What You Should Know About Queueing Models To Set Staffing Requirements in Service Systems

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University
304 S. W. Mudd Building
500 West 120th Street
New York, NY 10027-6699
ww2040@columbia.edu

Abstract

One traditional application of queueing models is to help set staffing requirements in service systems, but the way to do so is not entirely straightforward, largely because demand in service systems typically varies greatly by the time of day. This paper discusses ways - old and new - to cope with that time-varying demand.

Keywords: setting staffing requirements, call centers, time-varying demand, queues with time-varying arrival rate, nonstationary queueing models, pointwise stationary approximation, modified-offered-load approximation, infinite-server queues.

April 26, 2007

1. Introduction

The purpose of this paper is to provide a brief high-level overview of one important topic involving stochastic models. We discuss queueing models that can be used to set staffing requirements in service systems. There are many possible applications, but we have in mind telephone call centers and their generalizations to customer contact centers, allowing contact by other means besides the telephone, such as email and web chat. Gans et al. [6] provide a good introduction to call centers with an operations research perspective. The traditional management perspective is nicely described by Cleveland and Mayben [2].

An illustrative specific context is a medium-sized financial-services call center, which employs about 200 agents at peak periods. An important feature of this call center, as well as most other service systems, is that demand for service varies greatly by time of day, as shown in Figure 1. (The peak agent requirement is about 200 because the average call holding time is about 6 minutes or 0.1 hour. The required number of agents is roughly the instantaneous offered load - the product of the arrival rate and the average service time: $2000 \times 0.1 = 200$.) The problem we discuss is: **How can we set appropriate staffing levels in the face of such time-varying demand?** Our discussion here is an abridged version of our recent survey in Green et al. [9], including recent research in Feldman et al. [5].

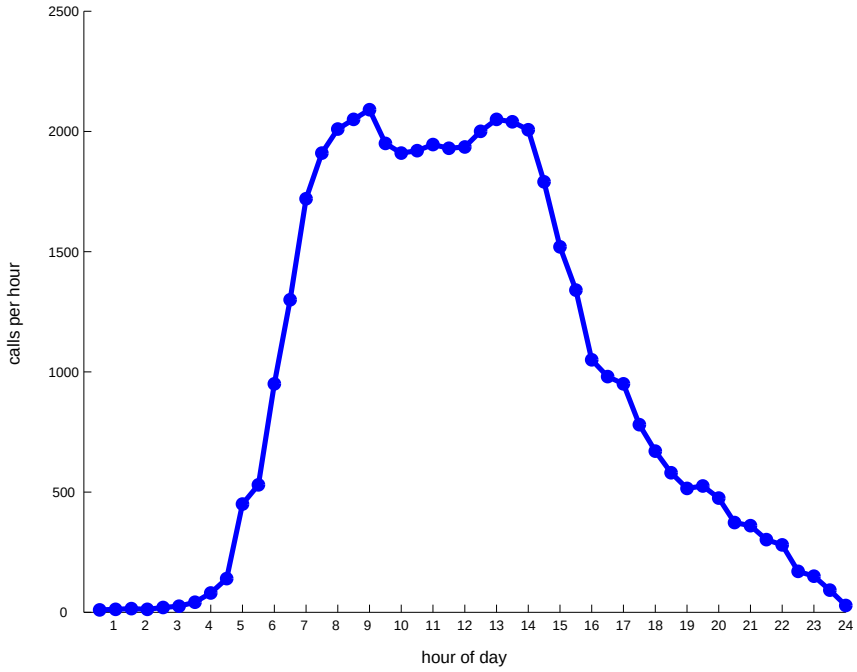

Figure 1: Arrivals per hour to a medium-sized financial-services call center.

2. The Staffing Problem

The staffing problem is to determine the required number of agents as a function of time. The goal is to provide a satisfactory quality of service at all times, without having more agents than necessary.

In call centers, agents often possess different call-handling skills, so that we need to determine the required numbers of agents with different skill sets. There often are multiple classes of customers as well, so the staffing problem is related to a complex "skill-based" routing problem; e.g., see Gans et al. [6]. However, in this discussion of how to cope with time-varying demand, we restrict attention to the single-skill special case. It is important, though, that the methods for coping with time-varying demand should be relevant for the more general multi-skill settings. That is so, and there are two reasons: First, with a limited amount of cross-training, the total staffing in multi-skill cases can often be set the same as for the single-skill case; see Wallace and Whitt [17], Gurvich and Whitt [10] and references therein. Second, the methods for coping with time-varying demand discussed here do indeed extend naturally to more complex service networks.

The single-skill staffing problem can be expressed as an optimization problem: Minimize the total agent hours assigned, subject to specified quality-of-service constraints holding at all times. We may have performance constraints for (i) the proportion of customers that abandon before an agent can respond, (ii) the average waiting time before an agent can respond, among served customers, and (iii) the proportion of served customers that have to wait more than 20 seconds (or some other threshold) before an agent can respond. In applications, the precise definition of performance targets can be important. For example, both poor service and inefficient staffing can occur if the performance requirements are allowed to be expressed as long-run averages, because those targets can then be met by alternating periods of understaffing and overstaffing. We assume that the performance targets are to be met locally at all times (or in all sufficiently short time intervals).

Stochastic queueing models can play an important role, because customer arrivals, abandonment and service times are variable and uncertain. It is important to recognize that there are two kinds of variability: There is the *predictable variability* of the demand rate as a function of the time of day and the day of the week, and there is the *stochastic variability* about the predictable average caused by the random behavior of customers and agents.

Here we are concerned with the number of agents required as a function of time, assuming

for simplicity that the number can change continuously through time. In practice, however, staffing changes typically can occur only periodically, such as once every 30 minutes, so that the staffing level is constant during staffing intervals. One simple staffing rule is to use throughout each staffing interval the maximum number of agents required at any single time in that staffing interval.

After having set staffing requirements, managers often may be able to make further adjustments in real time, moving agents in and out of the line of duty, to respond to unanticipated deviations in demand. Such adjustments are made possible by having extra agents on site doing alternative work or being trained, or by being able to use remote agents on short notice. Without that extra flexibility, extra agents are needed to provide insurance against unexpected high demand. With that extra flexibility, management may be able to circumvent the entire staffing problem. From a practical perspective, it is important to recognize that it may be more effective to provide appropriate flexibility to do real-time adjustments than to carefully determine the "best" staffing level in advance. Here we are assuming that such extra flexibility is not available.

3. Queueing Models

The first thing to observe when we consider ways to cope with the time-varying demand is that there is a fundamental disconnect between basic queueing theory and practice: Standard textbook queueing theory does not apply directly because it is concerned with the long-run steady-state behavior of stationary models. If we were to act as if that perspective were relevant, then we would presumably use the long-run average arrival rate and necessarily use one fixed staffing level throughout time. Needless to say, with typical time-varying demand such as in Figure 1, that approach - called the simple stationary approximation (SSA) - usually fails badly, producing alternating periods of understaffing and overstaffing.

The Base Model. To represent a single-skill call center with time-varying demand, we consider the $M_t/GI/s_t + GI$ queueing model. The initial M_t indicates that the arrival process is assumed to be a nonhomogeneous Poisson process with (deterministic) arrival-rate function $\lambda(t)$; i.e., the arrival rate at time t is $\lambda(t)$ and the number of arrivals in the interval $[t_1, t_2]$ has a Poisson distribution with mean $\int_{t_1}^{t_2} \lambda(t)\, dt$. Because of the commonly occurring daily cycle, it is natural to assume that $\lambda(t)$ is a periodic function, but that extra assumption is not too important because we usually are concerned with performance within a single day.

The first GI in the model indicates that the service times are independent and identically distributed (i.i.d.), independent of the arrival process, each distributed as a random variable S with cumulative distribution function (cdf) $G(x) \equiv P(S \le x)$ having finite mean $\mu^{-1} \equiv E[S]$. The s_t in the model indicates that the number of servers is allowed to be time-dependent, which we assume is a deterministic function $s(t)$, which is for us to determine. We assume that there is unlimited waiting space and that customers enter service in order of arrival.

The final $+GI$ in the model indicates that we allow customer abandonments. We assume that each waiting customer may elect to abandon before starting service, but no customer abandons after service has begun. We assume that customer times to abandon after arrival are i.i.d. random variables, independent of the arrival process and the service times, each distributed as a random variable T with cdf $F(x) \equiv P(T \le x)$ having finite mean $\theta^{-1} \equiv E[T]$. The independence assumption is realistic for the invisible queues usually occurring in call centers; then customers do indeed make abandonment decisions without knowing what other customers are doing. If our targeted quality of service is high, then we might elect to leave abandonment out of the model, but it often is better to take account of customer abandonment when it is present, because it can reduce the required staffing level. More importantly, as we will explain next, taking account of customer abandonment can actually make analysis easier!

Solution Methods. In general, the $M_t/GI/s_t + GI$ model is difficult to analyze mathematically, so that the staffing problem is challenging. However, there is one special case that is amazingly tractable: the Markovian $M_t/M/s_t + M$ model in which $\theta = \mu$ (the individual abandonment rate equals the individual service rate). In that special case, the stochastic process representing the number of customers in the system is distributed the same as for the associated infinite-server $M_t/M/\infty$ model, which is tractable, as we explain in §6. The whole problem becomes very manageable if we can work with that special case.

Measurements show that it can be important to consider non-exponential service-time and time-to-abandon distributions; see Brown et al. [1]. In practice, both distributions can be non-exponential, but the non-exponentiality in the time-to-abandon distribution has a greater impact upon performance; e.g., see Whitt [19].

For any given staffing function $s(t)$, it is not difficult to analyze the performance of the $M_t/GI/s_t+GI$ model by computer simulation. For the special Markovian cases $M_t/M/s_t+M$ and $M_t/M/s_t$, with or without abandonment, where the cdf's G and F are exponential (but we need not have $\theta = \mu$), the number in system is a nonstationary continuous-time Markov

chain (CTMC). For that special case, we can calculate the the transition function of the CTMC numerically by solving a system of ordinary differential equations (after truncating the state space at an appropriate level). Both simulation and the nonstationary-CTMC-ODE approaches have been used extensively over the years.

However, with these last two approaches, it remains to examine the extraordinarily large number of alternative staffing functions $s(t)$. Hence, until recently, those computational approaches have only been applied to evaluate alternative pre-determined staffing strategies. Feldman et al. [5] show that the computational approaches can be used to identify a good staffing function in a remarkably efficient manner, as we will explain in §7. But before we discuss that, we review the traditional way to cope with time-varying demand.

4. The Pointwise Stationary Approximation (PSA)

There is a long history of using queueing models to set staffing requirements in the face of time-varying demand. The classical call center was a group of telephone operators. In the early days of telephony, a human telephone operator set up each telephone call.

The standard way to cope with time-varying demand is to use a *pointwise stationary approximation* (PSA) - it provides a time-dependent description of performance based on the steady-state behavior of a stationary model, using the arrival rate and other model parameters that prevail at the time at which we want to describe the performance. That is, we approximate the distribution of the number of customers in the system at time t in the $M_t/GI/s_t + GI$ model by the steady-state distribution of the number of customers in the associated $M/GI/s+GI$ model, having the same service-time and time-to-abandon distributions, but with the (constant) arrival rate and number of servers equal to the values of the functions $\lambda(\cdot)$ and $s(\cdot)$ at time t. The term PSA was coined by Green and Kolesar [8], who conducted research in a series of papers investigating how it and variants perform.

Whitt [18] showed that PSA is asymptotically correct as the arrival rate changes less rapidly; a proper formulation is not quite as obvious as the basic idea. Massey and Whitt [16] went further to develop asymptotic "uniform-acceleration" asymptotic expansions, where PSA appears as the leading term. From the expansions, we can see when PSA will perform well: when the second and higher terms are negligible.

5. Staffing with Stationary Models

Given that we do apply PSA (or use an alternative method, such as the modified-offered-load approximation to be discussed in §6), we succeed in replacing our initial $M_t/GI/s_t+GI$ model by a stationary $M/GI/s+GI$ model. With PSA, at time t, we use the limiting steady-state distribution for the model with fixed arrival rate $\lambda(t)$. But even the stationary $M/GI/s+GI$ model is challenging in general. The Markovian cases $M/M/s$ (Erlang-C or delay model) and $M/M/s+M$ (Erlang-A or Palm model) are not difficult to analyze, because the number of customers in the system is a birth-and-death process. The $M/M/s+GI$ model is substantially more complicated, but it too can be analyzed exactly; see Zeltyn and Mandelbaum [21]. Easily-computed approximations for all standard performance measures in the $M/GI/s+GI$ model have been provided by Whitt [19].

The Normal Approximation. Experience has shown that, when the offered load is not too small (say at least 5) and the targeted quality of service is high, the number of customers in the system is approximately normally distributed. A revealing derivation of the normal approximation is to first approximate the $M/GI/s$ and $M/GI/s+GI$ models by an infinite-server $M/GI/\infty$ model, having the same arrival rate and the same service-time distribution. The steady-state number of busy servers in the $M/GI/\infty$ model has a Poisson distribution with mean equal to the offered load $a \equiv \lambda E[S]$, independent of the service-time distribution beyond its mean. The Poisson distribution in turn can be approximated by the normal distribution. Since the actual distribution is Poisson, the variance necessarily equals the mean, so that the offered load $a \equiv \lambda E[S]$ is the only parameter in the normal approximation.

With abandonments, there is additional justification for this infinite-server approximation: If customers abandon at the same rate they are served, i.e., if $\theta = \mu$, then the steady-state number of customers in the Markovian $M/M/s+M$ model has the same distribution as the number of customers in the associated $M/M/\infty$ model. (We have already observed that this property holds in the more general time-dependent setting.)

The Square-Root-Staffing Formula. From the normal approximation, we immediately obtain the *square-root-staffing formula*:

$$s = a + \beta\sqrt{a} \ , \tag{5.1}$$

where $a \equiv \lambda E[S]$ is the offered load - the mean number of busy servers in the $M/GI/\infty$ model - and β is a parameter reflecting the quality of service (QoS). A feasible integer staffing level is the least integer greater than or equal to s in (5.1).

To specify the QoS parameter β, it is convenient to focus on the delay probability, i.e., the probability that a customer must wait before starting service. With the normal approximation, we can directly relate the QoS parameter β in (5.1) to any desired steady-state delay-probability, which we denote by α. Letting Q be the steady-state number of busy servers in the infinite-server model, we approximate the steady-state delay probability α by

$$\alpha \equiv P(Delay) \approx P(Q \geq s) = P\left(\frac{Q-a}{\sqrt{a}} \geq \frac{s-a}{\sqrt{a}}\right) \approx 1 - \Phi(\beta) \ , \tag{5.2}$$

where Φ is the cdf of the standard (mean 0 and variance 1) normal distribution.

Many-Server Heavy-Traffic Limits. In the actual $M/GI/s+GI$ model, the steady-state number of customers in the system is usually not exactly normally distributed. Thus, it is often desirable to refine the normal approximation outlined above. Fortunately, there is an effective way to do so based many-server heavy-traffic limits, as in Halfin and Whitt [11], Garnett et al. [7], Whitt [20] and references therein.

The idea is to let $s \to \infty$ and $\lambda \to \infty$, while leaving the service-time cdf G and the time-to-abandon cdf F unchanged. (Note that this is exactly how a typical call center becomes large.) But we need to specify how the limits for λ and s are related. Halfin and Whitt showed for the $M/M/s$ model that we should let $s \to \infty$ and $\lambda \to \infty$, so that

$$\frac{s-a}{\sqrt{a}} \to \beta \ , \tag{5.3}$$

where again $a \equiv \lambda/\mu = \lambda E[S]$ is the offered load. In that limit, the steady-state delay probability $\alpha \equiv \alpha(\lambda, \mu, s)$ in the $M/M/s$ model approaches a limit strictly between 0 and 1. (Such a limit holds if and only if (5.3) holds.)

This implies that the delay probability is a good performance measure, because it tends to have meaning independent of scale. That is not true for most other performance measures. For example, the mean waiting time is asymptotically of order $1/\sqrt{s}$ in the limiting regime (5.3).

From the defining limit in (5.3), we see that the many-server heavy-traffic regime also produces a square-root-staffing law, for in the limit we have $s \approx a + \beta \cdot \sqrt{a}$, which coincides with (5.1).

As a consequence of the many-server heavy-traffic limit for the $M/M/s$ model, there is a continuous strictly increasing function mapping the QoS parameter β into the limiting delay probability α, now commonly called the *Halfin-Whitt delay function*:

$$P(Delay) \equiv \alpha \approx HW(\beta) \equiv [1 + (\beta\Phi(\beta)/\phi(\beta))]^{-1} \ , \quad 0 < \beta < \infty \ , \tag{5.4}$$

where, again, Φ is the cdf and ϕ is the associated probability density function (pdf) of the standard normal distribution. Jennings et al. [13] proposed using the Halfin-Whitt delay function in (5.4) instead of the normal delay function in (5.2) to represent the relation between β and α.

For the stationary Markovian $M/M/s+M$ model with customer abandonment and general abandonment rate θ, Garnett et al. [7] established a corresponding many-server heavy-traffic limit and showed that a corresponding continuous strictly increasing function maps the QoS parameter β and the ratio of the abandonment rate to the service rate, $\theta_{rat} \equiv \theta/\mu$, into the limiting delay probability α. This is now commonly called the *Garnett delay function*:

$$P(Delay) \equiv \alpha \approx Garnett(\beta, \theta_{rat}) \equiv \left[1 + \sqrt{\theta_{rat}} \cdot \frac{h(\beta/\sqrt{\theta_{rat}})}{h(-\beta)}\right]^{-1} \ , \quad -\infty < \beta < \infty \ , \tag{5.5}$$

where $h(x) \equiv \phi(x)/(1 - \Phi(x))$ is the *hazard rate* of the standard normal distribution.

The QoS parameter β can be based on the targeted probability of delay, α, because they can be related, as shown in Figure 2 below. We plot the Garnett delay function for five different values of $\theta_{rat} \equiv \theta/\mu$: 1/16, 1/4, 1, 4 and 16. When the abandonment rate is low, as in the case when θ_{rat} is equal to 1/16, the Garnett function is close to the Halfin-Whitt delay function, provided that β is not too small. Without customer abandonment, the system is unstable for $\beta \leq 0$; hence $HW(\beta) = 1$ for $\beta \leq 0$.

The normal approximation in (5.2) is also plotted in Figure 2 because, as noted previously, the infinite-server model coincides exactly with the $M/M/s+M$ model when $\theta = \mu$, i.e., when $\theta_{rat} = 1$. Figure 2 shows the error caused by using formula (5.2) when $\theta_{rat} \neq 1$. We see that the degree of abandonment, as measured by θ_{rat} can make a big difference in the staffing, when the quality of service is not too high.

6. The Infinite-Server Model

We have just described how we can staff with stationary models once we decide to apply PSA. Now we go on to consider when it makes sense to apply PSA and what to do if it does not.

An Amazingly Tractable Model. As emphasized in Eick, Massey and Whitt [3, 4] and subsequent papers, we can learn a lot about how to cope with time-varying demand in many-server queues by considering the associated infinite-server $M_t/GI/\infty$ model, having the same arrival process and the same service times, but infinitely many servers. When the required number of servers is large in the $M_t/GI/s_t + GI$ model and the quality of service provided is consistently good, it is evident that the $M_t/GI/\infty$ model should behave similarly. Indeed, for the $M_t/M/s_t + M$ model with $\theta = \mu$, the stochastic process representing the number of customers in the system has the same law (finite-dimensional distributions) as in the corresponding $M_t/M/\infty$ model, because the time-varying birth-and-death process has death rate $k\mu$ in state k, independent of the number of servers.

In general, from the perspective of staffing, the $M_t/GI/\infty$ model provides an *offered-load perspective*. The time-dependent distribution of the number of customers in the infinite-server system shows how many agents would actually be used if unlimited agents were available. We thus might staff in a time-varying way so that, at any time with the infinite-server model, there would be a small fixed probability that the demand would exceed the supply. Indeed, that is the approach advocated by Jennings et al. [13].

The main reason that it is good to look at the closely related $M_t/GI/\infty$ model is that it is so tractable, yielding insightful closed-form formulas. The number of busy servers at time t

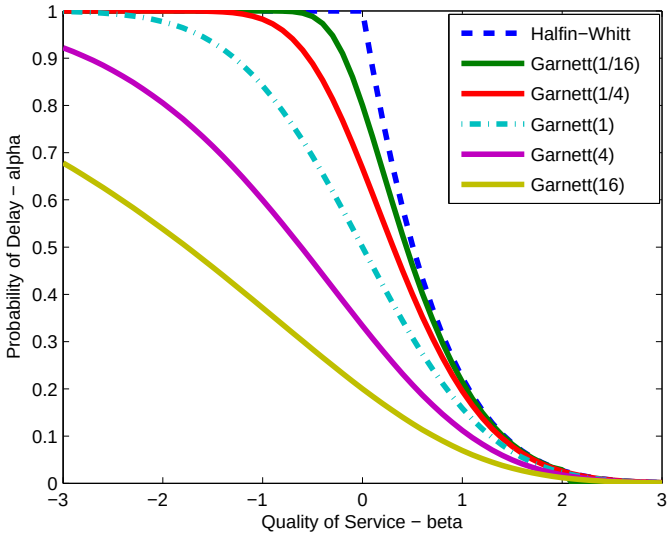

Figure 2: The Halfin-Whitt and Garnett functions mapping the QoS parameter β into the steady-state delay probability α. Five different values are considered for the parameter $\theta_{rat} \equiv \theta/\mu$: 1/16, 1/4, 1, 4 and 16.

in the time-dependent $M_t/GI/\infty$ model has a Poisson distribution with mean

$$m_\infty(t) = E[\lambda(t - S_e)]E[S] = E\left[\int_{t-S}^{t} \lambda(u)\, du\right] = \int_{-\infty}^{t} [1 - G(t-u)]\lambda(u)\, du\ , \tag{6.1}$$

where S_e is a random variable with the stationary-excess (or residual lifetime) cdf associated with the service-time cdf G, i.e.,

$$P(S_e \le t) \equiv \frac{1}{E[S]} \int_0^t [1 - G(u)]\, du, \quad t \ge 0, \quad \text{and} \quad E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}\ ; \tag{6.2}$$

see Theorem 1 of Eick et al. [3]. Equation (6.1) applies to the situation in which the system began operation in the distant past. If we want to start at time 0, we just define $\lambda(u) = 0$ for $u < 0$.

Since a Poisson distribution is characterized by its mean, the time-dependent distribution of the number of busy servers in the $M_t/GI/\infty$ model is completely characterized by the deterministic time-dependent mean function $m_\infty(t)$ in (6.1). Moreover, this Poisson distribution supports the normal approximation and the square-root staffing formula in (5.1), but now with a single modification: Instead of using the PSA mean $m_{PSA}(t) \equiv \lambda(t)E[S]$ for a at time t, we should use the (exact) time-dependent mean $m_\infty(t)$ in (6.1) in the associated $M_t/GI/\infty$ model. When the arrival rate is constant, both of these agree with the offered load $a = \lambda E[S]$ in the stationary $M/GI/\infty$ model.

Interpretation of the Mean Formula. We can interpret the three components of equation (6.1) by relating them to the PSA mean $m_{PSA} \equiv \lambda(t)E[S]$ and the lagged PSA mean $m_{LaggedPSA} \equiv \lambda(t - E[S])E[S]$. It is natural to expect that there should be a lag, because each customer remains in the system a random time (his service time) after his arrival time.

The first component of (6.1) shows that the PSA is correct except for a *random time lag*, with the random time lag being the stationary-excess variable S_e defined in (6.2), rather than just S. Thus, the mean $E[S_e]$ is a natural candidate for the approximate lag, but this interpretation is not direct, because the expectation appears outside the arrival-rate function instead of inside. We discuss how to *move the expectation inside* later in this section.

The second component of (6.1) shows that the mean is the integral of the arrival rate over a random interval before time t, specifically, over the interval $[t - S, t]$. The second formula can be interpreted as saying that PSA is correct except that $\lambda(t)$ should be replaced by an *average of the arrival rate in an interval before time* t, where the length of that interval should be about $E[S]$. The second formula also supports the notion of a time lag, showing that the

extent of the lag is related to the random service time S. (Here S appears instead of S_e, but the results are actually not inconsistent.) Finally, the third component of (6.1), an integral, shows that the exact mean can be computed numerically, given the arrival-rate function $\lambda(t)$ and the service time cdf G.

An Idealized Mathematical Model. The methods introduced so far apply to arbitrary arrival-rate functions, but we can gain insight into system physics from a structured mathematical model that captures the spirit of typical arrival-rate functions. The dynamic character of the demand function is reasonably characterized by a *sinusoid*:

$$\lambda(t) = \bar{\lambda} + A \cdot \sin(ct), \quad 0 \le t \le T \ , \tag{6.3}$$

where $\bar{\lambda}$, A and c are positive constants, with $\bar{\lambda}$ being the *average arrival rate*, $A \le \bar{\lambda}$ being the *amplitude* and c specifying the frequency or time scale; in particular, a cycle of this sinusoidal arrival-rate function is $2\pi/c$. An important issue is the rate of fluctuation in the arrival-rate function compared to the mean service time. We will fix the time scale by letting the mean service time be $E[S] = 1$. Then $m_{PSA}(t) \equiv \lambda(t)E[S] = \lambda(t)$. Having specified $E[S]$, the rate of fluctuation depends only on c.

Theorem 4.1 of Eick et al. [4] provides an explicit formula for the time-dependent infinite-server mean, namely,

$$m_\infty(t) = \bar{\lambda} + A\left(\sin(ct)E[\cos(cS_e)] - \cos(ct)E[\sin(cS_e)]\right) \ . \tag{6.4}$$

We can apply formula (6.4) to understand the system physics. For example, the extreme values of $\lambda(t)$ in (6.3) occur at times $t_\lambda = \pi/2c + \pi n/c$ for integer n. The corresponding extreme values of $m_\infty(t)$ in (6.4) occur at the later times

$$t_m = t_\lambda + \frac{1}{c}\tan^{-1}(E[\sin(cS_e)]/E[\cos(cS_e)]) \ , \tag{6.5}$$

where $E[S] = 1$, while the extreme values themselves are

$$m_\infty(t_m) = \bar{\lambda} \pm A((E[\cos(cS_e)])^2 + (E[\sin(cS_e)])^2)^{1/2} \ . \tag{6.6}$$

For the special case of exponential service times, (6.4) reduces to

$$m_\infty(t) = \bar{\lambda} + \frac{A}{1+c^2}[\sin(ct) - c \cdot \cos(ct)] \ , \tag{6.7}$$

while (6.5) and (6.6) become

$$t_m = t_\lambda + \frac{1}{c}\cot^{-1}(1/c) \quad \text{and} \quad m_\infty(t_m) = \bar{\lambda} \pm \frac{A}{\sqrt{1+c^2}} \ . \tag{6.8}$$

From equations (6.7) and (6.8), we can readily see how performance depends on the parameters.

Taylor-Series Approximations. We can also obtain important insights without making such strong sinusoidal assumptions. The first representation $m_\infty(t) = E[\lambda(t-S_e)]E[S]$ in (6.1) is complicated since the random time lag S_e appears inside the general function $\lambda(t)$, inside the expectation. We could move the expectation inside to produce the deterministic time lag $E[S_e]$ if $\lambda(t)$ were linear and, more generally, we could directly express $m_\infty(t)$ in terms of moments of S_e if the arrival-rate function $\lambda(t)$ were a polynomial. Of course, the arrival-rate function $\lambda(t)$ will usually not be a polynomial, but a smooth function can be approximately by polynomials in the neighborhood of individual arguments, by virtue of Taylor-series approximations. We proceed on this basis, following Eick et al. [3]; also see Massey and Whitt [15].

Suppose that we are interested in the performance at some time t. We can approximate the arrival-rate function in a time interval before time t by using a first-order Taylor-series approximation for $\lambda(t)$ centered at t:

$$\lambda(t-u) \approx \lambda(t) - \lambda^{(1)}(t)u \quad \text{for} \quad u \geq 0 \ , \tag{6.9}$$

where $\lambda^{(k)}(t)$ is the k^{th} derivative of $\lambda(t)$ evaluated at time t, from which we obtain from (6.1) the approximation

$$m_\infty(t) \approx \lambda(t - E[S_e])E[S] \ , \tag{6.10}$$

showing that $m_\infty(t)$ is approximately $m_{PSA}(t)$ modified by the *deterministic time lag* $E[S_e]$.

We can also consider a *second-order Taylor-series approximation* for the arrival-rate function $\lambda(t)$:

$$\lambda(t-u) \approx \lambda(t) - \lambda^{(1)}(t)u + \lambda^{(2)}(t)\frac{u^2}{2} \quad \text{for} \quad u \geq 0 \ , \tag{6.11}$$

from which Eick et al. [3, Theorem 9] obtain the approximation

$$m_\infty(t) \approx \lambda(t - E[S_e])E[S] + \frac{\lambda^{(2)}(t)}{2} Var(S_e)E[S] \ . \tag{6.12}$$

The first term in (6.12) is the first-order linear approximation given in (6.10), with the deterministic time lag, and the second term can be interpreted as a *deterministic magnitude shift*.

The Modified-Offered-Load (MOL) Approximation. It is often possible to apply the infinite-server model as the *first step in a two-step procedure* to generate a better approximation for the time-dependent performance measures and the required staffing than can be provided by either PSA or a direct application of the infinite-server model: This is the *modified-offered-load* (MOL) approximation, first proposed by Jagerman [12] for the time-dependent Erlang

loss model $M_t/M/s/0$, but the approach generalizes to the $M_t/GI/s_t/r_t + GI$ model with time-dependent staffing s_t, and possibly general time-dependent finite waiting room of size r_t. The MOL approximation is suggested and analyzed in Massey and Whitt [14, 15], Jennings et al. [13] and Feldman et al. [5].

With MOL, we approximate the performance in the $M_t/GI/s_t + GI$ model at time t by the performance in an associated stationary $M/GI/s+GI$ model, just as with PSA, except we replace the instantaneous offered load $m_{PSA}(t) \equiv \lambda(t)E[S]$ by the exact infinite-server mean $m_\infty(t)$. In other words, we use a stationary finite-server $M/GI/s + GI$ model at each t with the "modified" time-dependent arrival rate

$$\lambda_{MOL}(t) \equiv \frac{m_\infty(t)}{E[S]} \ , \tag{6.13}$$

where $m_\infty(t)$ is the infinite-server mean in (6.1). This enables us to apply both the square-root-staffing formula in (5.1) and the refinements to the normal approximation in (5.2) based on the Halfin-Whitt and Garnett delay functions in (5.4) and (5.5), but with a better approximation for the time-dependent offered load, now using $m_\infty(t)$ for a instead of $m_{PSA}(t)$.

Example 6.1. *Comparing PSA, Lagged PSA and MOL.* Consider the $M_t/M/s_t + M$ model with $\theta = \mu$, for which the infinite-server approximation $M_t/M/\infty$ is exact. In Figure 3 we compare three approximations for the time-dependent mean number of customers in the $M_t/M/s_t + M$ model: PSA, lagged PSA and MOL. It is easy to see that the MOL approximation applied to the $M_t/GI/\infty$ model is exact.

We assume a sinusoidal arrival-rate function, as in (6.3), where the average offered load is $\bar{\lambda} \cdot E[S] = 100$ and the amplitude is $A = 50/E[S]$. Following Green et al. [7], we assume that there is a half cycle over each day, corresponding to a peak but no trough as in Figure 1. Measuring time in minutes, a half cycle is of length 1440. We consider the difficult case in which $E[S] = 300$ minutes, making $c = (2\pi)300/2880 = 0.654$. Figure 3 shows that both PSA and lagged PSA are inadequate for this challenging example. At the time of peak congestion, which lags after the peak arrival rate, lagged PSA errs by providing 10 extra agents, because it does not include the magnitude shift in (6.12).

If we reduce the mean service time to $E[S] = 30$ minutes (and increase the arrival rate correspondingly to keep $m_{PSA}(t) \equiv \lambda(t)E[S]$ fixed), then the curves for lagged PSA and the exact mean values fall on top of each, so lagged PSA is sufficient, but ordinary PSA overstaffs during the initial period of rising demand, but understaffs in the final period of declining demand, by as much as three agents. If we reduce the mean service time even further to

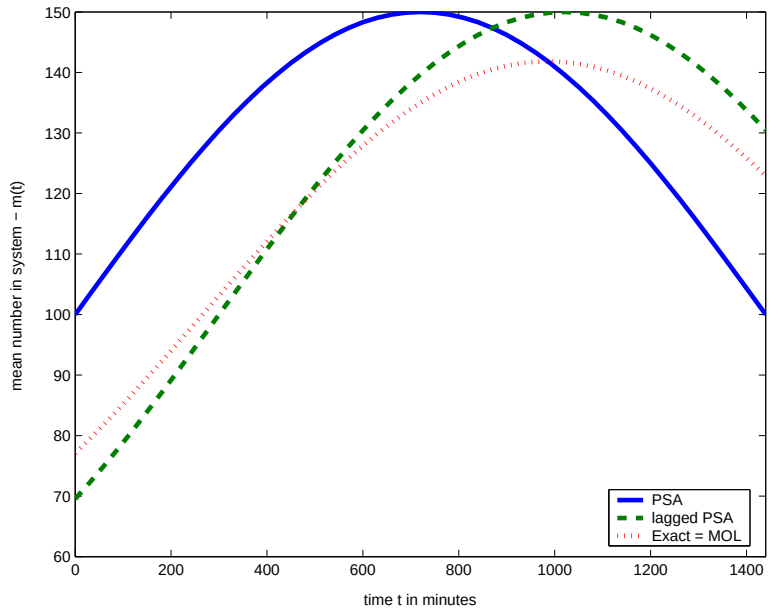

Figure 3: A comparison of PSA, lagged PSA and MOL for mean service time $E[S] = 300$ minutes with the sinusoidal arrival-rate function having $\bar{\lambda} = 100/E[S] = 1/3$, $A = 50/E[S] = 1/6$.

$E[S] = 3$ minutes (and again increase the arrival rate), then all three curves fall on top of each other, so that PSA is sufficiently accurate. ∎

7. Simulation-Based Iterative Staffing Algorithms

The MOL method suggests staffing in the $M_t/GI/s_t + GI$ model by (the least integer above) $s(t) = m_\infty(t) + \beta\sqrt{m_\infty(t)}$, where $m_\infty(t)$ is the time-dependent mean in the associated infinite-server model, which is easily computed via (6.1). We can apply (5.4) or (5.5) to choose the QoS parameter β given any desired delay probability α, but the validity of that refinement depends on exponential-distribution assumptions. More generally, we can apply simulation to search over possible values of β, aiming to reach time-stable performance with any desired α. We greatly simplify the search over staffing functions by restricting attention to the one-dimensional family indexed by β. For more complex models, we can even use simulation to estimate $m_\infty(t)$ as the time-dependent mean in the associated infinite-server model.

Feldman et al. [5] developed an alternative simulation-based iterative staffing algorithm (ISA) for the $M_t/GI/s_t + GI$ model, which works directly with the targeted delay probability α. The ISA approach can be extended directly to more general models, for which analytic results are unavailable. It is self-validating, because we directly verify that the performance will be as desired (assuming of course that the simulation model itself is appropriate). The ISA

keeps staffing constant over small subintervals. It does a sequence of iterations, starting with an infinite-server system at iteration 0. Let $Q_n(t)$ be the number of customers in the system at time t in iteration n; and let $s_n(t)$ be the staffing function in iteration n, with $s_0(t) = \infty$ (or some large value) for all t. Given the staffing function $s_n(t)$ in the n^{th} iteration, we perform multiple (say 5000) independent replications of the full planning period (the day) in order to estimate the distribution of $Q_n(t)$, the number of customers in the system, at each time t. Given that estimated distribution of $Q_n(t)$, we then create a new staffing function $s_{n+1}(t)$, by choosing the value at each time t that just meets the specified delay-probability target at each time t:

$$P(Q_n(t) \geq s_{n+1}(t)) \leq \alpha < P(Q_n(t) \geq s_{n+1}(t) - 1) \ . \tag{7.1}$$

Having found the new staffing function $s_{n+1}(t)$, we simulate again to find the distribution of $Q_{n+1}(t)$ for each t. We continue to iterate until there is negligible change (e.g., at most a single agent) in the staffing function from one iteration to the next. For the special case of the $M_t/M/s_t + M$ model, Feldman et al. proved that ISA (without estimation error) converges.

Feldman et al. showed through experiments that the ISA is remarkably effective in achieving time-stable performance in face of time-varying demand, even with long service times and relatively low QoS targets. They showed that ISA produces time-stable performance for delay-probability targets ranging from 0.1 to 0.9. They also showed that the ISA results are consistent with MOL, thus demonstrating the power of the infinite-server offered-load perspective.

Finally, Figure 4 shows that the refined staffing methods are helpful even for the relatively well behaved example in Figure 1, where the service times were quite short, having mean only 6 minutes. In this case, ISA and lagged PSA produce essentially the same desired time-stable result (as does MOL, which is not shown), but PSA fails to achieve time-stable performance, but not too badly. In real service systems, PSA or lagged PSA will often be effective, but the methods here help us understand when and why, and what to do when these simple methods fail.

Acknowledgments

I thank my co-authors for their important contributions. Recent research was supported by NSF grant DMI-0457095.

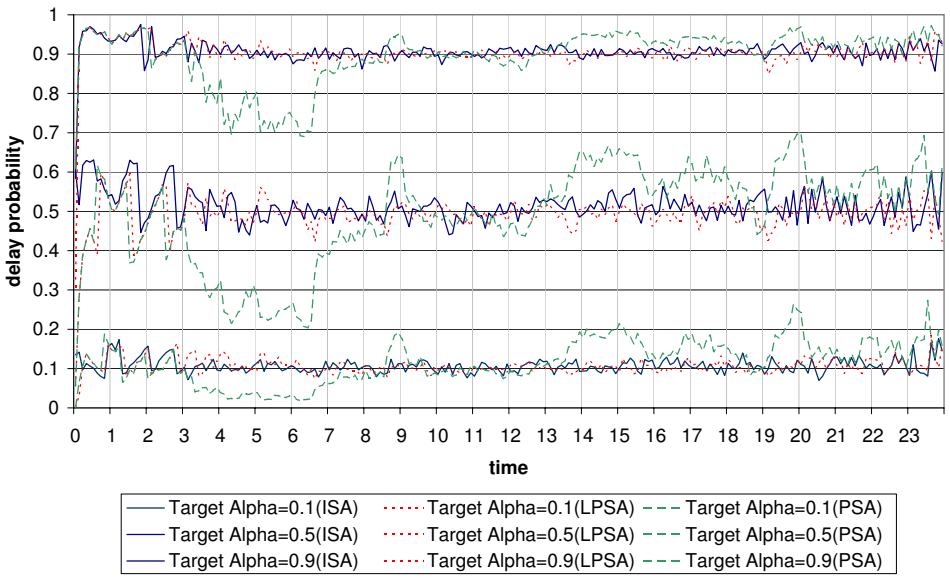

Figure 4: A comparison of the delay probabilities produced by ISA, PSA and lagged PSA for the example in Figure 1 (with mean service time 0.1 hour) for three different delay-probability targets: **0.1**, **0.5** and **0.9**.

References

[1] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, S. Zeltyn, L. Zhao, S. Haipeng. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association* (JASA) 100 (1), 36–50.

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

[3] Eick, S., W. A. Massey, W. Whitt. 1993a. The Physics of The $M_t/G/\infty$ Queue. *Operations Research* 41 (4), 731–742.

[4] Eick, S., W. A. Massey, W. Whitt. 1993b. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* 39 (2), 241–252.

[5] Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2007. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, forthcoming. Available at http://columbia.edu/∼ww2040.

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

[7] Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4 (3), 208–227.

[8] Green, L., P. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* 37 (1), 84–97.

[9] Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, forthcoming.

[10] Gurvich, I., W. Whitt. 2007. Service-level differentiation in many-server service systems: a solution based on fixed-queue-ratio routing. Submitted to *Operations Research*. Available at http://columbia.edu/∼ww2040.

[11] Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29 (5), 567–587.

[12] Jagerman, D. L. 1975. Nonstationary blocking in telephone traffic. *Bell System Tech. J.* 54 (3), 625–661.

[13] Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* 42 (10), 1383–1394.

[14] Massey, W. A., W. Whitt. 1994. An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Annals of Applied Probability* 4 (4), 1145–1160.

[15] Massey, W. A., W. Whitt. 1997. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems* 25 (1-4), 157–172.

[16] Massey, W. A., W. Whitt. 1998. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability* 8 (4), 1130–1155.

[17] Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management* 7 (4), 276–294.

[18] Whitt, W. 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rate increases. *Management Science* 37 (3), 307–314.

[19] Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Science* 51 (2) 221–235.

[20] Whitt, W. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. Available at http://columbia.edu/∼ww2040.

[21] Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue. *Queueing Systems* 51 (3-4), 361–402.