SERVER STAFFING TO MEET TIME-VARYING DEMAND

by

Otis B. Jennings,[1] *Avishai Mandelbaum,*[2] *William A. Massey*[3] *and Ward Whitt*[4]

AT&T Bell Laboratories

September 12, 1994

Revision: July 11, 1995

[1] School of Industrial and System Engineering, Georgia Institute of Technology, Atlanta, GA 30332. This work was done while serving as a consultant to AT&T Bell Laboratories.

[2] Faculty of Industrial and Management Engineering, The Technion, Haiffa 32000, ISRAEL. This work was done while serving as a consultant to AT&T Bell Laboratories.

[3] AT&T Bell Laboratories, Room 2C-120, Murray Hill, NJ 07974-0636.

[4] AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636.

Abstract

We consider a multiserver service system with general nonstationary arrival and service-time processes in which $s(t)$, the number of servers as a function of time, needs to be selected to meet projected loads. We try to choose $s(t)$ so that the probability of a delay (before beginning service) hits or falls just below a target probability at all times. We develop an approximate procedure based on a time-dependent normal distribution, where the mean and variance are determined by infinite-server approximations. We demonstrate that this approximation is effective by making comparisons with the exact numerical solution of the Markovian $M_t/M/s_t$ model.

Keywords: operator staffing, queues, nonstationary queues, queues with time-dependent arrival rates, multi-server queues, infinite-server queues, capacity planning.

1. Introduction

We propose a procedure to determine how many servers are needed, as a function of time, in a *nonstationary* stochastic service system. We assume that any number of servers can be assigned as a function of time in response to *projected* loads, but that the server assignment cannot be changed adaptively in real time in response to *observed* loads. This problem is often referred to as the *operator staffing problem;* e.g., see Andrews and Parsons (1989, 1991), Brigandi, Dargon, Sheehan and Spencer (1994), Buffa, Cosgrove and Luce (1976), Gaballa and Pearce (1979), Grassmann (1986, 1988), Chapter 7 of Hall (1991), Holloran and Byrn (1986), Kolesar (1986), Kolesar, Rider, Crabill and Walker (1975), Larson (1972), Quinn, Andrews and Byrne (1991), Segal (1974), Sze (1984) and Thompson (1993, 1994).

We investigate the operator staffing problem in the context of the $G_t/GI_t/s_t$ queueing model. There is unlimited waiting room and the service discipline is first-come first-served (FCFS). The arrival process is partially characterized by a time-dependent arrival-rate function $\lambda(t)$ and a time-dependent variability function $c_a^2(t)$ (to be defined below); the service times are mutually independent and independent of the arrival process; the cumulative distribution function (cdf) of the service time of an arrival at time t is $G_t(x)$; and the number of servers as a function of time is $s(t)$, which is subject to control, given the functions $\lambda(t)$, $c_a^2(t)$ and $G_t(x)$. This model contains as an important special case the fully specified $M_t/GI/s_t$ model with nonhomogeneous Poisson arrival process.

We do not consider the important problem of *forecasting uncertainty*, which means uncertainty about the arrival-rate function and other elements of the model. We also do not consider customer abandonments and retrials. These phenomena can be important, but we are primarily interested in providing a sufficiently high grade of service that these phenomena will be negligible.

We develop a procedure for determining the required number of servers, $s(t)$, as a function of time t. We do not discuss the subsequent problems of determining the actual work schedules to provide these servers. *We choose $s(t)$ so that the probability of delay* (the probability that an arrival at time t would have to wait before beginning service) *is approximately some target probability at all times.* Our experience indicates that other performance measures of interest, such as the mean and tail probabilities of the number in queue, do not fluctuate greatly when we control the delay probability.

2. A Challenging Example

Consider a Markovian $M_t/M/s_t$ system with exponential service times having mean 1 and a nonhomogeneous Poisson arrival process with sinusoidal arrival-rate function $\lambda(t) = 30 + 20\sin(5t)$. (We use the sinusoidal form for simplicity; our methods apply to general arrival-rate functions estimated from data.) Suppose that the target delay probability is 0.13. (We explain the use of 0.13 later.) This is a challenging (if not entirely realistic) example, because the arrival rate fluctuates rapidly relative to the mean service time, e.g., each cycle is $2\pi/5 \approx 1.256$. If we think of daily cycles, then the mean service time is about 0.8 days. Such long service times are uncommon, but they do occur, e.g., in equipment repair and loan processing. However, the point of this example is to show that there exist circumstances in which available procedures fail miserably, while our proposed procedure performs well.

To evaluate alternative staffing algorithms for the $M_t/M/s_t$ model, we calculate the time-dependent distribution of the number of customers present by numerically solving the Chapman-Kolmogorov forward equations (a system of ordinary differential equations), using a variant of the algorithm in Davis, Massey and Whitt (1995). We choose a finite capacity sufficiently large that it has negligible influence.

There are two relatively simple approximations that are natural to consider with time-varying arrival rates. First, if the arrival rate changes sufficiently slowly (relative to the service times), then it is natural to consider a *pointwise stationary approximation (PSA),* which is the steady-state distribution of the stationary M/M/s model using the instantaneous arrival rate $\lambda(t)$ at time t; e.g., see p. 178 of Hall (1991), Green and Kolesar (1991), Green, Kolesar and Svoronos (1991) and Whitt (1991). However, in this example the arrival-rate function varies much too quickly, as can be seen from Figure 1. Figure 1 plots $s(t)$ with the offered load (the arrival rate times the mean service time), which here is just $\lambda(t)$. Since the offered load varies from 10 to 50, PSA with a target delay 0.13 has the number of servers vary from 15 to 60, and the delay probability oscillates nearly over the full range between 0 and 1. Similar large oscillations occur with other performance measures such as the mean number in queue. (There is an initial startup effect, because we start the system empty at time 0, but this effect does not last long.) The average delay probability with PSA is 0.46.

Insert Figure 1 here

An alternative approximation is the *simple stationary approximation* (SSA), which uses the stationary model with the long-run average arrival rate. In contrast to PSA, SSA tends to be appropriate when the arrival rate changes very rapidly (again relative to the service times), provided that the arrival rate is nearly constant over the longer time scale over which the average is taken. In this example, the average arrival rate is 30. With a constant arrival rate of 30, the steady-state delay probability is 0.112 with 38 servers and 0.155 with 37 servers, so that 38 meets the target delay probability 0.13. However, with the actual arrival rate in this example, the delay probability oscillates between 0.04 and 0.30 when we always use 38 servers, as shown in Figure 2. Hence both PSA and SSA perform poorly for this example.

Insert Figure 2 here

As suggested in Section 5.3 of Whitt (1991), it is natural to look for approximations in between these two extremes. In particular, to determine the effective arrival rate at time *t*, it is natural to average the arrival rate over an appropriate interval before time *t*. With PSA, the interval has zero length; with SSA, the interval has infinite length. Assuming that there is no anticipated major queue buildup before time *t*, this interval should perhaps have a length of about 1-10 mean service times, but it is not clear precisely how this averaging should be done.

Our main idea is to use an infinite-server (IS) *approximation.* The IS approximation can be regarded as an averaging procedure that produces an effective arrival-rate at all times, as will be clear from Section 2. Moreover, the IS approximation clearly reveals the proper role of the service-time distribution in this averaging.

Figures 3 and 4 show the result of applying our IS algorithm to choose $s(t)$ to keep the delay probability below 0.13. The offered load is plotted with the number of servers in Figure 3. Even though the offered load fluctuates between 10 and 50, our algorithm dictates that the number of servers should vary only from 34 to 42. Figures 3 and 4 show the exact delay probability, the probability that there are at least five customers in queue, and the expected number in queue (not in service) as functions of time. Observe that the time-dependent delay probability is indeed close to the target value 0.13 at all times. Moreover, all three performance measures are nearly constant in time. (The observed fluctuations are primarily due to the fact that the change of one server changes the delay probability quite substantially.)

Insert Figures 3 and 4 here

We have just shown by example that PSA and SSA can perform poorly, while the IS approximation performs well. However, the example is not too realistic because the arrival rate fluctuates rapidly. When the arrival rate fluctuates more slowly, PSA performs much better. Indeed, for many applications the arrival rates fluctuate sufficiently slowly that PSA is excellent; e.g., see Green and Kolesar (1991) and Eick et al. (1993a). It is significant, though, that the IS approximation still works well when the arrival rate fluctuates more slowly. Then the IS approximation tends to coincide with PSA.

3. A Normal Approximation

So how do we achieve these results? We act as if there are always infinitely many servers and develop an approximation for the distribution of the number $Q_\infty(t)$ of busy servers at time t. Given an approximate distribution of $Q_\infty(t)$ for each t, we try to choose $s(t)$ so that

$$P(Q_\infty(t) \geq s(t)) \leq \alpha \quad \text{and} \quad P(Q_\infty(t) \geq s(t)-1) > \alpha \quad \text{for all } t \tag{1}$$

for some prescribed target probability α.

We use a *normal approximation* to solve (1). Let $N(m,\sigma^2)$ denote a normal random variable with mean m and variance σ^2. We approximate the distribution of $Q_\infty(t)$ by the distribution of $N(m(t), v(t))$ for a suitable mean function $m(t)$ and variance function $v(t)$ developed below. Hence, $[Q_\infty(t)-m(t)]/\sqrt{v(t)}$ is distributed approximately as $N(0,1)$, so that the number of servers is taken to be

$$s(t) = \left\lceil m(t) + z_\alpha \sqrt{v(t)} \right\rceil , \tag{2}$$

where $\lceil x \rceil$ is the least integer greater than x and z_α satisfies

$$P(N(0,1) > z_\alpha) = \alpha . \tag{3}$$

It is also natural to incorporate a refinement to account for the discreteness ($s(t)$ must be an

integer) as in (3.16) on p. 185 of Feller (1968). Thus, we actually use

$$s(t) = \left\lceil m(t) + 0.5 + z_\alpha \sqrt{v(t)} \right\rceil \tag{4}$$

instead of (2).

Formula (4) is the procedure when we allow the server levels to be changed at arbitrary times. In our programs implementing (4) we also allow changes to be made only at user specified times $t_1, t_2, \ldots$. Then $s(t)$ throughout the interval $[t_i, t_{i+1})$ is set equal to the maximum value of $s(t)$, computed by (4), over that interval. We also allow other constraints, such as servers being brought in or removed only in groups of a certain size. In all cases, $s(t)$ is the least integer such that all constraints are satisfied. With (4) it is easy to study the impact of such extra constraints.

We have evaluated the performance of our staffing algorithm based on (4) by making comparisons with exact numerical solutions for the $M_t/M/s_t$ model with nonhomogeneous Poisson arrival process and a fixed exponential service-time distribution. (In this case, our approximations yield $v(t) = m(t)$.) Thus, for Markovian $M_t/M/s_t$ models we compute the exact delay probabilities and related performance measures, such as the mean number in queue, at the same time that we compute the server staffing levels. For the examples we consider (e.g., like the one above), all these calculations are typically completed in a few seconds.

Normal approximations for multi-server models have a long history, e.g., see p. 191 of Feller (1968), Iglehart (1965), Section II.2 of Newell (1973), Whitt (1984, 1992), and references therein. The normal approximation with the IS mean and variance can be developed via heavy-traffic limit theorems in which the entire arrival-rate function increases (and the number of servers increases if there are only finitely many); see Glynn and Whitt (1991) and references therein. Glynn and Whitt only discuss the stationary model, but the argument there extends to nonstationary models. Central limit theorems for nonstationary arrival processes are discussed in Massey and Whitt (1994a).

The normal approximation performs well when the offered load, and thus the number of servers, is relatively large, because then we are in the domain of the heavy-traffic limit theorems, but the normal approximation also performs remarkably well when the number of servers is quite small, e.g., 5. This should not be too surprising, because the central limit theorem also often performs remarkably well for small sample sizes.

Since the normal approximation is natural and has a substantial history, our main contribution is providing an appropriate mean function $m(t)$ and variance function $v(t)$ in (4). Our main idea is to base these functions on the exact values, and subsequent approximations, in the associated $G_t/GI_t/\infty$ infinite-server model. We give the details in Sections˜5 and 6.

As background, it is useful to review how the IS approximation and the normal approximation (4) work for the stationary M/M/s model. The IS approximation is just the Poisson distribution. Table 1 displays the number of servers needed to meet several target blocking probabilities as a function of the offered load.

Insert Table 1 here

Notice that the three methods differ by at most one server in all cases. The difference is small in part because a change of one server changes the delay probability quite substantially. This is illustrated in Table 2, where we show the actual delay probabilities associated with a target delay probability $\alpha = 0.005$ as well as the delay probabilities with one fewer and one more server. Additional numerical comparisons for stationary models appear in Whitt (1992).

Insert Table 2 here

4. Refined Delay Probability Approximations

The IS approximation here extends the IS approximation for stationary models in Section 2 of Whitt (1992). As pointed out there, even though it is natural to interpret the target probability α in (1) as a direct approximation for the probability of delay in the $G_t/GI_t/s_t$ model, it is possible to obtain a better approximation for the probability of delay in the finite-server $G_t/GI_t/s_t$ model by exploiting the relation between the tail probabilities in the two models. For the stationary GI/M/s delay model, the probability of delay converges to a constant as the offered load increases when the number of servers (which also must increase) is specified according to (2) or (4), by virtue of a heavy-traffic limit theorem in Halfin and Whitt (1981). For example, as the arrival rate and number of servers increase in the M/M/s model with (2) holding, the probability of delay approaches

$$p_D(\alpha) \equiv [1 + \sqrt{2\pi}\, z_\alpha (1-\alpha)\exp(z_\alpha^2/2)]^{-1} \ . \tag{5}$$

for α in (1) and z_α in (3). Moreover, Whitt (1992, 1993) showed that (5) provides an excellent approximation for the actual steady-state delay probability even when the number of servers is not large; e.g., see Table 13 of Whitt (1993).

Hence, *when the target probability is α in (1), we propose $p_D(\alpha)$ in (5) as a better approximation for the actual delay probability.* Table 3 displays typical values of α and $p_D(\alpha)$. We already applied (5) in the example with rapidly fluctuating sinusoidal arrival rate in Figure 1. We actually used $\alpha = 0.1$ and thus $z_\alpha = 1.282$. By Table 3, this corresponds to a delay probability of 0.13, which was our stated target value. Table 3 shows that $p_D(\alpha)$ is very close to α when α is small.

Insert Table 3 here

The initial approximation for the delay probability at time t using (1)–(4) is the target IS delay probability α and a second refined approximation is $p_D(\alpha)$ in (5). Another refined approximation is the *modified offered load* (MOL) approximation, as in Jagerman (1975), Massey and Whitt (1994b) and Davis et al. (1995). (These papers consider loss models, but the MOL approximation can be applied in the same way to delay systems.) The basic MOL approximation for the distribution of the number of customers in an $M_t/M/s_t$ system at time t is the steady-state M/M/s distribution with $s(t)$ as the number of busy servers and $m(t)/s(t)$ as the traffic intensity, where $m(t)$ is the time-dependent IS mean. Our experience indicates that the MOL approximation yields more accurate approximations for the delay probability than the IS approximation refined by (5), but the improvement tends not to be important for staffing. The MOL approximation applies directly to $M_t/M_t/s_t$ models, but can be used more generally. We do not discuss MOL approximations further.

5. The Mean Function

For the $G_t/GI_t/\infty$ model, assumed to start empty in the distant past, the mean number of busy servers as a function of time is (exactly)

$$m(t) = \int_{-\infty}^{t} G_u^c(t-u)\,\lambda(u)\,du\ , \tag{6}$$

where $G_u^c(t) \equiv 1 - G_u(t) \equiv P(S(u) > t)$ with $S(u)$ being the service time of an arrival at time u and $\lambda(u)$ is the arrival rate at time u. Formula (6) is reasonably well known for the case of nonhomogeneous Poisson arrivals, but formula (6) actually remains valid *without* the Poisson assumption, provided that the arrival-rate function $\lambda(t)$ is well defined. The case of Poisson arrivals is reviewed in Section 1 of Eick, Massey and Whitt (1993a). The extension to non-Poisson arrivals is treated in Theorem 2.1 and Remark 2.3 of Massey and Whitt (1993).

Thus, we can compute $m(t)$ using (6), either analytically or numerically, depending on the

nature of the functions λ and G_t. In the rest of this section we discuss further simplifying assumptions and/or approximations that can facilitate the computation. If the service-time distribution G_t does not depend on t, then (6) can be rewritten as

$$m(t) = E[\int_{t-S}^{t} \lambda(u)\,du] = E[\lambda(t-S_e)]E[S] , \tag{7}$$

where S is a generic service time and S_e is a generic service-time stationary-excess random variable, i.e.,

$$P(S_e \le t) = \frac{1}{E[S]}\int_0^t P(S > u)\,du ; \tag{8}$$

see Theorem 1 of Eick et al. (1993a). Note from formulas (6) and (7) that the IS mean $m(t)$ can be regarded as a *smoothing* of $\lambda(t)$ multiplied by the mean service time. (This is the averaging referred to in Section 1.) Our experience is that the IS approximation succeeds in finding an *appropriate smoothing*. This is illustrated by the example in Section 2.

We now develop an approximation for time-dependent service times, assuming that $G_t(x)$ changes relatively slowly compared to $\lambda(t)$. We note that this often seems to be the case. For example, measurements indicate that the holding-time distribution of telephone calls changes throughout the day, having a bigger mean in the evening than in the day, but this rate of change is slow compared to the rate of change of the arrival rate $\lambda(t)$. Assuming that the service-time distribution does indeed change relatively slowly, we introduce a natural approximation to (6) based on (7), namely,

$$m(t) \approx E[\lambda(t-S_e(t))]E[S(t)] , \tag{9}$$

where $S(t)$ has cdf G_t and $S_e(t)$ has cdf

$$P(S_e(t) \le x) = \frac{1}{E[S(t)]}\int_0^x P(S(t) > u)\,du , \quad x > 0 . \tag{10}$$

We have associated the service-time distribution with customers, by letting it be determined upon arrival. If instead we want to think of each server completing service at rate $\mu(t)$ at time t, then we suggest approximating $S(t)$ by an exponential random variable with mean $1/\mu(t)$. Of course, if the service rate changes relatively quickly, then we could use the exact (time-dependent exponential) cdf

$$G_t^c(x) = e^{-\int_t^{t+x} \mu(u)\,du}, \; x \geq 0 . \tag{11}$$

To systematically apply our staffing procedure (1)–(4), we need to be able to efficiently compute the mean $m(t)$ as a function of t for all times t of interest. The formulas (6)–(11) can be used directly, but they are not especially convenient for this purpose. However, if we make further assumptions, then we can obtain very convenient algorithms to compute $m(t)$.

We can obtain an efficient algorithm if we make simplifying assumptions about the service-time distribution. First, if the service-time distribution is exponential with constant rate μ, then $m(t)$ can be obtained simply by solving the *ordinary differential equation* (ODE)

$$m'(t) = \lambda(t) - m(t)\mu , \tag{12}$$

where $m(0)$ is given, e.g., $m(0) = 0$ if we start empty at time 0; see Corollary 4 on p. 737 of Eick et al. (1993a). (Note that (12) extends to time-dependent service rates as well.) *It is significant that (12) applies with a general arrival-rate function.*

More generally, if there is a fixed phase-type service-time distribution with k phases, then the mean $m(t)$ can be obtained by solving a system of k ODEs, as was done in Davis et al. (1995); also see Massey and Whitt (1994b). The $M_t/PH_k/\infty$ model is equivalent to a special case of the $(M_t/M/\infty)^k/M$ network treated in Section 8 of Massey and Whitt (1993).

Even for the numerical examples in this paper, involving the $M_t/M/s_t$ model with sinusoidal arrival-rate functions, we use the ODE (12) to calculate the mean $m(t)$. It is somewhat faster than

repeatedly evaluating the explicit formula for $m(t)$ given in (15) of Eick et al. (1993b). (Computing $\lambda(t)$ requires evaluating one trigonometric function, while computing $m(t)$ directly from (15) of Eick et al. (1993b) requires evaluating two.)

Second, instead of making special assumptions about the service-time distribution, we can assume special structure for the arrival rate function $\lambda(t)$. Explicit formulas for $m(t)$ when $\lambda(t)$ is sinusoidal are given in Eick et al. (1993b). Explicit formulas for $m(t)$ when $\lambda(t)$ is a polynomial or a step function are given in Eick et al. (1993a).

However, in practice $\lambda(t)$ typically will not have a well-defined mathematical structure. What we want is a convenient algorithm to calculate $m(t)$ when $\lambda(t)$ is a general function estimated from data. We can treat general arrival-rate functions arising in practice by making polynomial (e.g., linear or quadratic) approximations to the arrival-rate function $\lambda(t)$ over subintervals. For example, we can compute $m(t)$ over the interval $[(k-1)c, kc)$ by using a polynomial approximation to $\lambda(t)$ over the interval $[(k-j)c, kc)$, where c is an appropriate constant such as 5 mean service times and j is perhaps about 8. The polynomial approximation for $\lambda(t)$ can be fit to data using least squares or iterated weighted least squares, as in Massey, Parker and Whitt (1995). (Iterated weighted least squares yields the maximum likelihood estimator, but ordinary least squares usually performs almost as well.)

The ODEs provide a simple effective way to calculate $m(t)$ for general $\lambda(t)$. Nevertheless, in some situations it may be desirable to have even more elementary approximations. For this purpose, we can exploit the *quadratic approximations* (QUAD-S) in Section 3 of Eick et al. (1993a) to further simplify (9). In particular, we can use the approximation

$$m(t) \approx \lambda(t - E[S_e(t)])E[S(t)] + \frac{\lambda''(t)}{2}\mathrm{Var}[S_e(t)]E[S(t)] , \qquad (13)$$

where, by (2) of Eick et al. (1993a),

$$E[S_e(t)] = E[S(t)^2]/2E[S(t)] \tag{14}$$

and

$$\mathrm{Var}(S_e(t)) = (E[S(t)^3]/3E[S(t)]) - (E[S(t)^2]/2E[S(t)])^2 , \tag{15}$$

with $\lambda''(t)$ being computed based on a suitably smoothed estimate of λ. A convenient simple approximation for $m(t)$, which tends to be appropriate when $\lambda(t)$ changes slowly, is (13) without the second term, i.e., (13) with $\lambda''(t) \approx 0$.

An even simpler approximation is the IS *pointwise stationary approximation* (PSA), namely,

$$m(t) \approx \lambda(t)E[S(t)] . \tag{16}$$

The quadratic approximation QUAD-S in (13) differs from PSA in (16) by a *time lag* $E[S_e(t)]$ and a *space shift* $\lambda''(t)\mathrm{Var}[S_e(t)]E[S(t)]/2$. We call (13) without the space shift *shifted PSA*.

We regard PSA in (16) as the natural initial approximation to $m(t)$. Evidently PSA has often been used in applications. Our goal is to go beyond (16) and indicate when going beyond (16) is desirable. The refinement (13) is only significantly better than (16) if the time lag and space shift are nonnegligible. *Since the time lag and space shift can easily be computed explicitly, we can easily apply (13) to determine when refinements to PSA should make a significant difference.*

In summary, (4) with the exact mean $m(t)$ in (6) (and the variance discussed below) is the IS approximation for the original finite-server problem. If we can assume a fixed exponential service-time distribution, then we can obtain $m(t)$ for any arrival-rate function by solving the ODE (12). If we use a phase-type service-time distribution, then we can obtain $m(t)$ by solving a corresponding system of ODEs. We propose (13) as a convenient approximation to (6), and PSA in (16) as a convenient approximation to (13). The time lag and space shift in (13) help us evaluate the importance of doing more than PSA.

6. The Variance Function

In the case of (nonhomogeneous) Poisson arrivals, the number of busy servers at time t in the infinite-server model has a Poisson distribution with the mean $m(t)$ in (6); e.g., see Theorem 2.1 of Massey and Whitt (1993). (Here the Poisson property of the arrival process *is* needed.) Hence, in this case, the variance $v(t)$ equals the mean $m(t)$. More generally, we regard $m(t)$ as the default value for $v(t)$. Our purpose in this section is to develop better approximations for $v(t)$ to cover cases when the arrival process is not nearly Poisson.

Paralleling the treatment of the stationary model in Section 2 of Whitt (1992), we suggest the approximation

$$v(t) \approx z(t)\, m(t) , \tag{17}$$

where

$$z(t) = 1 + \frac{(c_a^2(t)-1)}{E[S(t)]} \int_0^\infty [1-G_t(x)]^2\, dx , \tag{18}$$

$$c_a^2(t) \approx \frac{\mathrm{Var}[A(t)-A(t-\eta)]}{\int_{t-\eta}^{t} \lambda(u)\, du} \tag{19}$$

for some suitable $\eta > 0$; in (19) $A(t)$ counts the number of arrivals before time t. The function $z(t)$ in (18) is a time-dependent generalization of the *heavy-traffic peakedness* in (13) of Whitt (1992), while $c_a^2(t)$ in (19) is a time-dependent generalization of the *asymptotic variability parameter* in (10) of Whitt (1992). For a stationary G/G/∞ model, the peakedness is the ratio of the variance to the mean of the steady-state number of busy servers. Hence, without time dependence, and with the true peakedness, (17) would be exact by definition. A detailed expression for the peakedness is available in considerable generality, e.g., see Eckberg (1983), but it is complicated. (See Chapter 7 of Wolff (1989) for background.) Hence we suggest (18) as a useful approximation.

The idea in (19) is to obtain $c_a^2(t)$ as the variance to mean ratio of the counting process in the neighborhood of *t*. We look backward from *t*, since $Q(t)$ depends on the arrivals before *t*. The estimate should be robust in the interval length η if $A(t)-A(t-\eta)$ is not too small and the variability is changing relatively slowly. If the arrival process is approximately a renewal process before time *t*, where interarrival times have squared coefficient of variation (SCV, variance divided by the square of the mean) c_a^2, then $c_a^2(t)$ should be just c_a^2.

The representation (17)–(19) is convenient because modellers often have ideas about the relevant peakedness $z(t)$ or the asymptotic variability parameter $c_a^2(t)$ from extensive experience with stationary models. In that case, we need not do the hard work to directly calculate or estimate $c_a^2(t)$ via (19). Thus, (17)—(19) permit the degree of stochastic variability to be specified directly in a relatively familiar way.

7. A Start-Up Example

In this section we consider a stationary M/G/s delay model with arrival rate λ starting up from an empty initial state. In this case, the mean function for the IS model has a very simple form, namely,

$$m(t) = \lambda E[S]P(S_e \le t) = \lambda \int_0^t G^c(u)\,du \tag{20}$$

for S_e in (8). The well-known insensitivity of the steady-state distribution to the service-time distribution beyond its mean is seen by letting $t \to \infty$ in (20), because $m(t) \to \lambda E[S]$ as $t \to \infty$.

Suppose that we consider the special case of arrival rate $\lambda = 100$ and exponential service times with mean μ^{-1}. Then

$$v(t) = m(t) = 100(1-e^{-\mu t}) \ , \ t \ge 0 \ . \tag{21}$$

Furthermore, suppose that $\mu = 1$ and we consider changing the number of servers only at integer time points. Then we must meet the capacity constraints at *k* throughout the interval $[k-1, k)$.

If $\alpha = 0.05$ ($z_\alpha = 1.645$), then by (4) the number of servers required in successive intervals $[k-1, k)$ is $77, 103, 112, 115, 117, 117, 117, \ldots$ for $k \geq 1$. Over the first four intervals, the average number of servers required is 101.8, which represents an average savings of about 15% from the steady-state value of $s = 117$. Note that the steady-state value of $s = 117$ is also what PSA prescribes (using $\lambda(t)E[S]$ instead of $m(t)$ in (4)), so we clearly see the improvement provided by the IS approximation over PSA. Of course, no savings at all are obtained after time 4 (i.e., after 4 mean service times), so the interval $[0, 4]$ must be significant in the overall problem for the time-dependent analysis to be important.

We considered the specific case of service rate $\mu = 1$. More generally, from (21), we see that if staffing is done at times k/μ, then savings will occur only over the time interval $[0, 4\mu^{-1}]$; i.e., the savings will be proportional to μ^{-1}. Thus, if μ is increased from 1 to 10, then the savings will be reduced by a factor of 10.

Figure 5 displays the actual time-dependent delay probabilities for the $M_t/M/s_t$ model with the staffing levels above for the case $\mu = 1$.

Insert Figure 5 here

Since $\alpha = 0.05$, from Table 3 we would predict that the maximum delay probability is 0.062. This is remarkably close to the actual steady-state blocking probability of 0.064 with $s = 117$. Jumps in the delay probabilities are to be expected, since changing the number of servers by 1 has a significant impact. To see this, note that the steady-state delay probabilities with $s = 116$, 117 and 118 are, respectively, 0.078, 0.064 and 0.052.

8. Other Sinusoidal Examples

We conclude by considering several other $M_t/M/s_t$ examples with sinusoidal arrival rates.

We refer to Eick et al. (1993b) for the IS theory. We use sinusoidal models because they are convenient for illustrative purposes, not because we believe that they are especially realistic. It is no more difficult to solve an $M_t/M/s_t$ model with a general arrival-rate function using the ODE (12).

As before, we assume that the service times are all exponential with mean 1, so that time is measured in mean service times. We consider the four arrival rate functions and delay probability targets α shown in Table 4. The range of realized delay probabilities for the IS approximation is also shown in Table 4. Figures like Figures 1-3 appear in an unpublished original longer version of this paper.

The first arrival-rate function $\lambda(t) = 20 + 10\sin(t)$ has period $2\pi \approx 6.28$. For a daily cycle this corresponds to a mean service time of 3.82 hours. In this example the mean service time is relatively long compared to the period, as might occur in repair operations, but it is not nearly as long as the sinusoidal example given in Section 1. The time variation still has a significant impact on the mean $m(t)$ in the IS approximation, which is $20 + 5(\sin t - \cos t)$. The peak delay probability is about 0.13 as predicted by (5) and Table 3. The time-average number of servers required with IS is 26.91, which is slightly *less* than the 27 servers required for the stationary model with $\lambda = 20$. The peak number of servers with PSA is 38 compared to 35 when we apply (4) directly with $m(t)$. The average number of servers used is essentially the same, but the delay probabilities are much more erratic with PSA, ranging from 0 to about 0.7 for PSA as opposed to from 0.09 to 0.013 for the IS approximation. The average delay probability is 0.264 for PSA and 0.089 for the IS approximation.

The other performance measures also fluctuate much more with PSA. For example, the mean number in queue ranges from 0 to 6 with PSA, as opposed to from 0.25 to 0.40 with the IS approximation. We also considered the simple shifted PSA approximation $m(t) \approx \lambda(t - ES_e)$

based on (13). It is better than PSA but still substantially worse than (4). Shifted PSA has a maximum delay probability of 0.4 and an average delay probability 0.151.

To illustrate that PSA is much closer to the IS approximation with longer cycles, our second example has arrival rate $\lambda(t) = 400 + 40\sin(0.2t)$. The larger size associated with an average arrival rate of 400 illustrates that the exact $M_t/M/s_t$ calculation still works well for such systems. The cycle is now 10π, which corresponds to a mean service time of 0.768 hours = 46 minutes in a daily cycle. The range of delay probabilities using a target of $\alpha = 0.1$ ($z_\alpha = 1.282$) using the IS approximation is between 0.12 and 0.13. In contrast, the range for PSA is 0.06 to 0.26. The dynamic steady-state delay probabilities still vary substantially more with PSA, but only a little bit more with shifted PSA. The average blocking is 0.117 with IS, 0.123 with shifted PSA and 0.140 with PSA. In this example the simple shift of PSA gains most of the benefits of IS.

To illustrate that the IS approximation also works remarkably well with quite small systems, our third example is the arrival-rate function $\lambda(t) = 3 + 2 \sin t$ and target $\alpha = 0.1$. The range of delay probabilities is 0.06 to 0.12.

The IS approximation tends to work better when the target delay probability α is small, but it also works quite well when α is not small. To illustrate, our last example is the arrival rate function $\lambda(t) = 20 + 10 \sin t$, as before, but now with $\alpha = 0.4$, which from Table 3 we see corresponds to $z_\alpha = 0.2533$ and $p_D(\alpha) = 0.7177$. The realized peak delay probability for IS is about 0.58, but the quality of service is again quite constant.

We conclude by pointing out that the difference between $m(t)$ and $\lambda(t)E[S]$, and thus IS and PSA, would be greater if the service-time distribution were more variable. The case of a hyperexponential service-time distribution is discussed in Section 7 of Eick et al. (1993b).

Acknowledgment. We thank Linda Green and Peter Kolesar for their interest and helpful

comments.

References

Andrews, B. H. and H. L. Parsons, ‘‘L. L. Bean Chooses a Telephone Agent Scheduling System,’’ *Interfaces* 19 (1989), 1-9.

Andrews, B. H. and H. L. Parsons, ‘‘Establishing Telephone Agent Staffing Levels Through Economic Optimization,’’ *Interfaces* 23 (1993), 14-20.

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.

Buffa, E. S., M. J. Cosgrove and B. J. Luce, ‘‘An Integrated Work Shift Scheduling System,’’ *Decision Sciences*, 7 (1976), 620-630.

Davis, J. L., W. A. Massey and W. Whitt, ‘‘Sensitivity to the Service-Time Distribution in the Nonstationary Erlang Loss Model,’’ *Management Sci.* 42 (1995), to appear.

Eckberg, A. E., ‘‘Generalized Peakedness of Teletraffic Processes,’’ *Proc. Tenth Int. Teletraffic Congress*, Montreal, p. 4.4b.3, June 1983.

Eick, S. G., W. A. Massey and W. Whitt, ‘‘The Physics of the $M_t/G/\infty$ Queue,’’ *Oper. Res.*, 41 (1993a), 731-742.

Eick, S. G., W. A. Massey and W. Whitt, ‘‘$M_t/G/\infty$ Queues with Sinusoidal Arrival Rates,’’ *Management Sci.*, 39 (1993b), 241-252.

Feller, W. *An Introduction to Probability Theory and Its Applications*, vol. I, third edition, Wiley, New York, 1968.

Gaballa, A. and W. Pearce, ‘‘Telephone Sales Manpower Planning at Quantas,’’ *Interfaces* 9 (1979), 1-9.

Glynn, P. W. and W. Whitt, ‘‘A New View of the Heavy-Traffic Limit Theorem for Infinite-Server Queues,’’ *Adv. Appl. Prob.*, 23 (1991), 188-209.

Grassmann, W. K., ‘‘Is the Fact That the Emperor Wears No Clothes a Subject Worthy of Publication?’’ *Interfaces* 16 (1986), 43-51.

Grassmann, W. K., ‘‘Finding the Right Number of Servers in Real-World Queueing Systems,’’ *Interfaces* 18 (1988), 94-104.

Green, L. and P. J. Kolesar, ‘‘The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals,’’ *Management Sci.* 37 (1991), 84-97.

Green, L. and P. J. Kolesar, ‘‘On the Accuracy of the Simple Peak Hour Approximation for Markovian Queues,’’ Graduate School of Business, Columbia University, 1993.

Green, L., P. J. Kolesar and A. Svoronos, ‘‘Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems,’’ *Oper. Res.* 39 (1991), 502-511.

Halfin, S. and W. Whitt, ‘‘Heavy-Traffic Limits for Queues with Many Exponential Servers,’’ *Oper. Res.* 29 (1981) 567-587.

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

Holloran, T. J. and J. E. Byrn, ‘‘United Airlines Station Manpower Planning System,’’ *Interfaces,* 16 (1986), 39-50.

Iglehart, D. L., ‘‘Limit Diffusion Approximations for the Many-Server Queue and Repairman Problem,’’ *J. Appl. Prob.* 2 (1965) 429-441.

Jagerman, D. L. ‘‘Nonstationary Blocking in Telephone Traffic,’’ *Bell System Tech. J.* 54 (1975), 625-661.

Kolesar, P. J., ‘‘Comment to ‘Is the Fact That the Emperor Wears No Clothes a Subject Worthy of Publication?’’ *Interfaces* 16 (1986) 50-51.

Kolesar, P. J., K. L. Rider, T. B. Crabill and W. E. Walker, ‘‘A Queueing-Linear Programming Approach to Scheduling Police Patrol Cars,’’ *Oper. Res.* 23 (1975), 1045-1062.

Larson, R. C., *Urban Police Patrol Analysis,* MIT Press, Cambridge, MA, 1972.

Massey, W. A., G. A. Parker and W. Whitt, ‘‘Estimating the Parameters of a Nonhomogeneous Poisson Process with Linear Rate,’’ *Telecommunication Systems,* 4 (1995), to appear.

Massey, W. A. and Whitt, W., ‘‘Networks of Infinite-Server Queues with Nonstationary Poisson Input,’’ *Queueing Systems*, 13 (1993), 183-250.

Massey, W. A. and W. Whitt, ‘‘Unstable Asymptotics for Nonstationary Queues,’’ *Math. Oper. Res.* 19 (1994a), 267-291.

Massey, W. A. and W. Whitt, ‘‘An Analysis of the Modified Offered Load Approximation for the Nonstationary Erlang Loss Model,’’ *Ann. Appl. Prob.* 4 (1994b), 1145-1160.

Newell, G. F., *Approximate Stochastic Behavior of n-server Service Systems with Large n*, Springer, New York, 1973.

Quinn, P., B. Andrews and H. Parsons, ‘‘Allocating Telecommunications Resources at L. L. Bean,’’ *Interfaces* 21 (1991), 75-91.

Segal, M. ‘‘The Operator Scheduling Problem: A Network Flow Approach,’’ *Oper. Res.,* 22 (1974), 808-823.

Sze, D. Y., ‘‘A Queueing Model for Telephone Operator Staffing,’’ *Oper. Res.,* 32 (1984), 229-249.

Thompson, G. M., ‘‘Accounting for the Multi-Period Impact of Service When Determining Employee Requirements for Labor Scheduling,’’ *J. Opns. Mgmt.* 11 (1993), 269-288.

Thompson, G. M., ‘‘Setting Staffing Levels in Pure Service Environments When the True Mean Daily Customer Arrival Rate Is a Normal Random Variate,’’ University of Utah, Salt Lake City, UT, 1994.

Whitt, W., ‘‘Heavy-Traffic Approximations for Service Systems with Blocking,’’ *AT&T Bell Lab. Tech. J.*, 63 (1984), 689-708.

Whitt, W., ‘‘The Pointwise Stationary Approximation for $M_t/M_t/s$ Queues Is Asymptotically Correct as the Rates Increase,’’ *Management Sci.* 37 (1991), 307-314.

Whitt, W., ‘‘Understanding the Efficiency of Multi-Server Service Systems,’’ *Management Sci.* 38 (1992), 708-723.

Whitt, W., ‘‘Approximations for the GI/G/m Queue,’’ *Production and Opns. Mgmt.* 2 (1993), 114-161.

Wolff, R. W., *Stochastic Modeling and the Theory of Queues*, Prentice Hall, Englewood Cliffs, NJ, 1989.

offered load m	target delay probability α											
	0.20			0.10			0.05			0.01		
	E	P	N	E	P	N	E	P	N	E	P	N
1	3	3	3	3	3	3	4	4	4	5	5	5
2	4	4	4	5	5	5	6	6	5	7	7	6
5	8	8	8	9	9	9	10	10	10	12	12	11
10	14	14	14	16	16	15	17	16	16	19	19	18
20	26	25	25	27	27	27	29	29	28	32	32	31

Table 1. A comparison of three methods for determining the required number of servers in the stationary M/M/s model as a function of the offered load m and target delay probability α: the exact Erlang delay formula (E), the infinite-server Poisson distribution (P) and the normal approximation (N).

m	target probability α			
	0.1	0.01	0.001	0.0001
1	3	5	6	7
2	5	7	9	10
3	6	9	11	12
4	8	10	12	14
5	9 (9)	11 (11)	14 (13)	16 (14)
6	10	13	16	18
7	11	15	17	20
8	13	16	19	21
9	14	18	21	23
10	15 (15)	18 (18)	22 (21)	25 (23)
11	16	20	24	26
12	18	22	25	28
13	19	23	26	29
14	20	24	28	31
15	21 (21)	26 (25)	29 (28)	32 (30)
16	22	27	31	34
17	23	28	32	35
18	25	30	33	37
19	26	31	35	38
20	27 (27)	32 (31)	36 (35)	40 (38)

Table 5. Least number s of servers required to have $P(Q(m) \geq s) \leq \alpha$, where $Q(m)$ has a Poisson distribution with mean m, as a function of α and m. The normal approximation (4) is given in parentheses for some cases.

offered load m	$m + 0.5 + 2.576\sqrt{m}$	number of servers s	probability of delay in M/M/s		
			$P(Q \geq s+1)$	$P(Q \geq s)$	$P(Q \geq s-1)$
1	4.1	5	—	0.004	0.060
2	6.1	7	0.0005	0.005	0.037
3	7.9	8	0.0005	0.0045	0.025
4	9.7	10	0.0012	0.0090	0.035
5	11.3	12	0.0010	0.0030	0.017
6	12.8	13	0.0022	0.0097	0.030
8	15.8	16	0.0025	0.0090	0.025
10	18.6	19	0.0026	0.0080	0.020
15	25.5	26	0.0035	0.0067	0.013
20	33	32	0.0052	0.0091	0.015
30	44.6	45	0.0044	0.0070	0.011
40	56.8	57	0.0049	0.0074	0.011
50	68.7	69	0.0048	0.0069	0.010
60	80.4	81	0.0044	0.0063	0.0088
70	92.1	93	0.0039	0.0054	0.0075
80	103.6	104	0.0047	0.0063	0.0085
90	114.9	115	0.0053	0.0071	0.0093
M (big)	$M + 2.576\sqrt{M}$		0.0056	0.0056	0.0056

Table 2. The probability of delay in the stationary M/M/s model as a function of the offered load m and the number of servers s when s is chosen to satisfy (4) with $\alpha = 0.005$ ($z_\alpha = 2.576$).

normal tail percentiles $P(N(0,1) > z_\alpha) = \alpha$ z_α	infinite-server $P(Q_\infty \geq s) = \alpha$ α	s-server $P(Q_s \geq s)$ $P_D(\alpha)$ in (5)
0	0.5	1.000
0.2533	0.4	0.7177
0.5244	0.3	0.4865
0.8416	0.2	0.2937
1.282	0.1	0.1320
1.645	0.05	0.0619
2.362	0.01	0.0115
2.576	0.005	0.00561
3.090	0.001	0.00109
3.719	0.0001	0.000107

Table 3. The limiting probability of s or more busy servers in the infinite-server and s-server models, where $s = m + z_\alpha\sqrt{m}$ with the offered load m increasing to infinity.

arrival rate function	targets		range of delay probabilities	
	α	$P_D(\alpha)$		
$20+10\sin(t)$	.1	.13	0.09	0.13
$400+40\sin(.2t)$	.1	.13	0.12	0.13
$3+2\sin(t)$	.1	.13	0.06	0.12
$20+10\sin(t)$	.4	.71	0.52	0.58

Table 4. Four examples of the IS approximation applied to the $M_t/M/s_t$ model with sinusoidal arrival rate and mean service time 1.

time					
mean service times	hours	$m(t)$	$s(t)$	five levels	three levels
0.000	0.00	7.50	15	15	16
0.114	0.44	7.80	16	17	16
0.347	1.33	8.49	17	17	20
0.554	2.12	9.19	18	19	20
0.783	2.99	9.90	19	19	20
0.959	3.66	10.61	20	21	20
1.171	4.48	11.33	21	21	23
1.407	5.38	12.06	22	23	23
1.695	6.48	12.79	23	23	23
2.356	9.00	13.53 (max)	23	23	23
3.017	11.52	12.79	22	23	23
3.305	12.63	12.06	21	21	23
3.541	13.53	11.33	20	21	20
3.753	14.34	10.61	19	19	20
3.929	15.01	9.90	18	19	20
4.158	15.89	9.19	17	17	20
4.365	16.68	8.49	16	17	16
4.598	17.57	7.80	15	15	16
4.837	18.49	7.12	14	15	16
5.498	21.00	6.47 (min.)	14	15	16
6.159	23.53	7.12	15	15	16
6.283	24.00	7.50	15	15	16
time average		10.00	18.54	19.06	19.68

Table 5. Server staffing with $\alpha = 0.005$ ($z_\alpha = 2.576$) for the sinusoidal example in Section 5.

Figure 1. The PSA approximation: The server-staffing levels and delay probabilities as functions of time for the $M_t/M/s_t$ example with rapidly fluctuating sinusoidal arrival-rate function $\lambda(t) = 30 + 20 \sin(5t)$ using the pointwise stationary approximation (PSA) with a delay probability target 0.13. The offered load $\lambda(t)$ is plotted with the number of servers.

Figure 2. The SSA Approximation: The server-staffing levels and delay probabilities as functions of time for the $M_t/M/s_t$ example with rapidly fluctuating sinusoidal arrival-rate function $\lambda(t) = 30 + 20\sin(5t)$ using the simple stationary approximation (SSA) with the average arrival rate 30 and a delay probability target of 0.13. The offered load $\lambda(t)$ is plotted with the constant number of servers, 38.

Figure 3. The IS approximation: The server-staffing levels and delay probabilities as functions of time for the $M_t/M/s_t$ example with rapidly fluctuating sinusoidal arrival rate $\lambda(t) = 30 + 20\sin(5t)$, using the IS approximation based on $z_\alpha = 1.282$ ($\alpha = 0.1$ and $p_D(\alpha) = 0.13$).

Figure 4. The IS approximation: The tail probability $P(Q_s(t) \geq s(t) + 5)$ and expected number in queue (not in service) for the example in Figure 3.

Figure 5. IS approximation for the start-up example in Section 4: The server-staffing levels and associated (exact) delay probabilities as functions of time in the $M_t/M/s_t$ example starting out empty based on $\alpha = 0.05$. The number of servers is adjusted at each integer time point. The (constant) offered load is plotted with the number of servers.

Figure 6. IS approximation for the first sinusoidal example in Section 5: The server-staffing levels and associated (exact) delay probabilities as functions of time in the $M_t/M/s_t$ example with sinusoidal arrival rate $\lambda(t) = 20 + 10\sin t$ based on $\alpha = 0.1$, using (4). The offered load $\lambda(t)$ is plotted with the number of servers.

Figure 7. IS approximation for the first sinusoidal example in Section 5: The time-dependent tail probabilities $P(Q_s(t) \geq s(t) + 5)$ and the expected number in queue (not in service) for the sinusoidal example in Figure 6.

Figure 8. PSA for the first sinusoidal example in Section 5: The server-staffing levels and actual delay probabilities as functions of time using the pointwise-stationary approximation (PSA), i.e., using $\lambda(t)$ instead of $m(t)$ in (4), in the $M_t/M/s_t$ example with sinusoidal arrival rate $\lambda(t) = 20 + 10\sin t$ based on $\alpha = 0.1$.

Figure 9. A comparison of IS, shifted PSA and PSA: The actual delay probabilities in the $M_t/M/s_t$ example with sinusoidal arrival rate $\lambda(t) = 400 + 40 \sin(0.2t)$ based on $\alpha = 0.1$ when the number of servers is specified by (4). We use (12) for IS, (13) with $\lambda''(t) \approx 0$ for shifted PSA and (16) for PSA.

Figure 10. IS approximation: The tail probabilities $P(Q(t) \geq s(t) + 5)$ and expected number in queue (not in service) for the IS approximation for the example in Figure 9.

Figure 11. IS approximation for the small sinusoidal example: The server staffing levels and the associated (exact) tail probabilities as functions of time in the $M_t/M/s_t$ example with sinusoidal arrival rate $\lambda(t) = 3 + 2 \sin t$ based on $\alpha = 0.1$ using (4). As before, the offered load is plotted with the number of servers.

Figure 12. The IS approximation: The tail probabilities $P(Q_s(t) \geq s(t) + 5)$ and expected number in queue (not in service) for the small example in Figure 11 using IS.

Figure 13. The IS approximation with a large target probability: The server staffing levels and the associated delay probabilities as functions of time in the $M_t/M/s_t$ example with sinusoidal arrival rate $\lambda(t) = 20 + 10 \sin t$ based on $\alpha = 0.4$ using (4).

Figure 14. The IS approximation: The tail probabilities $P(Q_s(t) \geq s(t) + 5)$ and mean number in queue (not in service) for the example in Figure 13.

PSA, using $\lambda(t)$ instead of $m(t)$ in (2)

the infinite-server approximation (2)

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

$\lambda(t)$ $s(t)$

IS

PSA

shifted PSA

Appendix to ‘‘Server Staffing to Meet Time-Varying Demand’’ by O. B. Jennings, A. Mandelbaum, W. A. Massey and W. Whitt

This appendix contains additional displays of time-dependent performance measures for $M_t/M/s_t$ queues. These new displays are similar to the 12 Figures in the paper. Some of these new displays provide extra detail for the examples in the paper. Other displays are different examples.