# DYNAMIC STAFFING IN A TELEPHONE CALL CENTER AIMING TO IMMEDIATELY ANSWER ALL CALLS

by

*Ward Whitt**

AT&T Labs–Research

August 28, 1997

Revision: January 19, 1999

*Operations Research Letters* 24 (1999) 205–212

*Room A117, AT&T Labs, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971; email: wow@research.att.com

## Abstract

This paper proposes practical modeling and analysis methods to facilitate dynamic staffing in a telephone call center with the objective of immediately answering all calls. Because of this goal, it is natural to use infinite-server queueing models. These models are very useful because they are so tractable. A key to the dynamic staffing is exploiting detailed knowledge of system state in order to obtain good estimates of the mean and variance of the demand in the near future. The near-term staffing needs, e.g., for the next minute or the next twenty minutes, can often be predicted by exploiting information about recent demand and current calls in progress, as well as historical data. The remaining holding times of calls in progress can be predicted by classifying and keeping track of call types, by measuring holding-time distributions and by taking account of the elapsed holding times of calls in progress. The number of new calls in service can be predicted by exploiting information about both historical and recent demand.

*Keywords:* server staffing, operator staffing, forecasting, predicting remaining call holding times, infinite-server models, nonstationary queues, telephone call centers

## 1. Introduction

In the last fifteen years there has been spectacular growth in retail business conducted by telephone. It has been recognized that a critical factor in business success in this environment is being able to respond rapidly to customer requests. At the same time, the costs of staffing telephone call centers have become a substantial part of business expense. Thus it is essential to efficiently manage telephone call centers, so that customer requests are met without excess staffing. This has led to sophisticated systems designed to balance the competing objectives; e.g., see Hall [12], Andrews and Parsons [4], Thompson [19], Brigandi, Dargon, Sheehan and Spencer [5], Jennings, Mandelbaum, Massey and Whitt [13], and references therein.

However, in highly competitive businesses, success often depends on being able to provide service of even higher quality. One way to do so in a telephone call center is to aim to immediately answer *all* calls. Of course, actually answering all calls immediately upon arrival may be an unrealistic objective, especially when demand is highly unpredictable, as is often the case when unusually high demand is periodically stimulated by special promotions. Nevertheless, we contend that it is often possible to come quite close to the goal of immediately answering all calls in a reasonably stable environment by dynamically staffing based on detailed information on the recent history of the call center. Modern computer systems make it possible to collect and apply this information.

The purpose of this paper is to discuss ways to manage a telephone call center and do dynamic staffing with the goal of immediately answering all calls. For this purpose, two key system attributes are: (i) scale and (ii) flexibility. By "scale," we mean large size. With large size, the demand fluctuations over time tend to be a smaller percentage of the average workload; i.e., with larger systems the workload tends to be more predictable. Under regularity conditions, the required staffing level at any time has, at least approximately, a Poisson distribution. The possible fluctuations of a Poisson distribution can be characterized roughly by its standard deviation, which is always the square root of the mean. As the mean increases, the standard deviation becomes a smaller proportion of the mean. Expressed differently, extra staffing to account for fluctuations tends to be about $c\sqrt{m}$ for some constant $c$, typically with $1 \leq c \leq 10$ when the mean is $m$. Thus, for very large $m$, we can staff very near the mean $m$. Then the problem reduces to predicting the mean, which is what much of this paper is about.

By "flexibility," we primarily mean the ability of the service system to dynamically control the staffing response. Flexible staffing can be achieved by ensuring that representatives (the staff of the call center) have alternative work. Natural forms of alternative work are training and after-call processing of previous calls. Flexibility is achieved by having representatives do alternative work when demand is relatively low. Idle representatives may also be used to make contact with customers by making calls themselves (having outbound as well as inbound calls). With substantial alternative work, there can be a large number of representatives in the center not currently answering calls who are available to start answering calls on short notice. Moreover, with advanced communication systems, physical proximity may not be a critical requirement.

If possible, it is desirable to classify calls according to their importance or value. for example, there might be two classes of calls: high-priority calls and low-priority calls. It might be decided only to answer all high-priority calls immediately upon arrival. In contrast, low-priority calls might be answered on a best-effort basis. Then, assuming that the high-priority calls can be identified, they can be considered the demand that must be met, and the low-priority calls can be considered as a form of alternative work.

Indeed, it is natural to consider several priority levels for different staff functions. The high-priority calls would have highest priority. Then the low-priority calls might have second priority, the after-call processing of previous calls might have third priority, and the training might have fourth priority. These priority assignments are natural because the different types of alternative work have very different degrees of delay tolerance. The low-priority calls may have a delay tolerance of a few seconds or minutes, while the after-call processing of previous calls may have a delay tolerance of a few minutes or hours, while the training may have a delay tolerance of a few hours or days. The different time scales of the requirements make it natural to use a priority scheme.

In this paper we are primarily concerned with dynamic staffing for a single class of demand. In the priority scheme, we are thus focusing on the highest priority class only. If there is enough lower-priority work, then it is natural to think that we should be able to immediately answer all calls for the highest priority. Indeed, with enough lower-priority work that also must be satisfied, it should be possible to meet the highest priority demand by assigning servers on quite short notice.

The goal of dynamic staffing dictates a different kind of stochastic analysis from the goal of longer-term capacity planning. For longer-term capacity planning, it is natural to use the classical Erlang loss and delay formulas and their relatives, which describe the steady-state performance. In contrast, dynamic staffing requires time-dependent analysis of a time-dependent model. The relevant perspective for dynamic staffing is optimal control given an appropriate system state. Thus, if a classical Erlang model is used, then it is the time-dependent behavior we want to know.

We conclude this introduction by mentioning our previous related work. The problem of longer-term staffing to meet time-varying demand is considered in Jennings, Mandelbaum, Massey and Whitt [13]. The problem of improving service by informing customers of anticipated delays, when delays are deemed necessary, is considered in Whitt [20]. Ways to actually predict queueing delays are proposed in Whitt [21]. Admission control schemes for immediate-request calls when some calls book ahead are proposed in Greenberg, Srikant and Whitt [11] and Srikant and Whitt [17]. These admission control schemes also involve real-time prediction of service load. Theoretical analysis of infinite-server queues closely related to the analysis here is contained in Duffield and Whitt [7]. The primary aim there is to gain insight into overload controls, but the supporting theory is similar to what we do here. That paper provides additional supporting theoretical results, such as large deviations principles. Ways to calculate the time-dependent characteristics of the Erlang loss model are developed in Abate and Whitt [2].

## 2. Current Calls Remaining in the Future

In doing our analysis, we assume that all calls are immediately answered, because that is what we are aiming to achieve. The immediate-answer property means that the usual performance-analysis concern about the impact of waiting before beginning service or blocking and retrials after blocking need not be considered; i.e., it suffices to use an infinite-server model.

In this context, the staffing requirement (number of required servers) in the near future can be divided into two parts: (i) the number of current calls that will remain in progress in the future and (ii) the number of new calls that will arrive and remain in service. Moreover, it is reasonable to regard these two components of future demand as being independent (and we do) and focus on them separately. We aim to describe the mean and variance of each component. The overall mean and variance of the total future staffing requirement will be the sums of the component means and variances.

If the lead time (the length of the interval until the time for which the prediction is made) is larger than all but a few call holding times, then current calls in progress will tend not to be significant. However, we are thinking of predicting for lead times less than many call holding times. For example, we might have a lead time of 5 minutes in an airline reservation or software support call center, where many calls exceed 30 minutes.

Assuming that both components of staffing are relevant, it is useful to treat the two components separately in order to actually determine how important each component is. This reveals how much attention in the control should be given to current calls in progress as opposed to new arrivals. As the lead time increases well beyond the average call holding time, the current calls obviously play a less important role. However, we are thinking of short lead times, so that the evolution of current calls can be the dominant component.

Current calls in progress also have the potential of providing useful information. Given the types of calls in progress and their elapsed holding times, we may be able to accurately predict the remaining holding times. To appreciate the great control opportunity here, it is important to break away from the conventional stochastic models that are traditionally used in performance analysis. The conventional stochastic model has all holding times exponentially distributed with a common mean. With that model, the only relevant information is the number of active calls. The number of active calls can of course be important for predicting future requirements, but other state information can be even more important.

In many settings, there are different classes of customers with very different holding-time distributions. One class might have a mean holding time of 1 minute, while another class has a mean holding time of 30 minutes. (Think of airline reservation and software support centers.) Often the customer class can be identified from the originating telephone number. Moreover, the representative may be able to further classify a call after it arrives. The representative can proceed to refine the classification while the call is in progress. (For example, think of a technical support service for a personal computer dealer. After answering the call, and after some conversation, the representative may be able to better predict how much longer the call will be.) The representatives might even directly estimate the remaining length of the current call and provide updates while the call remains in progress.

The representatives may not only be able to *classify* the calls, but the representatives may be able to *control* the remaining holding times of calls. When the system is heavily congested, representatives could be informed and some might be able to take actions to shorten calls in progress. If such a policy is used, we presume that it is properly taken account of when remaining holding times are predicted.

The holding time could also depend strongly upon the particular representative handling the call. Representatives may have special skills. The assignment system may attempt to assign calls to representatives with the appropriate special skills, but if that assignment is not possible, then the assignment will be made to an alternative representative. When future staffing requirements are being considered, it is possible to take account of the assignments in progress. Again, available sophisticated computer systems make it possible to obtain and apply this information.

We now propose a general model to predict the number of current calls remaining in future. We assume that the number $n$ of current calls is known. We assume that the remaining holding time for call $i$ conditional on all available system state information is a random variable $T_i$ with cumulative distribution function (cdf) $H_i$, i.e.,

$$P(T_i \le t) = H_i(t)\ , \quad t \ge 0\ . \tag{2.1}$$

We assume that these random variables $T_i$, $1 \le i \le n$, are mutually independent.

Let $C(t)$ be the number of the current calls in progress $t$ time units in the future. By the assumptions above, for each $t$, $C(t)$ is the sum of $n$ non-identically-distributed Bernoulli ($\{0,1\}$-valued) i.i.d. random variables. Thus, the mean and variance of $C(t)$ can be computed. In particular,

$$EC(t) = \sum_{i=1}^{n} H_i^c(t)\ , \quad t \ge 0\ , \tag{2.2}$$

and

$$Var\ C(t) = \sum_{i=1}^{n} H_i(t) H_i^c(t)\ , \quad t \ge 0\ , \tag{2.3}$$

where $H_i^c(t) = 1 - H_i(t)$. Assuming that $n$ is relatively large, it is natural to regard $C(t)$ as normally distributed with mean and variance in (2.2) and (2.3), by virtue of the central limit theorem for independent non-identically-distributed random variables, see p. 262 of Feller [10].

Obviously $EC(t)$ in (2.2) is decreasing in $t$ for all $t$. The behavior of the variance $Var\ C(t)$ in (2.3) is somewhat more complicated. If $H_i(t)$ is differentiable at $t$ and if $H_i(t) \le (\ge) 1/2$, then $H_i(t)H_i^c(t)$ is increasing (decreasing) in $t$. Thus $Var\ C(t)$ is concave, first increasing and then decreasing. If $t$ is sufficiently short, then $EC(t)$ is relatively large while $Var\ C(t)$ is relatively small, so that prediction is important and accurate prediction is possible.

Implementation depends on being able to appropriately identify the cdf's $H_i(t)$. There are several natural scenarios. If current call $i$ is known to have holding-time cdf $G_i$ upon arrival, and nothing more is known except that the elapsed holding time is $t_i$, then we let $H_i$ be the conditional remaining service time given the elapsed holding time $t_i$, i.e.,

$$H_i(t) = \frac{G_i(t + t_i) - G_i(t_i)}{G_i^c(t_i)}\ , \quad t \ge 0\ , \tag{2.4}$$

where $G_i^c(t) = 1 - G_i(t)$. Formula (2.4) exploits two pieces of information: the original cdf $G_i$ for this call (which may depend on other factors, such as the representative handling it) and the elapsed holding time $t_i$.

Of course, if $G_i$ is an exponential cdf, then $H_i$ in (2.4) is just $G_i$ again, by the lack-of-memory property of the exponential distribution. However, if

$$G_i(t) = 1_{[c,\infty)}(t) \ , \tag{2.5}$$

where $1_A(t) = 1$ if $t \in A$ and 0 otherwise, corresponding to a *constant holding time* of length $c$ associated with a very well defined task, then $H_i(t) = 1_{[c-t_i,\infty)}(t)$, corresponding to a constant remaining holding time of length $c - t_i$. Obviously we can predict very accurately with low-variability holding times. In many scenarios, low variability can be achieved *after* the call has been properly classified. Many tasks have highly predictable durations. High variability often stems from having uncertainty about which of two or more possible predictable tasks is required. Thus, there is reason to expect that the variance $Var\ C(t)$ will be small when the proper information is brought to bear.

On the other hand, if $G_i$ is highly variable, then the elapsed holding time can greatly help in future prediction. With highly variable holding-time distributions, *a very long elapsed holding time tends to imply a very long remaining holding time.* To illustrate, let $Y(a,b)$ denote a random variable with a Pareto distribution, i.e.,

$$P(Y(a,b) \le t) = 1 - (1+bt)^{-a} \ , \quad t \ge 0 \ . \tag{2.6}$$

The high variability of $Y(a,b)$ is indicated by the fact that the tail decays as a power instead of exponentially. Now let $Y_t(a,b)$ denote the conditional remaining holding time given an elapsed holding time $t$. It turns out that $Y_t(a,b)$ is distributed as $(1+bt)Y(a,b)$; see Theorem 8 of [7]. Hence,

$$EY_t(a,b) = (1+bt)EY(a,b) \ ; \tag{2.7}$$

i.e., the mean remaining holding time $EY_t(a,b)$ is approximately proportional to the elapsed holding time $t$.

If the cdf $G_i$ is known but the elapsed holding time is not available, then it is still possible to exploit the fact that service is in process. To do so, it is it is natural to use the equilibrium-excess cdf associated with $G_i$, namely,

$$H_i(t) = G_{ie}(t) \equiv \frac{1}{ET_i}\int_0^t [1 - G_i(u)]du \ . \tag{2.8}$$

For an $M/G/\infty$ queueing model in steady state, formula (2.8) is in fact the *exact* distribution of the remaining holding time; e.g., see p. 161 of Takács [18].

As indicated above, it may be possible to directly predict the remaining holding time cdf $H_i$ while the call is in progress. If only partial information is given, then it is natural to fit the cdf to the available information. For example, if only the mean $m_i$ of $H_i$ is specified, then we can fit $H_i$ to an exponential cdf with mean $m_i$ by setting

$$H_i(t) = 1 - e^{-t/m_i} \ , \quad t \geq 0 \ . \tag{2.9}$$

However, even if only the mean of $H_i$ is given directly, it should be possible to do better by exploiting historical data. Depending on the application, we should be able to assess the variability. More generally, we can assume that

$$H_i(t) = F_i(t/m_i) \ , \quad t \geq 0 \ , \tag{2.10}$$

where $F_i$ is a cdf with mean 1 and the right shape. Then the scaling by $m_i$ in (2.10) makes the cdf $H_i$ have mean $m_i$ and the shape of $F_i$. For example, highly variable holding times might have the Pareto shape in (2.6) for some $a > 1$, where $b$ is chosen to produce mean 1. Note that (2.9) is a special case of (2.10) with $F_i(t) = 1 - e^{-t}$.

The goal in bringing information to bear on the cdf's is to have $H_i(t)$ either be close to 1 or close to 0 for the desired lead time $t$. If $H_i(t) \approx 1$ for many $i$, while $H_i(t) \approx 0$ for the remaining $i$, then the mean $EC(t)$ in (2.2) will be substantial, while the variance $VarC(t)$ in (2.3) will be small.

For a concrete example, suppose that $H_i(t) = 1 - \epsilon$ for a fraction $p$ of the calls, while $H_i(t) = \epsilon$ for the remainder of the calls. Then $EC(t) = (1-p)n + n\epsilon(2p-1)$ and $VarC(t) = n\epsilon(1-\epsilon)$. The degree of uncertainty can be characterized approximately by the ratio of the standard deviation to the mean, which here is

$$\frac{SD(C(t))}{EC(t)} \approx \frac{1}{\sqrt{n}} \left( \frac{\sqrt{\epsilon(1-\epsilon)}}{1-p+(2p-1)\epsilon} \right) \ . \tag{2.11}$$

The ratio (2.11) tends to be small if *either* $n$ is large or one of $\epsilon$ and $(1-\epsilon)$ is small, provided that $p$ is not near 0 or 1.

## 3. New Calls in Progress in the Future

In this section we develop a procedure to predict how many new calls will be in progress in the future. Part of this prediction involves predicting the future arrival rate of new calls, but we must also consider how long the new calls will remain in service, since some new arrivals may depart before the specified lead time. We approach this second problem by using the $M_t/G/\infty$ queueing model, assumed to start empty. We assume that the system starts empty because we have already accounted for the calls initially in service in Section 2. It is natural to assume a Poisson arrival process, which can be justified by the assumption that arriving customers act independently of each other. However, reality usually dictates that the arrival-rate function $\lambda(t)$ should be time-varying. We thus think of the arrival-rate function $\lambda(t)$ as time-varying but deterministic. In fact, $\lambda(t)$ is not known, so that $\lambda(t)$ should properly be thought of as the realization of a stochastic process $\{\Lambda(t) : t \geq 0\}$. However, as an approximation, we let $\lambda(t)$ be the mean $E\Lambda(t)$ and aim to estimate it. Then use this mean value as the deterministic arrival-rate function in the $M_t/G/\infty$ model. We hope to exploit system state to reduce the uncertainty about the future arrival rate.

A useful first step is to classify the arrivals into different call types. We assume that the separate call types are independent, so that we can simply add the means and variances of the different types to obtain the mean and variance of the total number of new arrivals present at time $t$. We now analyze a single call type.

For any one call type, we assume that the arrival process is a nonhomogeneous Poisson process with deterministic arrival-rate function $\lambda(t)$ and that the holding-time is a random variable $T$ with cdf $G$. Let $N(t)$ be the number of these calls in the system at time $t$ in the future. Using basic properties of infinite-server queues, e.g., see Eick, Massey and Whitt [9], we conclude that $N(t)$ has a Poisson distribution with mean (and variance)

$$EN(t) = Var\, N(t) = m(t) = \int_0^t G^c(t-u)\lambda(u)du \ . \tag{3.1}$$

It is significant that (3.1) remains the valid formula for the mean (but not the variance) when $\lambda(t)$ is replaced by a stochastic process $\Lambda(t)$. Because of the linearity associated with the infinite server model,

$$EN(t) = E\int_0^t G^c(t-u)\Lambda(u)du = \int_0^t G^c(t-u)E\Lambda(u)du \ . \tag{3.2}$$

Henceforth, we focus on (3.1).

Formula (3.1) can easily be calculated numerically. It should suffice to use the simple trapezoidal rule approximation

$$m(t) \approx \frac{1}{2n}G^c(t)\lambda(0) + \frac{1}{n}\sum_{k=1}^{n-1} G^c(t-(k/n))\lambda(k/n) + \frac{1}{2n}G^c(0)\lambda(t) \ , \tag{3.3}$$

where $n$ is chosen large enough to produce negligible error; e.g., see Davis and Rabinowitz [6].

Formula (3.1) applies with general arrival-rate functions, but given the relatively short time scale for prediction, it might be possible to consider as an approximation a constant arrival rate function, i.e., a step function, which is zero before time 0. In that case, (3.1) becomes

$$m(t) = \lambda ET G_e(t) = \lambda \int_0^t G^c(u)du \ , \quad t \geq 0 \ , \tag{3.4}$$

where $G_e$ is the equilibrium-excess cdf associated with the holding-time cdf $G$, defined as in (2.8). For practical applications of (3.4), it is significant that the equilibrium-excess cdf $G_e$ often inherits the structure of the cdf $G$; see Section 4 of Duffield and Whitt [8]. For example if $G$ is a mixture of exponentials or phase type, then so is $G_e$. Moreover, given the Laplace-Stieltjes transform of $G$, i.e.,

$$\hat{g}(s) = \int_0^\infty e^{-st} dG(t) \ , \tag{3.5}$$

we can easily compute $G_e^c(t) \equiv 1 - G_e(t)$ by numerically inverting its Laplace transform

$$\hat{G}_e^c(s) = \int_0^\infty e^{-st} G_e^c(t)dt = \frac{sET - 1 + \hat{g}(s)}{s^2 ET} \ ; \tag{3.6}$$

e.g., see Abate and Whitt [1].

Given (3.1) or (3.4), the remaining problem is to estimate the appropriate arrival-rate function $\lambda(t)$ and the holding-time cdf $G$ for the single customer class under consideration. It is important, though, to recognize that these should depend on the observation time. We assume that the holding-time cdf $G$ can be adequately estimated from historical data (over a much longer time interval than the prediction lead time). We estimate the arrival-rate function $\lambda(t)$ in two steps. In the first step we obtain a preliminary estimate of $\lambda(t)$ over a day, based on seasonal, day-of-the-week and known promotion effects. We call this estimate $\lambda_a(t)$; it is the standard prediction that can be done a day in advance or possibly even a week in advance. Given arrival counts in subintervals of previous days, the arrival rate function $\lambda_a(t)$ might be a linear or quadratic function estimated by least squares; e.g., see Massey, Parker and Whitt [14]. In the second step we make adjustments to the anticipated demand by taking account of the observed demand during previous times on the same day. A simple approach is to estimate a multiplier $r(t)$ based on the observed history over the day, i.e., we represent the observed demand as $r(t)\lambda_a(t)$. We thus estimate $r(t)$ from the ratio

$$r(s) = \lambda(s)/\lambda_a(s) \tag{3.7}$$

for times $s$ in the past. In practice, we can divide time into equally spaced intervals. Let $\gamma(k)$ and $\gamma(k)$ denote the predicted and observed demand in interval $k$. We then can predict the ratio $r(n) = \lambda(n)/\lambda_a(n)$ for the $n^{\text{th}}$ interval by using exponential smoothing, i.e.,

$$r(n) = \alpha \frac{\gamma(n-1)}{\gamma_a(n-1)} + (1-\alpha) r(n-1) = \frac{\sum_{k=1}^m \alpha^k [\gamma(n-k)/\gamma_\alpha(n-k)]}{\sum_{k=1}^m \alpha^k} \tag{3.8}$$

for some constants $m$ and $\alpha$. We can choose $m$ and $\alpha$ by finding the best fit using historical data; e.g., we see what values minimize means squared error.

Given that current time is 0, we would use $\lambda(t) \equiv r(0)\lambda_a(t)$ as the predicted arrival rate for future times $t$, where $r(0)$ is determined from the recent history up to the current time 0, as in (3.8).

In this section we have emphasized methods to determine the mean and variance (which turns out to be equal to the mean) of the number $N(t)$ of new calls in the system at time $t$ in the future, assuming the $M_t/G/\infty$ model where the arrival rate function $\lambda(t)$ and service-time cdf $G$ are known (can be estimated). We thus account for uncertainty naturally associated with the $M_t/G/\infty$ queueing model, given specified model components. In doing so, we have only briefly considered statistical forecasting required in estimating $\lambda(t)$, $t \geq 0$, and $G$. However, in some contexts the critical problem may be estimating these quantities; e.g., see Abraham and Ledolter [3], Pankratz [16] and references therein. We have indicated one possible approach to estimating $\lambda(t)$, $t \geq 0$, and $G$. Our $M_t/G/\infty$ analysis will remain valid if some other estimation procedure is used, provided that the estimates are sufficiently accurate (have approximately the correct expected value and are accompanied by sufficiently low uncertainty).

The most likely shortcoming in our analysis seems to be the assumption that $\lambda(t)$ can be estimated without significant remaining uncertainty. We now indicate a simple modification in the analysis for cases when that is not possible. The general idea is to inflate the variance $Var\ N(t)$ to account for additional uncertainty about $\lambda(t)$. We propose one specific approach that can be considered: We assume that the stochastic arrival rate $\Lambda(t)$ has the following special form

$$\Lambda(t) = X\lambda(t), \quad t \geq 0 \ , \tag{3.9}$$

where $X$ is a positive random variable, independent of the queueing process with $EX = 1$. Let $N_X(t)$ be the number of new calls in the system at time $t$ associated with $X$. The $M_t/G/\infty$ model structure and the special assumption (3.9) imply that

$$N_X(t) = XN_1(t), \quad t \geq 0 \ , \tag{3.10}$$

so that

$$E[N_X(t)^k] = E[X^k]E[N_1(t)^k], \quad t \geq 0 \ , \tag{3.11}$$

for all $k \geq 1$. The first important conclusion is that the mean $EN(t)$ is unaltered by the random variable $X$ in (3.9). As noted in (3.2), this holds more generally. The specific form (3.9) allows us to obtain a convenient simple expression for the variance, in particular,

$$Var\ N_X(t) = m(t) + [m(t) + m(t)^2]Var\ (X)\ , \tag{3.12}$$

where $m(t) = EN_1(t)$ as before. As a consistency check, note that $Var\ N_X(t) \to m(t) = Var\ N_1(t)$ when $Var(X) \to 0$.

## 4. Predicting Future Demand

The total future demand, say $D(t)$, is the sum of the current calls remaining in the future and the new calls in progress in the future, i.e.,

$$D(t) = C(t) + N(t), \quad t \geq 0\ . \tag{4.1}$$

The mean is the sum of the means in (2.2) and and (3.1). The variance is the sum of the variances in (2.3) and (3.1) or (3.12). Since both components should have approximately normal distributions, so should the sum. Let $N(0,1)$ denote a standard (mean 0, variance 1) normal random variable. Thus we let the required number of servers at time $t$ be

$$s(t) = \lceil ED(t) + z_\alpha \sqrt{Var\, D(t)} + 0.5 \rceil \tag{4.2}$$

where $P(N(0,1) > z_\alpha) = \alpha$ and $\lceil x \rceil$ is the least integer greater than $x$. We choose $\alpha$ suitably small ($z_\alpha$ suitable large) so that the likelihood of demand exceeding supply at time $t$ is suitably small.

Formula (4.2) is the same as formula (4) in Jennings et al. [13]. The difference is that here we dynamically exploit all available information up to the current time in order to more accurate predict the mean $ED(t)$ and the variance $Var\, D(t)$ a relatively short time $t$ in the future.

The estimation procedure can be said to be working if the demand $t$ units in the future is indeed distributed approximately as $N(ED(t), VAR\, D(t))$. Thus the procedure can be checked with historical data. The estimation scheme is *effective* if (i) $ED(t)$ is indeed estimated accurately and (ii) $Var\, D(t)$ is suitably small (relative to $ED(t)$) and estimated accurately. Then the required number of servers $s(t)$ will be only slightly greater than actually needed. The *overhead due to uncertainty* can be described by the percentage the difference $s(t) - ED(t)$ is of $ED(t)$. For example, the overhead is 10% if $ED(t) = 400$ and $s(t) = 440$. The overhead represents the cost of providing the high quality of service.

Finally, it remains to ensure that at least the required number $s(t)$ of servers will be available at time $t$ in the future. Greater efficiency can be achieved if some number of servers that are quite sure to be needed, such as $ED(t) - z_\alpha\sqrt{Var\,D(t)}$, are committed, while another number, say $2z_\alpha\sqrt{Var\,D(t)}$ are made available, but not committed, by being placed on alert. The servers on alert might pursue other tasks, but be ready to answer calls immediately upon notice. This analysis shows the degree of flexibility needed, and how scale can help.

## 5. Verification and Examination

It is important to recognize that uncertainty plays an important role in the staffing decision. The number $s(t)$ of servers in (4.2) depends upon the standard deviation $\sqrt{Var\ D(t)}$ as well as the mean $ED(t)$. We thus want to make comparisons with historical data to ensure that both $ED(t)$ and $\sqrt{Var\ D(t)}$ are being estimated properly.

We also want to examine the historical data to better understand where $ED(t)$ and $\sqrt{Var\ D(t)}$ come from. Our approach has provided a means for identifying sources of uncertainty in the prediction. For a future time $t$ of interest, we can see how much of the mean $ED(t)$ and variance $Var\ D(t)$ are due to the number $C(t)$ of current calls and the number $N(t)$ of new calls. An analysis of the models determines what to anticipate. An analysis of historical data can confirm the predictions about these two sources of uncertainty.

For both $C(t)$ and $N(t)$, from an analysis of historical data we can see how much uncertainty is due to (i) stochastic fluctuations for the specified model and (ii) uncertainty about the model components (confounded with an improper model).

Thus, to substantiate the prediction process, we suggest not only comparing the prediction of overall demand $D(t)$ with observations, but also comparing predictions of component demand, $C(t)$ and $D(t)$, plus predictions of the model elements. Such performance feedback provides the basis for prediction improvements in the future. It might also be possible to take other actions (controls) to improve system performance.

## References


[1] Abate, J. and W. Whitt, "Numerical Inversion of Laplace Transforms of Probability Distributions," *ORSA J. Computing* 7 (1995) 36–43.

[2] Abate, J. and W. Whitt, "Calculating Transient Characteristics of the Erlang Loss Model by Numerical Transform Inversion," *Stochastic Models* 14 (1998), 663–680.

[3] Abraham, B. and J. Ledolter, *Statistical Methods for Forecasting*, Wiley, New York, 1983.

[4] Andrews, B. H. and H. L. Parsons, "Establishing Telephone Agent Staffing Levels Through Economic Optimization," *Interfaces*, 23 (1993) 14–20.

[5] Brigandi, A. J., D. R. Dargon, M. J. Sheehan and T. Spencer III, "AT&T's Call Processing Simulator (CAPS) Operational Design for Inbound Call Centers," *Interfaces*, 24 (1994) 6–28.

[6] Davis, P. J. and P. Rabinowitz, *Methods of Numerical Integration*, second edition, Academic Press, New York, 1984.

[7] Duffield, N. and Whitt, W. "Control and Recovery from Rare Congestion Events in a Large Multi-Server System," *Queueing Systems*, 26 (1997) 69–104.

[8] Duffield, N. G. and W. Whitt, "A Source Traffic Model and its Transient Analysis for Network Control," *Stochastic Models*, 14 (1998), 51–78.

[9] Eick, S. G., W. A. Massey and W. Whitt, "The Physics of the $M_t/G/\infty$ Queue," *Oper. Res.* 41 (1993) 731–742.

[10] Feller, W., *An Introduction to Probability Theory and its Applications*, vol. II, second edition, Wiley, New York, 1971.

[11] Greenberg, A. G., R. Srikant and W. Whitt, "Resource Sharing for Book-Ahead and Instantaneous-Request Calls," *IEEE/ACM Trans. Networking*, to appear.

[12] Hall, R. W., *Queueing Methods for Services and Manufacturing*, Prentice Hall, Englewood Cliffs, NJ, 1991.

[13] Jennings, O. B., A. Mandelbaum, W. A. Massey and W. Whitt, "Server Staffing to Meet Time-Varying Demand," *Management Sci.* 42 (1996) 1383–1394.

[14] Massey, W. A., G. A. Parker and W. Whitt, "Estimating the Parameters of a Nonhomogeneous Poisson Process with Linear Rate," *Telecommunication Systems*, 5 (1996) 361–388.

[15] Massey, W. A. and W. Whitt, "Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates," *Queueing Systems*, 25 (1997) 157–172.

[16] Pankratz, A., *Forecasting with Univariate Box-Jenkins Models*, Wiley, New York, 1983.

[17] Srikant, R. and W. Whitt, "Resource Sharing for Book-Ahead and Instantaneous-Request Calls Using a CLT Approximation," *Telecommunication Systems*, to appear.

[18] Takács, L., *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.

[19] Thompson, G. M., "Accounting for the Multi-Period Impact of Service When Determining Employee Requirements for Labor Scheduling," *J. Oper. Management*, 11 (1993) 269–288.

[20] Whitt, W., "Improving Service by Informing Customers About Anticipated Delays," *Management Sci.*, to appear.

[21] Whitt, W., "Predicting Queueing Delays," *Management Science*, to appear.