

Acknowledgments

A.M. thanks Prof. I. Meilijson of Tel-Aviv University, and Prof. 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.

Appendix

Following are outlines for the proofs of Theorems 1-5. For more details see [16].

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 $N \uparrow \infty$, 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 1 (see Remark 2.1 for a more general condition), 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 [2]:

(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 [22] (p.538), that

$$P_N\{Bl\} \sim \left[\frac{\rho_\infty}{\rho_\infty - 1} - \frac{\rho_\infty}{(\rho_\infty - 1)^3}\frac{1}{N} + \frac{2\rho_\infty^2 + \rho_\infty}{(\rho_\infty - 1)^5}\frac{1}{N^2}\right]^{-1} \tag{13}$$

We deal with $\lambda_N = N\mu \cdot \rho_\infty + o(N)$ as follows. Choose $0 < \varepsilon < \rho_\infty - 1$, and define $\lambda_N^+ = N\mu \cdot (\rho_\infty + \varepsilon)$, $\lambda_N^- = N\mu \cdot (\rho_\infty - \varepsilon)$. 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 $\varepsilon \downarrow 0$ 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: Since $P_N\{Ab\}$ with $\rho_\infty \leq 1$ must be smaller than with $\rho_\infty = 1 + \varepsilon$ for any $\varepsilon > 0$ we have that for $\rho_\infty \leq 1$

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

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 ([33]) hold, hence the limiting process is easily found through the convergence of the infinitesimal expectation and variance. The $\theta = \infty$ case is more difficult since the state space "shrinks".

We omit discussion of uniqueness and refer readers to Dupuis [11] and Mandelbaum and Pats [25] .

$\underline{\theta = 0}$: Here the abandonment rate converges to 0. As N grows, the abandonments become less significant, and indeed the limiting process is identical to the heavy traffic limit of a sequence of $M/M/N$ queues ([19]). The proof in this case is almost identical to that in Halfin and Whitt [19], by using Stone's criteria (see proof of Theorem 2 in [19]): The state space of the rescaled process q_N becomes dense in $\mathbb{R}$ as $N \uparrow \infty$; The infinitesimal expectation (μ_N) and variance (σ_N^2) are given by

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

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

converging, as $N \uparrow \infty$

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

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

<u>$0 < \theta < \infty$</u> : This case appears in Fleming, Stolyar and Simon [14] as a conjecture without a proof, with slightly different centering. It can be proved either as in the case $\theta = 0$ or using [14].

<u>$\theta = \infty$</u> : Here the proof is more complex. Stone's criteria does not hold since 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 [25]. Now since 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 $N \uparrow \infty$. The proof is then completed using an Inverse Random Time Change theorem (see Appendix in Nguyen [27]).

Proof of Theorem 2, part 2 :

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 ("Lhs"). Here too we 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 using results by Browne and Whitt [9] (section 18.3). They provide a simple procedure for calculating the stationary distribution's density function ($f(x)$) of diffusion processes which have piecewise continuous parameters; reflecting boundary points, if finite, or inaccessible, if infinite. Following their procedure we obtain:

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

$$0 < \theta < \infty \quad : \quad f(x) = \begin{cases} \sqrt{\theta/\mu}\cdot h(\beta\sqrt{\theta/\mu})\cdot w(-\beta,\sqrt{\mu/\theta})\cdot\frac{\phi(x+\beta)}{\phi(\beta)}\ , & x \le 0 \\ \sqrt{\theta/\mu}\cdot h(\beta\sqrt{\theta/\mu})\cdot w(-\beta,\sqrt{\mu/\theta})\cdot\frac{\phi(x\sqrt{\theta/\mu}+\beta\sqrt{\mu/\theta})}{\phi(\beta\sqrt{\mu/\theta})}\ , & x > 0 \end{cases}$$

$$\theta = \infty \quad : \quad f(x) = \begin{cases} \frac{\phi(x+\beta)}{\Phi(\beta)}\ , & x \le 0 \\ 0\ , & x > 0 \end{cases}$$

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

Lhs: Now we find the weak limit (if exists) of the sequence of stationary distributions $\{q_N(\infty),\ N = 1, 2, \ldots\}$. Note that these distributions always exist. Our discussion is in terms of the sequence of cumulative distribution functions, denoted $\{F_N\}$, 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 abandonments. Hence the conditional distribution (denoted F^+) is identical to that emerging from a sequence of $M/M/N/N$ queues, namely

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

This leaves us with determining F on $x > 0$.

For the $0 < \theta < \infty$ case we quote the result by Fleming, Stolyar and Simon [14], 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

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

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 $\hat{F}_N = P\{q_N(\infty) \le x | q_N(\infty) > 0\}$, we now 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 $\{F_N^-\}$. According to Halfin and Whitt [19] this sequence has a limit

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

As stated above, The "upper" sequence corresponds to a sequence of $M/M/N+M$ queues with $0 < \theta < \infty$, and is denoted $\{\bar{F}_N\}$. 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 $t \uparrow \infty$, we get

$$\lim_{\theta \to 0} \bar{F}(x) = 1 - \lim_{\theta \to 0} \frac{x\sqrt{\theta/\mu} + \beta\sqrt{\mu/\theta}}{\beta\sqrt{\mu/\theta}} e^{-(x^2\frac{\theta}{\mu} + 2x\beta)/2} = 1 - e^{-\beta x} \ .$$

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

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

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

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

Proof of Theorem 3 :

This result relies on a corollary by Puhalskii [31] dealing with first passage times. Most of the notation we use here follows the example in [31] (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:

$$Q_N = \{Q_N(t), t \geq 0\} \ , \ A_N = \{A_N(t), t \geq 0\} \ , \ D_N = \{D_N(t), t \geq 0\} \ ,$$

as the queue, arrival and departure processes, respectively.

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

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

We define rescaled processes

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

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

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

Now introduce

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

and a first passage time

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

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

Finally, let

$$\begin{aligned} U^3(t) &= q(0) - \mu\beta t + \sqrt{\mu} b(t) - q(t) \ , \\ V^3(t) &= -\mu\beta t + \sqrt{\mu} b(t) + q(0) \ . \end{aligned}$$

From here, applying [31] and the result of Theorem 2 for the $0 < \theta < \infty$ case we get

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

which yields, through continuous mapping

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

completing the proof.

Proof of Theorem 4 **:**

Referring to Figure 4 (substituting "q" with "ν") we have that Part 1 deals with the "path" going down and then right, and Part 2 with that going right and then down.

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 $0 \le t \le \infty$. Now using the result of Theorem 3 the proof is complete.

Proof of Theorem 5 **:**

The directions going from the center ($-\infty < \beta < \infty$) outward are by-products of Lemmas 1 and 2, 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 $h(t)$: $h(t) \sim t$, $t \uparrow \infty$.

<u>$\beta \uparrow \infty$:</u>

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

$$\begin{aligned}
\alpha &\le \lim_{\beta\to\infty} w(-\beta, \sqrt{\mu/\theta}) = \lim_{\beta\to\infty} \frac{\sqrt{\mu/\theta}\,h(-\beta)}{h(\beta\sqrt{\mu/\theta}) + \sqrt{\mu/\theta}\,h(-\beta)} = 0 \ , \\
\Delta &\le \lim_{\beta\to\infty}\lim_{N\to\infty} \sqrt{N} P_N\{Ab\} \le \lim_{\beta\to\infty}\lim_{N\to\infty} \sqrt{N} P_N\{Bl\} = \lim_{\beta\to\infty} h(-\beta) = 0 \ .
\end{aligned}$$

<u>$\beta \downarrow -\infty$:</u>

Here we use the reverse argument, bounding from below:

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

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

Lemma 1

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

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 $0 < \theta < \infty$

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

proof:

First we express $P_N\{Ab\}$ as a function of $P_N\{W > 0\}$ and $P_N\{Bl\}$, whose asymptotic behavior is known ([22] and Lemma 1).

In §5.1 we give the balance equation $P_N\{Ab\} = \theta \cdot E[W]$, which can be rewritten as

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

Inserting Riordan's [32] expression for the conditional expectation we get

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

Through Palm's [29] representation (see Remark 1 at the end of §3) we obtain after a few simple manipulations

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

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

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

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

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

which completes the proof since $h(x)(1 - w(x,y)) = \frac{1}{y} h(-xy) w(x,y)$.

References

[1] Baccelli, F., Hebuterne, G. (1981). On queues with impatient customers. In: *Performance '81,* ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 159-179.

[2] Bhattacharya, P.P., Ephremides, A. (1991). Stochastic monotonicity properties of multiserver queues with impatient customers. *J. Appl. Probab.* **28,** 673-682.

[3] Blumenthal, R.M., Getoor, R.K., (1968). *Markov Processes and Potential Theory.* Academic Press.

[4] Billingsley, P., (1968). *Convergence of Probability Measures.* Wiley, NY.

[5] Borst, S., Mandelbaum, A., Reiman, M., (1998). Dimensioning of large call centers. *Preprint* .

[6] Boxma, O.J., de Waal, P.R. (1994). Multiserver queues with impatient customers. *ITC* **14,**

[7] Brandt, A., Brandt, M., (1997). On the $M(n)/M(m)/s$ queue with impatient calls. *Preprint.*

[8] Brandt, A., Brandt, M., (1998). On a two-queue priority system with impatience and its application to a call center. *Preprint.*

[9] Browne, S., Whitt, W., (1995). Piecewise-linear diffusion processes. *Probability and Stochastic Series: Advances in Queueing. Theory, Methods, and Open Problems.* 463-480. Ed. J.H. Dshalalow, CRC Press.

[10] Chen, H., Mandelbaum, A., (1991). Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *The Annals of Probability,* Vol. **19** , No. **4** , 1463-1519.

[11] Dupuis, P., Ishii, H., (1993). SDE's with oblique reflection on nonsmooth domains. *The Annals of Probability,* Vol. **21** , 554-580.

[12] Erlang, A.K., (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift Mat.* **B 20** , 33-39.

[13] Erlang, A.K., (1917). Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Electroteknikeren* (Danish) **13** , 5-13. English translation: *P.O. Elec. Eng. J.* **10** , 189-197, 1917-1918.

[14] Fleming, P.J., Stolyar, A., Simon, B., (1994). Heavy traffic limit for a mobile phone system loss model. *Proc. of 2nd Int'l Conf. on Telecomm. Syst. Mod. and Analysis*, Nashville, TN.

[15] Garnett, O., (1998). Designing a telephone call center with impatient customers. *Msc. thesis,* Technion - Israel Institute of Technology.

[16] Garnett, O., Mandelbaum, A., Reiman, M., (1999). Designing a call center with impatient customers. *Technical Report,* Technion, Haifa.

[17] Grimmett, G.R., Stirzaker, D.R., (1992). *Probability and Random Processes.* Ed. 2. Clarendon Press.

[18] Help Desk and Customer Support Practice Report; May 1997 survey results. The Help Desk Institute, SOFTBANK Forums, 1997.

[19] Halfin, S., Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29,** 567-587.

[20] Harris, C.M., Hoffman, K.L., Saunders, P.B., (1987). Modeling the IRS telephone taxpayer information system. *Operations Research* **35,** 504-523.

[21] Hoffman, K.L., Harris, C.M., (1986). Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research* **27,** 207-214.

[22] Jagerman, D.L., (1974). Some properties of the Erlang loss function. *The Bell System Technical Journal* Vol. **53,** No. **3,** 525-551.

[23] Mandelbaum, A., Massey, W.A., Reiman, M., (1998). Strong approximations for Markovian service networks. *Queueing Systems: Theory and Applications (QUESTA),* **30** , 149-201.

[24] Mandelbaum, A., Massey, W.A., Reiman, M., Rider (1999). Time varying multiserver queues with abandonments and retrials. *ITC* **16** , Edinburgh Scotland.

[25] Mandelbaum, A., Pats, G., (1995). State-dependent queues: approximations and applications. *Stochastic Networks,* 239-282. Ed. F.P. Kelly, R.J. Williams, Springer-Verlag.

[26] Mandelbaum, A., Shimkin, N., (1999). A model for rational abandonments from invisible queues. *Technical Report* , Center for Communication and Information Technologies, Electronical Engineering, Technion, Israel; CC PUB#**266** , January 1999.

[27] Nguyen, V., (1993). Processing networks with parallel and sequential tasks: heavy traffic analysis and Brownian limits. *The Annals of Applied Probability.* **3,** 28-55.

[28] Palm, C., (1937). Etude des delais d'attente. *Ericsson Technics* **5,** 37-56.

[29] Palm, C., (1943). Intensitatsschwankungen im fernsprechverkehr, *Ericsson Technics* **44(1),** 1-189.

[30] Palm, C., (1953). Methods of judging the annoyance caused by congestion. *Tele.* **4,** 189-208.

[31] Puhalskii, A., (1994). On the invariance principle for the first passage time. *Mathematics of Operations Research,* Vol. **19,** No. **4,** 946-954.

[32] Riordan, J., (1962). *Stochastic Service Systems,* Wiley.

[33] Stone, C., (1963). Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois Journal of Mathematics* **7,** 638-660.

[34] Sze, D.Y., (1984). A queueing model for telephone operator staffing. *Operations Research* **32,** 229-249.

[35] Whitt, W., (1992). Understanding the efficiency of multi-server service systems. *Management Science* Vol. **38,** No. **5,** 708-723.