Predicting Queueing Delays

Ward Whitt
Room A117, AT&T Labs, Shannon Laboratory, 180 Park Avenue, Florham Park, New Jersey 07932-0971
wow@research.att.com

This paper investigates the possibility of predicting each customer's waiting time in queue before starting service in a multiserver service system with the first-come first-served service discipline, such as a telephone call center. A predicted waiting-time distribution or an appropriate summary statistic such as the mean or the 90th percentile may be communicated to the customer upon arrival and possibly thereafter in order to improve customer satisfaction. The predicted waiting-time distribution may also be used by the service provider to better manage the service system, e.g., to help decide when to add additional service agents. The possibility of making reliable predictions is enhanced by exploiting information about system state, including the number of customers in the system ahead of the current customer. Additional information beyond the number of customers in the system may be obtained by classifying customers and the service agents to which they are assigned. For nonexponential service times, the elapsed service times of customers in service can often be used to advantage to compute conditional-remaining-service-time distributions. Approximations are proposed to convert the distributions of remaining service times into the distribution of the desired customer waiting time. The analysis reveals the advantage from exploiting additional information.

(Service Systems; Telephone Call Centers; Predicting Delays; Communicating Anticipated Delays; Predicting Response Times)

1. Introduction

It is common practice in service systems to have customers who cannot be served immediately upon arrival wait in queue until system resources become available to the customer. Traditionally, customers have not been given estimates of their required waiting times, i.e., the time until they can begin to receive service. When the waiting times are sufficiently short, there is usually little need for such information, but if waiting times can be long (where what is "long" depends on the context), then prediction can be important.

The queue may physically contain the customers, as in the lines in a bank or a supermarket, or it may not, as in calls to a telephone call center or requests for emergency (police, fire or medical) service. When the queue physically contains the customers, the customers are often able to directly estimate the waiting time, so that customers may not gain much from additional delay predictions. However, even if the queue physically contains the customers, it may be difficult for customers to do the estimation. For example, in amusement parks, customers are often unable to see the full line, and they may not have the experience to know how fast it moves. Thus, amusement parks have increased customer satisfaction by having signs indicating the expected waiting time from that point.

Predicting delays for customers is especially important when customers do not have direct access to system state information. For a long time in emergency services, telephone dispatchers have tried to indicate how long it will take before assistance will arrive. Even with less critical services provided via telephone, it has been recognized that customers be-

0025-1909/99/4506/0870$05.00

come dissatisfied with the service provider when they are forced to wait "on hold" for long indeterminate periods. Thus, it is now becoming more common in telephone call centers to make announcements to customers containing delay predictions. Making delay predictions is one of several ways to improve customer waiting experience; see Hui and Tse (1996), Katz et al. (1991), Taylor (1994) and references therein.

The purpose of this paper is to study delay predictions. A first question is whether or not uncertainty about delays can be sufficiently small that point-estimate delay predictions such as the mean can be of value. We will show that the uncertainty about delays tends to be large if no state information is used, but that the uncertainty about delays typically can be dramatically reduced if state information is used, so that point-estimate prediction can indeed be of value.

Given that delay prediction can help, the second question is how to actually make these predictions. In this paper we propose several methods for making delay predictions. We propose more than one method, because we have found that there is not a single best method for all circumstances. In proposing these prediction methods, we are primarily concerned with improving customer satisfaction, but delay predictions can also be used for other purposes. For instance, a service provider might provide additional service at some other facility after the first service is complete. The predicted delay at the first facility might enable the service provider to better plan for the subsequent service. The service provider also might use the delay predictions to adjust its available service capacity, e.g., by adding agents when large delays are predicted.

To put our proposed prediction procedures in perspective, we first discuss current practice. Some service providers identify a customer's place in queue. However, position information may not enable the customer to determine how long the customer will have to wait before beginning service. The customer may not be able to determine how many agents are fielding service requests or the rate at which the agents are completing service requests. Other service providers may generate an estimate of the rate at which its agents complete service requests from its customers. A system having s agents each of whom, on average, complete service requests in r minutes, may predict that a customer placed in queue at the kth position will be able to begin service in kr/s minutes and complete service in $(1 + (k/s))r$ minutes. However, prediction based on such long-run averages may be subject to gross inaccuracies in specific instances. If long-run averages are not met in specific instances, customers may wait for a much longer time than is predicted. Such customers, too, may become dissatisfied with the service provider.

We assume that customers are served one at a time, each by one of several servers. A customer's service time may depend on the type of customer and the type of server, but otherwise (conditional on such information) the successive service times are assumed to be mutually independent. In that setting, we aim to provide more accurate predictions of required waiting times. Our main idea is to exploit additional information, which often is readily available with modern computer information systems. We propose classifying each waiting customer and each customer being served based upon known attributes of the customers and possibly upon the agents that are providing the service. Based upon these classifications, we estimate the cumulative distribution function (cdf) of each customer's remaining service time. We then convert these estimated remaining-service-time cdf's into an estimated waiting-time cdf.

A second main idea is to estimate the full waiting-time cdf. Recognizing that the waiting time is usually uncertain, we do more than provide a single-number prediction, such as the expected value. However, we are primarily interested in the waiting-time cdf to understand the reliability of point estimates such as the mean. In predicting the entire waiting-time probability distribution, we recognize that there is a separate question about what to say to customers. We believe that often a good case can be made for full and accurate disclosure, but that is a separate question which we do not address; see Hui and Tse (1996) for further discussion.

Here is how the rest of this paper is organized. We start in §2 by discussing delay predictions in queues with the first-come first-served discipline and exponentially distributed service times. We indicate how

the predictions can be enhanced to account for reneging, drawing upon recent work in the companion paper, Whitt (1999), which investigates how system performance in $M/M/s/r$ models with balking and reneging is affected by communicating system-state information. We also show how the value of information can be quantified.

In §3 we consider the case in which customers can be classified into two types, with different exponentially distributed service times. We develop efficient recursive algorithms for calculating the mean, variance, and the Laplace transform of the waiting time. Values of the full cumulative distribution function of the waiting time can be computed by numerically inverting the Laplace transform, e.g., by the algorithm in Abate and Whitt (1992, 1995). In §4 we develop stochastic bounds and approximations for the waiting-time cdf when the customers in the system have exponential remaining service times with different means.

In §5 we specify initial rough approximations for the mean, variance, and Laplace transform of the waiting time, exploiting the number of customers in the system ahead of the customer in question, when the service-time distribution is nonexponential. These approximations serve as reference points for more refined approximations. In §6 we develop a prediction method for general (nonexponential) service-time distributions, exploiting the elapsed service times of customers in service. We compute the conditional remaining service-time cdf's and use them to generate an approximation for the waiting-time cdf.

In §7 we consider the special case in which all remaining service times are known when the prediction is to be made. It is significant that the general approximation in §6 coincides exactly with a known deterministic recursion in this case. Thus we see that the approximation in §6 will be effective when there is little uncertainty about the remaining service times. We also suggest using the deterministic recursion applied to the mean remaining service times as a crude approximation requiring less information. As a refinement, we suggest a fast simulation performed in real time, randomly generating each remaining service time.

In §8 we consider special prediction methods for customers near the front of the queue when there are many servers. In §9 we consider simulation experiments to validate the approximations. We make concluding remarks in §10.

We close this introduction by mentioning previous work on prediction by Stanford et al. (1983) and Woodside et al. (1984). They focus on optimal predictors of *future* waiting times and queue lengths given past observations of these processes in a *stationary* setting. In contrast, we emphasize prediction of the waiting time of *current* customers based on extra information in what may be a *nonstationary* setting.

2. Prediction with Exponential Service Times

Throughout this paper we make three important assumptions: First, there is a fixed number s of servers. Second, the service discipline is first-come first-served (FCFS). Third, the service times at each server are mutually independent (possibly depending on the type of customer and/or the type of server) and independent of the arrival process. In this section, we assume in addition that all service times have a common exponential distribution with mean μ^{-1}.

If we are given no state information, then the natural prediction is the cdf of the steady-state waiting time W_∞, which requires that the queueing model be fully specified and that the arrival process be stationary with the property that the steady-state waiting-time cdf can be calculated. For example, if there is unlimited waiting space and if the arrival process is a renewal process, then we have the $GI/M/s$ model. Then the steady-state waiting-time cdf has the form $P(W_\infty \le t) = 1 - pe^{-bt}$, where the two parameters p and b depend on the interarrival-time cdf, s and μ; e.g., see Gross and Harris (1985). The most natural special case is a Poisson arrival process, yielding the $M/M/s$ model.

If we learn only that a customer must wait before beginning service, it is natural to use the conditional steady-state waiting-time cdf given that the wait is strictly positive, i.e., $P(W_\infty \le t \mid W_\infty > 0)$. For the $GI/M/s$ model, this conditional cdf is exponential. In this setting it is evident that a single point estimate

such as the mean is not very reliable. In contrast, we will show that additional information tends to make the conditional waiting-time cdf concentrate more about its mean, so that a single point estimate such as the mean becomes much more reliable.

Throughout this paper we assume that we learn at least the number of customers already in the system when each arrival occurs. We may also learn more. For example, we may learn the elapsed service times for each customer in service. However, with exponential service times, the elapsed service times provide no benefit, because by the lack-of-memory property of the exponential distribution, the remaining service time for each customer in service has an exponential distribution with mean μ^{-1}, independent of the elapsed service time.

For a specified model, such as $GI/M/s$, the additional information clearly helps, as we will discuss below. Perhaps, even more important, however, is the fact that prediction depends much less upon the model when we learn additional state information. When we learn the current state information, we do not need to know anything about the arrival process. The arrival process can be *arbitrary*, which we denote by A, even nonstationary. The prediction also does not depend upon blocking and balking (customers leaving immediately upon arrival). Thus the prediction applies to the $A/M/s/r$ model, where r is the number of waiting spaces. Moreover, the prediction can be done any time after arrival, not just upon arrival. In addition, the relevant information is only s, μ^{-1} and the number of customers in the system.

For the $A/M/s/r$ model, whenever all servers are busy, the time until the next service completion is exponentially distributed with mean $1/s\mu$, independent of the previous history. Hence, the waiting time before starting service for an arrival finding $s + k$ customers in the system is the sum of $k + 1$ i.i.d. exponential random variables each with mean $1/s\mu$, which has an Erlang (gamma) distribution. In this context, we can immediately see the advantage of the additional state information, because for $k > 0$ the Erlang distribution is much more concentrated about its mean than the exponential steady-state distribution (conditional on all servers being busy) in the $GI/M/s$ model. Conditioning on the state information, the mean and standard deviation of the steady-state waiting time before starting service, W, are

$$EW = \frac{k+1}{s\mu} \quad \text{and} \quad SD(W) = \frac{\sqrt{k+1}}{s\mu}. \tag{2.1}$$

We can also calculate the full cumulative distribution function (cdf) of the waiting time $P(W \le t)$, $t \ge 0$. The Erlang cdf is available in closed form, e.g., see p. 11 of Feller (1971). It can also be computed easily by numerical transform inversion, as we indicate for a more general case below. Except for very small k, a normal approximation works well, which depends only on the mean and standard deviation in (2.1). The cdf and the standard deviation also reveal the potential error in any single point estimate such as the mean. It is natural to describe the reliability of the mean as a point estimate by the ratio $SD(W)/EW$. Note that with i.i.d. exponential service times this ratio has the remarkably simple form

$$SD(W)/EW = 1/\sqrt{k+1}, \tag{2.2}$$

independent of μ and s. Formula (2.2) is a useful rough guide more generally. For example, when $k + 1 = 25$ (100), we anticipate about 20% (10%) error due to uncertainty.

It is reasonable to normalize by the mean service time, so that waiting times are viewed in relation to mean service times. This is equivalent to setting $\mu = 1$ in (2.1). The formulas in (2.1) show the advantage of large scale. When s is big, either EW and $SD(W)$ are both small (when k is small) or the ratio $SD(W)/EW$ is small (when k is large).

EXAMPLE 2.1. Thinking of a large telephone call center, natural values for s and k might be $s = 400$ and $k = 80$. By (2.1), with a mean service time of $\mu^{-1} = 5$ minutes, the mean wait before beginning service is about 1 minute. Using the normal approximation, a 95% confidence interval for the mean is approximately $EW \pm 2SD(W)$ or (47 seconds, 74 seconds). The mean is a reasonably reliable prediction because the upper and lower limits of the confidence interval differ from it by only 20%. The customer might be told that his expected wait is 1 minute and that the chance of the

wait exceeding 75 seconds is only 0.025. The 90th percentile is 70 seconds. To keep things simple while trying not to disappoint customers, the customer might be given the single estimate of 70 seconds.

Note that there is possible room for improvement even if the $A/M/s/r$ model provides an excellent fit, because we could obtain still more information. For example, immediately upon arrival, triage might be performed to determine the nature of a customer's service requirements; see p. 311 of Hall (1991). Triage has traditionally been used to assign priority to customers needing especially prompt service, as in emergency medicine, but it can also be used to improve prediction. In the most favorable circumstance, we might be able to learn the full service requirements of each customer upon arrival (which could still be consistent with the overall exponential distribution). If we do not use this information, then (2.1) is still correct. However, if we do use this information, then we can compute the waiting time exactly (with no variance) at every time. We have observed that when there are $s + k$ customers in the system, the waiting time has an Erlang distribution with mean and variance in (2.1). If we actually learn the service times, then we know the value of the Erlang-distributed waiting time at each instant.

It is naturally of interest to *quantify the value of information.* One way to do this is to compare the variance with and without the information. To do so, we consider the case in which there is a well-defined steady-state waiting time W_∞. Without any information, we have the variance $\mathrm{Var}(W_\infty)$. When we receive information, we want to consider not one instance, but its long-run impact in steady state. Thus let I_∞ represent the steady-state form of the information. The influence of this information can be quantified by comparing the expected conditional variance $E[\mathrm{Var}(W_\infty \mid I_\infty)]$ to the unconditional variance $\mathrm{Var}(W_\infty)$. Assuming that I_∞ and W_∞ are jointly distributed,

$$\mathrm{Var}(W_\infty \mid I_\infty) = E[(W_\infty - E(W_\infty \mid I_\infty))^2 \mid I_\infty]. \quad (2.3)$$

Since

$$\mathrm{Var}(W_\infty) = E[\mathrm{Var}(W_\infty \mid I_\infty)] + \mathrm{Var}(E(W_\infty \mid I_\infty)), \quad (2.4)$$

we have

$$\mathrm{Var}(W_\infty) \geq E[\mathrm{Var}(W_\infty \mid I_\infty)]; \quad (2.5)$$

the variance can only decrease with conditioning.

EXAMPLE 2.2. Consider the $M/M/1$ queue with arrival rate λ and traffic intensity $\rho = \lambda/\mu < 1$. The steady-state waiting time W_∞ has the distribution

$$P(W_\infty > t) = \rho e^{-(\mu-\lambda)t}, \quad t > 0, \quad (2.6)$$

so that

$$\mathrm{Var}(W_\infty) = \frac{\rho(2-\rho)}{\mu^2(1-\rho)^2}. \quad (2.7)$$

In contrast, assuming that I_∞ is the number of customers in the system, so that $P(I_\infty = k) = (1-\rho)\rho^k$, $k \geq 0$,

$$E[\mathrm{Var}(W_\infty \mid I_\infty)] = \sum_{k=0}^{\infty} (1-\rho)\rho^k \frac{k}{\mu^2} = \frac{\rho}{\mu^2(1-\rho)} \quad (2.8)$$

and

$$\frac{E[\mathrm{Var}(W_\infty \mid I_\infty)]}{\mathrm{Var}(W_\infty)} = \frac{1-\rho}{2-\rho}. \quad (2.9)$$

From (2.9), we see that the variance ratio is decreasing in ρ, starting at $\frac{1}{2}$ at $\rho = 0$ and is asymptotically $(1-\rho)$ as $\rho \to 1$. Hence conditioning provides a big variance reduction in heavy traffic. Also note that in those rare instances in which the number k of customers in the system is large, and delays will tend to be large, the prediction is reliable and differs dramatically from the steady-state mean. Thus the value of delay prediction actually may be much greater than predicted by (2.9). (Formula (2.9) describes the *average benefit per customer* for all customers. We would see a greater benefit if we considered the average benefit per customer experiencing a large delay.)

Another important consideration is reneging (customer abandonment after waiting). Some customers in line may actually renege before receiving service, making the waiting time less than it otherwise would be. Fortunately, it is also possible to predict waiting times when there is significant reneging, provided that the service times are i.i.d. and exponentially distributed. First, suppose that each waiting customer in position j of the queue tends to renege at a constant rate δ_j'. If $\delta_j' = \alpha$, then this is tantamount to each

customer being willing to wait a random time that is exponentially distributed with mean α^{-1}. We can use a pure death process (birth-and-death process with no births) to represent the evolution of the system when an arrival finds $s + k$ customers in the system; e.g., see Whitt (1999). When a customer finds $s + k$ customers in the system, the delay can again be represented as the sum of $k + 1$ independent exponential random variables, but now they are not identically distributed. The total reneging rate when there are k customers in queue is

$$\delta_k = \sum_{j=1}^{k} \delta_j'. \tag{2.10}$$

The mean and standard deviation of the waiting time for a new arrival (assuming that the arrival does not himself renege) become

$$EW = \sum_{j=0}^{k} \frac{1}{s\mu + \delta_j} \quad \text{and}$$

$$SD(W) = \left[\sum_{j=0}^{k} \frac{1}{(s\mu + \delta_j)^2} \right]^{1/2}, \tag{2.11}$$

for δ_j in (2.10), $j \geq 1$, and $\delta_0 = 0$.

The sum of $k + 1$ independent exponential random variables with different means has a rather complicated cdf, being a k-fold convolution of the component exponential cdf's, but the cdf of the sum is remarkably (perhaps surprisingly) easy to calculate by numerical transform inversion, as shown in §12 of Abate and Whitt (1982). The Laplace transform of the waiting time in state $s + k$ is

$$\hat{w}(z) \equiv Ee^{-zW} \equiv \int_0^\infty e^{-zt} dP(W \leq t)$$

$$= \prod_{j=0}^{k} \left(\frac{s\mu + \delta_j}{s\mu + \delta_j + z} \right), \tag{2.12}$$

so that the complementary cdf $P(W > t)$ can be computed by numerically inverting its Laplace transform

$$\hat{W}^c(z) \equiv \int_0^\infty e^{-zt} P(W > t) dt = \frac{1 - \hat{w}(z)}{z}. \tag{2.13}$$

The numerical inversion algorithm can compute several cdf values $P(W \leq t)$ in less than a second, so that it can be used on a real-time basis to predict the cdf.

EXAMPLE 2.3. A typical simplification of the reneging model is to assume that $\delta_j = j\alpha$, $j \geq 1$. If furthermore we assume that $\alpha = \mu$, then the $M/M/s$ model with reneging behaves like the $M/M/\infty$ model, and (for $s > 1$)

$$EW = \mu^{-1} \sum_{j=s}^{j=s+k} j^{-1}$$

$$\approx \mu^{-1}[\log(s + k) - \log(s - 1)]; \tag{2.14}$$

see 6.3.1 of Abramowitz and Stegun (1972). From (2.1) and (2.14), we see that it can be important to account for reneging when it is present to a significant degree.

3. Identifiable Customer Classes

We now start to consider improved predictions that can be made with additional information beyond the number of customers in the system. In this section we suppose that there are two classes of customers. Let customers of class 1 have exponential service times with mean μ_1^{-1}, and let customers of class 2 have exponential service times with mean μ_2^{-1}. For example, local telephone calls might be classified into data (e.g., Internet) calls and voice calls, with data calls tending to have longer holding times. Such a classification can be based on the called number. For another example, calls to airline reservation centers might be classified according to whether they are from travel agents or private individuals. Such a classification can be based on either calling or called numbers. For yet another example, Internet usage might be classified into sessions retrieving email and web browsing, with the classification based on customer usage determined from real-time traffic data.

In this section, let each successive customer be class 1 with probability p, independent of all other events. Thus the service times of successive customers are

i.i.d. with hyperexponential (H_2, mixture of two exponentials) distributions. The overall model is thus $A/H_2/s/r$ instead of $A/M/s/r$. As at the end of §2, we let customers in queue renege; the total reneging rate with k customers in queue is δ_k.

In this setting we suppose that the customer class identity is learned when the customer starts service, but not before. Thus, when there is a queue, we know the class identity of the s customers in service. Each customer in queue is class 1 with probability p, independent of other events. In this setting, just as in §2, we can calculate the mean, standard deviation and Laplace transform of the waiting time for a new arrival (or any customer already in queue), but the recursive calculation here is substantially more complicated. Before developing the recursive algorithm, we consider a revealing example.

EXAMPLE 3.1. To understand the advantages of the customer-class information and to obtain a simple approximation for the recursions that may sometimes be appropriate, it is useful to consider the limiting case in which class 2 has service times of length 0. We obtain simplicity because we can conclude that all customers in service at an arbitrary time are necessarily of class 1. For additional simplicity, suppose that there is no reneging. If there are $k + s$ customers in the system ahead of a customer of interest, that customer must wait for one plus a binomially distributed random number of exponential random variables each with mean $1/s\mu_1$. Hence the mean is

$$EW = \frac{(1 + kp)}{s\mu_1}, \tag{3.1}$$

the Laplace transform is

$$\hat{w}(z) \equiv Ee^{-zW} = \left(\frac{s\mu_1}{z + s\mu_1}\right) \sum_{j=0}^{k} \binom{k}{j} \times p^j(1 - p)^{k-j}\left(\frac{s\mu_1}{z + s\mu_1}\right)^j, \tag{3.2}$$

and the standard deviation is

$$SD(W) = \frac{(1 + 2kp - kp^2)^{1/2}}{s\mu_1}. \tag{3.3}$$

If $SD(W)/E(W) = (1 + 2kp - kp^2)^{1/2}/(1 + kp)$ is small, which occurs if kp is large, then reliable prediction is possible. From formulas (3.1)–(3.3), we can also see that there can be great value in learning the class identities of the customers in queue too. Then we would know the actual value of the binomial random variable. For instance, if the actual number of class-1 customers in queue is $j = kp$, then the mean is unchanged, but the standard deviation is reduced to $(1 + kp)^{1/2}/s\mu_1$. And when j is not near kp, the conditional means are not close.

To develop the recursive algorithm for the general case (with $\mu_2^{-1} > 0$), let $W(k, j)$ denote the waiting time of an arrival that finds $s + k$ customers in the system with j class-1 customers in service (assuming that this selected customer does not renege). We now are faced with a two-dimensional generalization of the pure-death process considered in §2. We calculate the mean recursively, considering what happens at each successive departure (ignoring future arrivals). Since we have assumed that (are conditioning upon the fact that) the customer of interest does not renege, we let the reneging rate be δ_k when the customer of interest is number $k + 1$ in queue. The time until the first departure has an exponential distribution with mean $(j\mu_1 + (s - j)\mu_2 + \delta_k)^{-1}$. Let $T(k, j)$ denote the remaining waiting time, not counting this time until the first departure. At the time of the departure, the number in queue ahead of the new arrival decreases from k to $k - 1$. With probability $j\mu_1/(j\mu_1 + (s - j)\mu_2 + \delta_k)$ the departing customer is due to a class-1 customer completing service. With probability $\delta_k/(j\mu_1 + (s - j)\mu_2 + \delta_k)$, the departure is a customer in queue reneging. With probability p, a customer entering service is class 1. Thus, we obtain the following recursion for the mean. For $k \geq 0$,

$$E[W(k, j)] = \frac{1}{j\mu_1 + (s - j)\mu_2 + \delta_k} + E[T(k, j)], \tag{3.4}$$

where, for $k \geq 1$,

$$E[T(k, j)] = \frac{p(s - j)\mu_2 E[W(k - 1, j + 1)]}{j\mu_1 + (s - j)\mu_2 + \delta_k} + \frac{(1 - p)j\mu_1 E[W(k - 1, j - 1)]}{j\mu_1 + (s - j)\mu_2 + \delta_k}$$

$$+\left[\frac{pj\mu_1 + (1-p)(s-j)\mu_2 + \delta_k}{j\mu_1 + (s-j)\mu_2 + \delta_k}\right] \times E[W(k-1, j)] \quad (3.5)$$

and

$$E[T(0, j)] = 0. \quad (3.6)$$

A related recurrence can be developed for the variance. Noting that the waiting time is the sum of the exponential variable with mean $(j\mu_1 + (s-j)\mu_2 + \delta_k)^{-1}$ and the independent variable $T(k, j)$, we see that

$$\text{Var}[W(k, j)] = \frac{1}{[j\mu_1 + (s-j)\mu_2 + \delta_k]^2} + \text{Var}(T(k, j)), \quad (3.7)$$

where

$$\text{Var}(T(k, j)) = E[T(k, j)^2] - (E[T(k, j)])^2 \quad (3.8)$$

with $E[T(k, j)]$ in (3.5),

$$E[T(k, j)^2] = \frac{p(s-j)\mu_2 E[W(k-1, j+1)^2]}{j\mu_1 + (s-j)\mu_2 + \delta_k} + \frac{(1-p)j\mu_1 E[W(k-1, j-1)^2]}{j\mu_1 + (s-j)\mu_2 + \delta_k} + \left(\frac{pj\mu_1 + (1-p)(s-j)\mu_2 + \delta_k}{j\mu_1 + (s-j)\mu_2 + \delta_k}\right) \times E[W(k-1, j)^2], \quad (3.9)$$

$$E[W(k-1, j)^2] = \text{Var}(W(k-1, j)) + (E[W(k-1, j)])^2 = \frac{1}{[j\mu_1 + (s-j)\mu_2 + \delta_{k-1}]^2} + \text{Var}(T(k-1, j) + (E[W(k-1, j)])^2 \quad (3.10)$$

and $E[T(0, j)^2] = 0$.

Let $\hat{w}(z; k, j)$ be the Laplace transform of $W(k, j)$, i.e.,

$$\hat{w}(z; k, j) \equiv Ee^{-zW(k,j)}. \quad (3.11)$$

Paralleling the recursions for the mean and variance, we obtain a recursion for the Laplace transform, namely,

$$\hat{w}(z; k, j) = \left(\frac{j\mu_1 + (s-j)\mu_2 + \delta_k}{j\mu_1 + (s-j)\mu_2 + \delta_k + z}\right) \cdot \left(\frac{p(s-j)\mu_2 \hat{w}(z; k-1, j+1)}{j\mu_1 + (s-j)\mu_2 + \delta_k} + \frac{(1-p)j\mu_1 \hat{w}(z; k-1, j-1)}{j\mu_1 + (s-j)\mu_2 + \delta_k} + \left(\frac{pj\mu_1 + (1-p)(s-j)\mu_2 + \delta_k}{j\mu_1 + (s-j)\mu_2 + \delta_k}\right) \times \hat{w}(z; k-1, j)\right) \quad (3.12)$$

and

$$\hat{w}(z; 0, j) = \frac{j\mu_1 + (s-j)\mu_2}{j\mu_1 + (s-j)\mu_2 + z}. \quad (3.13)$$

We thus can calculate values of the complementary cdf $P(W(k, j) > t)$ by numerically inverting its Laplace transform

$$\hat{W}^c(k, j) \equiv \int_0^\infty e^{-zt} P(W(k, j) > t) dt = \frac{1 - \hat{w}(z; k, j)}{z}. \quad (3.14)$$

By calculating $E[W(k, j)]$, $SD(W(k, j))$ and $P(W(k, j) > t)$ for various j, we can determine the impact of the state information. We can describe the extreme cases directly when there is no queue: Note that $W(0, s)(W(0, 0))$ has an exponential distribution with mean $(s\mu_1)^{-1}((s\mu_2)^{-1})$. When $s \gg k$ and δ_k is small, $W(k, s)(W(k, 0))$ is approximately distributed as the sum of $k + 1$ i.i.d. exponential variables each with mean $(s\mu_1)^{-1}((s\mu_2)^{-1})$. Hence, we see that the information can greatly improve predictions when μ_1^{-1} and μ_2^{-1} differ significantly.

REMARKS. Several variations and extensions can be treated by essentially the same reasoning. First, there could be more than two classes. Second, customer class identity could be determined upon arrival

instead of upon starting service. A similar recursion holds for this case, letting the state indicate the customer class of each successive customer in queue (paying attention to the order in queue) as well as the customers in service. Finally, the two classes considered could be determined by the servers instead of the customers; i.e., the customers could be homogeneous but there could be two classes of servers, one working at rate μ_1 and the other at rate μ_2. Then the number of class-1 customers in service would not change after each successive service completion, and the analysis closely parallels §2.

4. Bounds When Classifying Individual Customers

We now suppose that additional classification is possible, so that there are more than two customer classes. Unlike §3, we assume that customers are classified upon arrival. We still assume that all remaining service times are mutually independent with exponential distributions, but now the means may all be different. We assume that the vector of $s + k$ individual service rates $(\mu_1, \mu_2, \ldots, \mu_{s+k})$ is known, with the s customers in service listed first followed in order by the customers in queue. We also suppose that each customer in queue has his own reneging rate. We assume that the vector of reneging rates $(\alpha_1, \ldots, \alpha_k)$ is known as well.

In this section, we provide stochastic upper and lower bounds on the waiting-time distribution, i.e., for the waiting-time cdf, which yield upper and lower bounds on the mean. When the two bounds are close, we are assured that we have a good approximation for the mean.

To develop the bounds, note that each successive departure is either a service completion by a customer in service or an abandonment (reneging) by a customer in queue. Given that we keep track of previous departure triggering events, the interdeparture-time distribution is exponential with a known rate (the sum of the relevant rates at that time). An upper (lower) bound is obtained by assuming that the remaining service and reneging rates are as small (large) as possible. Thus the bounding waiting-time cdf's are the distributions of sums of $k + 1$ independent exponential random variables, but with different means.

Let $\{\mu_{n,1}, \ldots, \mu_{n,s}\}$ denote the bounding set of s service rates for the nth interdeparture time, with $\{\mu_{1,1}, \ldots, \mu_{1,s}\} \equiv \{\mu_1, \ldots, \mu_s\}$. Then the $(n + 1)$st upper (lower) bound set $\{\mu_{n+1,1}, \ldots, \mu_{n+1,s}\}$ contains the s smallest (largest) elements from the set $\{\mu_1, \ldots, \mu_{s+n}\}$, $1 \le n \le k$. Let $\{\alpha_{n,1}, \ldots, \alpha_{n,k+1-n}\}$ denote the bounding set of $k + 1 - n$ reneging rates for the nth interdeparture time, with $\{\alpha_{1,1}, \ldots, \alpha_{1,k}\} = \{\alpha_1, \ldots, \alpha_k\}$. Then the nth upper (lower) bound set $\{\alpha_{n,1}, \ldots, \alpha_{n,k+1-n}\}$ contains the $k + 1 - n$ smallest (largest) elements from the set $\{\alpha_1, \ldots, \alpha_k\}$. Then the bounding waiting times W_b have means

$$EW_b \approx \sum_{n=1}^{k+1} \left[\sum_{i=1}^{k+1-n} \alpha_{n,i} + \sum_{i=1}^{s} \mu_{n,i} \right]^{-1} . \qquad (4.1)$$

Since the means of each exponential variable are bounded, we have stochastic bounds on the entire waiting-time cdf, i.e.,

$$P(W_b^l > t) \ge P(W > t) \le P(W_b^u > t) \quad \text{for all } t, \qquad (4.2)$$

where W_b^u and W_b^l are the upper and lower bounds.

It may also be desirable to have an approximation in between the bounds. A simple one for the mean is the average of the two bounds. We obtain an associated intermediate approximation for the entire waiting-time distribution by again using the sum of $k + 1$ independent exponentials, but with the means being the average of the upper and lower bound means. Alternatively, we could use the average of the rates.

EXAMPLE 4.1. To illustrate, let $s = 100$, $k = 20$, and suppose that there is no reneging. Suppose that at the prediction time there are 100 customers in service, where 30 have rate $\mu_1 = 0.04$, 40 have rate $\mu_2 = 0.06$, and 30 have rate $\mu_3 = 0.08$. Let the 20 customers in queue all have rate $\mu_2 = 0.06$. The successive departure rates for the upper bound are then 6.00, 5.98, ..., 5.60, while the successive departure rates for the lower bound are 6.00, 6.02, ..., 6.40. The upper and lower bounding mean waiting times are 3.62 and 3.39. Their average, 3.51, differs from the bounds by only about 3%. The successive rates for the intermediate approximation with average rates are 6.00, ..., 6.00, so that

this intermediate approximating mean is 3.50. The standard deviations for the lower bound, intermediate approximation and upper bound are 0.74, 0.76, and 0.79. Hence, the error in the bounds for the mean is relatively small compared to the overall uncertainty. The lower bound, intermediate approximation and upper bound for the 90th percentile of the waiting-time cdf are 4.34, 4.47, and 4.63. In this setting, the customer might be told that his estimated waiting time is 3.5, but there is some uncertainty; there is only a 10% chance it will exceed 4.5. Note that customer means (and standard deviations) of the individual service times are 25, 16.7, or 12.5, so that the prediction is relatively reliable in relation to the service-time mean and standard deviation.

EXAMPLE 4.2. We now add reneging to the previous example. Suppose that 10 of the 20 customers in queue have reneging rates 0.01, while 10 others have reneging rates 0.02. The 21 successive lower total reneging rates for the upper bound are 0.30, 0.28, 0.26, . . . , 0.12, 0.10, 0.09, . . . , 0.01, 0.00, while the 21 successive higher total reneging rates for the lower bound are 0.30, 0.29, . . . , 0.21, 0.20, 0.18, . . . , 0.02, 0.00. The 21 successive lower total interdeparture-time rates for the upper bound are 6.30, 6.26, 6.22, . . . , 5.94, 5.90, 5.87, . . . , 5.60, while the 21 successive higher total interdeparture-time rates for the lower bound are 6.30, 6.31, 6.32, . . . , 6.40, 6.40, . . . , 6.40. Finally, the upper and lower bounds for the means are 3.55 and 3.14. The average 3.35 differs from the bounds only by about 6%. The average with the reneging rates considered here is about 5% less than the average in Example 4.1. The standard deviations for these two bounds are 0.775 and 0.719.

5. The Departure-Renewal-Process Approximation

We now start to develop waiting-time predictions without making exponential-distribution assumptions for the service times. We do not consider reneging here. Note that the formulas in §§2–4 all yield approximations for general service-time distributions if we just act as if the service-time distributions were exponential with the given means. Such approximations are clearly reasonable if the service-time distributions are not very different from exponential, and may be useful more generally.

As indicated in the introduction, a natural candidate approximation for the mean waiting time of an arrival finding $s + k$ customers in an $A/GI/s/r$ system with i.i.d. service times having a general distribution with mean μ^{-1} is

$$EW \approx \frac{k+1}{s\mu}, \tag{5.1}$$

just as in (2.1). We regard (5.1) as the standard approximation when the number in system is known and used in the prediction. It is no doubt frequently used in practice. We will want to develop better approximations than (5.1) exploiting additional information when the service-time distribution is not exponential.

EXAMPLE 5.1. To show how approximation (5.1) can perform suppose that the service times are deterministic, all equal to μ^{-1}. Moreover, first suppose that $k + 1 \le s$. The remaining service times of the s customers in service are necessarily less than μ^{-1}. If these s customers all entered service together, then the actual waiting time would be $W = (k + 1)r$, where r is the common remaining service time, $0 < r < \mu^{-1}$. Thus W can vary from 0 to $(K + 1)/\mu$, with both extremes differing substantially from (5.1). On the other hand, when the s remaining service times are $j/s\mu$, $1 \le j \le s$, formula (5.1) is exact. A more realistic case is the s service times being i.i.d. with the stationary-excess cdf G_e in (6.2), which in this case is uniform on $[0, \mu^{-1}]$. Then W is distributed as the $(k + 1)$st smallest among the s uniform random variables, which has mean $(k + 1)/(s + 1)\mu$, which is very close to (5.1) when s is not too small. Similar observations can be made for $k \ge s$. If $k = ns$, then $W = n\mu^{-1} + r_{\min}$, where $r_{\min}$ is the minimum of the initial remaining service times. The error in (5.1) is then $|r_{\min} - (s\mu)^{-1}|$. In summary, with deterministic service times, the mean wait depends on the way previous customers entered service, which in turn depends upon the arrival process. The error and the uncertainty are removed if we know the remaining

service times, which in this case would occur if we know the elapsed service times.

EXAMPLE 5.2. Example 3.1 also provides a basis for evaluating approximation (5.1). There the mean service time is $\mu^{-1} = p\mu_1^{-1}$. Approximation (5.1) yields $EW \approx (1 + k)p/s\mu^{-1}$, whereas the exact formula in (3.1) is $EW = (1 + kp)/s\mu^{-1}$. If $kp \gg 1$, the approximation is close, but not otherwise.

Now we want to extend approximation (5.1) to an approximation for the full waiting-time cdf and, especially, the standard deviation. To do so, we think of the case in which all servers are busy for an extended period of time. During this period, the departure process can be identified with the superposition of s i.i.d. renewal processes, where the interrenewal times are the service times. In that situation, the number $D(t)$ of departures in the interval $[0, t]$ has mean $s\mu t + o(1)$ as $t \to \infty$ and has variance $s\mu c_s^2 t + o(1)$ as $t \to \infty$, where c_s^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of a generic service time S; e.g., see p. 372 of Feller (1971). This asymptotic behavior is matched by a *single* renewal process with interrenewal-time S/s. Hence, we propose approximating the waiting time by the sum of $k + 1$ i.i.d. random variables distributed as S/s. This coincides with the waiting time in a single-server queue with service times S/s if we think of the customer in service just starting service at time 0. (For a single-server queue, we could improve upon (5.1) by using the mean conditional remaining service time for the customer in service.) This gives approximation (5.1) for the mean. The associated approximating Laplace transform of W (assuming $k + s$ customers are ahead of the current customer) is

$$Ee^{-zW} \approx (E[e^{-zS/s}])^{k+1}. \tag{5.2}$$

Similarly, we approximate the standard deviation by

$$SD(W) \approx \sqrt{k+1}\,\frac{SD(S)}{s}. \tag{5.3}$$

Note that (5.1)–(5.3) are all correct for an exponential service-time distribution. The waiting time should be asymptotically normally distributed with mean in (5.1) and standard deviation in (5.3) as $k \to \infty$, by virtue of the central limit theorem. Hence, we propose the normal distribution with mean in (5.1) and standard deviation in (5.3) as our approximating waiting-time cdf based only on the number of customers ahead of the customer of interest. We anticipate that this approximation will perform better for smaller s and larger k. As noted in §2, if k is large, then the standard deviation tends to be small compared to the mean, so that the mean in (5.1) becomes a very accurate prediction.

If the service-time distribution is approximately exponential, then (5.1)–(5.3) should be good approximations, but in general, if k is not large, then they can be crude approximations. To illustrate, in Example 3.1 the approximate variance by (5.3) is $(k + 1)(2p - p^2)/(s\mu_1)^2$, whereas the exact formula is $(1 + 2kp - kp^2)/(s\mu_1)^2$.

Formulas (5.1) and (5.3) are important for judging whether or not additional information should be beneficial. Assuming that the mean in (5.1) is relatively accurate, if the estimated standard deviation $SD(W)$ is small compared to the mean, i.e., if $SD(S)/E(S)\sqrt{(k+1)}$ is small, then there should be relatively little need to condition on extra information. We can see the advantage of the two-class algorithm in §3 by noting that the approximation for $SD(W)$ is relatively large for H_2 service-time distributions with component exponentials having very different means.

6. Exploiting Remaining Service-Time CDFs

The delay predictions in §§2–4 depend strongly upon the exponential assumptions. For other service-time cdf's, the remaining-service-time cdf depends on the elapsed service time. Thus, for non-exponential service-time cdf's, we can more accurately predict the delay of a new arrival if we exploit the elapsed service times (ages) of the customers in service. This step becomes even more effective if we can also classify customers into different types, where different types have very different service-time distributions. (Examples were mentioned at the beginning of §3.) This classification may be done before or after service has begun.

Henceforth, assume that the classification has been done before service begins, so that customer i before

starting service has service-time cdf G_i. (We label the customers starting with those in service.) The actual service-time cdf G_i should be easily estimated directly from the observed service times, assuming that there is no reneging after service has begun and that service times in progress are not altered by system state. In practice, this last possibility should be checked. It can be checked by estimating service-time distributions conditional on the number in system when service starts. With significant reneging, the estimation procedures should account for censoring.

Suppose that the service provider keeps track of the starting time for each service in process, so that at the time of a new arrival, the elapsed service times (ages) of the service times of all customers in service are known. Let x_i be the elapsed service time of customer i in service. Let $G_i(t \mid x_i)$ be the cdf of the conditional remaining service time, conditional on an elapsed service time x_i. Clearly,

$$G_i(t \mid x_i) = \frac{G_i(t + x_i) - G_i(x_i)}{1 - G_i(x_i)}, \quad t \geq 0, \ 1 \leq i \leq s. \tag{6.1}$$

If additional prediction is done after service has started (using service time), then $G_i(t \mid x_i)$ could be estimated directly instead of by (6.1). It is important to recognize that new information might well be gained once service has started. First, the customer's service time might depend significantly upon the service agent assigned to the task. There might be different service-time distributions for different combinations of customer type and service-agent type. Moreover, additional classification may be possible once service has begun. An initial step in providing service may involve customer classification. Service agents could even be generating updated predictions of remaining service times for the customers they are serving while service is in progress.

The importance of conditioning upon the ages clearly increases as the service-time distribution differs more from an exponential distribution. The difference is clearly dramatic when the original service time is deterministic or, more generally, has low variability. The difference is also dramatic when the service-time distribution is a long-tail distribution such as the Pareto distribution. Indeed, suppose that $Y(a, b)$ has the Pareto cdf $G(t) = 1 - (1 + bt)^{-a}$, $t \geq 0$. Let $Y_x(a, b)$ have the conditional cdf $G(t \mid x)$. Then, by Theorem 8 of Duffield and Whitt (1997), $Y_x(a, b)$ is distributed the same as $(1 + bx)Y(a, b)$. Hence, the mean residual life is approximately proportional to the age. In this setting the age can greatly help in predicting the residual life.

It might happen that we have nonexponential service-time cdf's, but we cannot observe the elapsed service times of the customers in service. If the service times are i.i.d. with cdf G having mean m, then we may elect to approximate the s remaining-service-times at any time by i.i.d. random variables with the stationary-excess cdf G_e, defined by

$$G_e(t) = m_1^{-1} \int_0^t G^c(u)du, \quad t \geq 0, \tag{6.2}$$

where $G^c(t) = 1 - G(t)$, which has mean $m_{e1} = m_2/2m_1$. If G is exponential, then $m_2 = 2m_1^2$ and $m_{e1} = m_1$, but more generally m_{e1} need not equal m_1. This approximation is exact for the $M/GI/\infty$ model, e.g., see Duffield and Whitt (1997), and tends to be correct if the servers have all been busy for a long time, e.g., see Coffman Jr. et al. (1996).

We also want to account for customer-dependent reneging. Let $H_i(t)$ be the probability that customer i (in queue) will abandon if he has not received service by time t. The problem now is to convert the estimated remaining-service-time cdf's—$G_i(t \mid x_i)$ for customers in service and $G_i(t)$ for customers in queue—and the reneging cdf's $H_i(t)$ into an estimated waiting-time distribution (before beginning service) for a new arrival (or any other customer in queue). First, note that if we have this information for each customer ahead of the customer of interest, then the waiting time does not depend on the arrival process or any blocking and balking that might occur, just as in §§2 and 3. Second, note that the prediction is relatively easy when there is no reneging in the case $s = 1$; then the waiting time is just the sum of the remaining independent service times, and for k not too small we can use a normal approximation based on the sums of the means and variances. Given the Laplace transforms of the

component distributions, we can express the Laplace transform of the waiting time as the product of the remaining-service-time Laplace transforms, just as in (2.12), and then calculate the waiting-time cdf by numerical transform inversion. Here we are primarily concerned with $s > 1$ and are thinking of large s, e.g., $s = 100$, as in a telephone call center.

Our problem then is to predict the waiting time of customer $k + 1$ in queue, under the assumption that the customer of interest does not renege. To develop our prediction, we use the fact that the waiting time can be expressed exactly in terms of the departure process, i.e., the number $D(t)$ of departures in the interval $[0, t]$, $t \geq 0$, where 0 is the initial time when the prediction is to be made and we include reneging in the departures. In particular, the actual waiting time of a customer with $s + k$ customers ahead of him is

$$W = \min\{t \geq 0 : D(t) = k + 1\}. \tag{6.3}$$

We approximate the mean EW by approximating $D(t)$ in (6.3) by its mean $ED(t)$ for all $t \geq 0$; i.e., we estimate the mean waiting time by

$$EW \approx \min\{t > 0 : ED(t) = k + 1\}. \tag{6.4}$$

Intuitively, we can justify approximation (6.4) for large waiting times by the observation that if t is suitably large, then $D(t)$ should be relatively close to its mean $ED(t)$, by an appropriate law of large numbers. In turn, if k is suitably large, then t in (6.4) should be suitably large. Formally, such large-t limits for the departure process can be related to large-t limits for the arrival process; e.g., see §2 of Whitt (1984) and §5 of Berger and Whitt (1992).

It now remains to develop an approximation for the mean $ED(t)$. For this purpose, let $D_s(t)$ be the number of the original s customers in service that will have departed t time units later. Then, given the s ages $x_1, \ldots, x_s$, its expected value is (exactly)

$$ED_s(t) = \sum_{i=1}^{s} G_i(t \mid x_i), \quad t \geq 0. \tag{6.5}$$

We then let t_j denote the estimated time when the jth customer in queue (originally) can enter service. We define t_j recursively by

$$t_1 \equiv \min\{t \geq 0 : ED_s(t) = 1\} \tag{6.6}$$

and, for $j \geq 2$,

$$t_j \equiv \min\left\{ t \geq 0 : ED_s(t) + \sum_{l=s+1}^{s+j-1} [H_l(t_l) + (1 - H_l(t_l))G_l(t - t_l)] \right\}. \tag{6.7}$$

The summands inside (6.7) represent the probability that customer l reneges before time t_l, $H_l(t_l)$, plus the probability that the customer does not renege by time t_l but then does complete service by time t, assuming these events to be independent. Given that the customer of interest sees $s + k$ customers ahead of him, our approximation for the mean wait is

$$EW \approx t_{k+1}. \tag{6.8}$$

Since the functions of time in (6.6) and (6.7) are monotone, we can use bisection search to quickly find the values t_j.

We now give alternative approximations for EW based on (6.4) and *bounds* for $D(t)$. We obtain a lower bound for $D(t)$ and thus $ED(t)$ by assuming that the customers in queue never renege or start service. We obtain an upper bound by assuming that they start service immediately at time 0 and are simultaneously subject to reneging. Thus,

$$ED_s(t) \leq ED(t) \leq ED_s(t) + \sum_{i=s+1}^{s+k} [G_i(t) + H_i(t)] \tag{6.9}$$

for $ED_s(t)$ in (6.5). Using these bounds in (6.4) yields more elementary approximations for EW that bound approximation (6.8) above and below. They tend to be reasonable rough approximations when k is substantially smaller than s.

EXAMPLE 6.1. Since we have convenient exact formulas for the $A/M/s/r$ model without reneging in §2 with i.i.d. exponential service times having mean μ^{-1}, we clearly do not need an approximation for that case, but it is useful to consider it in order to evaluate the performance of the approximation. Then all remaining service times are exponential with mean μ^{-1}. By

induction, it follows from (6.7) that $t_j = -j\mu^{-1}\log(1 - s^{-1})$, so that

$$EW \approx t_{k+1} = -\frac{(k+1)}{\mu}\log(1 - s^{-1})$$
$$\approx \frac{k+1}{\mu s}\left(1 + \frac{1}{2s}\right), \tag{6.10}$$

which in general is an overestimate of the true mean in (2.1), but is very accurate when s is not too small. Using the lower bound $ED_s(t)$ in (6.9) with $H_i(t) = 0$, we obtain the upper-bound for (6.8)

$$EW \approx -\frac{1}{\mu}\log\left(1 - \frac{(k+1)}{s}\right)$$
$$\approx \frac{k+1}{\mu s}\left(1 + \frac{k+1}{2s}\right). \tag{6.11}$$

Using the upper bound in (6.9), we obtain the lower bound for (6.8)

$$EW \approx -\frac{1}{\mu}\log\left(1 - \frac{(k+1)}{s+k}\right)$$
$$\approx \frac{k+1}{\mu(s+k)}\left(1 + \frac{k+1}{2(s+k)}\right)$$
$$= \frac{k+1}{\mu s}\left(1 - \frac{(k-1)}{2(s+k)} - \frac{k(k+1)}{(s+k)^2}\right). \tag{6.12}$$

When $k \ll s$, (6.10), (6.11), and (6.12) are very close. For instance, if $s = 400$, $k = 80$ and $\mu^{-1} = 5$ minutes as in Example 2.1, then the exact conditional mean wait EW is 60.9 seconds, while the three approximations in (6.12), (6.10), and (6.11) are 55.5, 60.8, and 67.8 seconds, respectively.

We now estimate the full distribution of W. What we have just done is equivalent to estimating the mean $ED(t)$ by

$$ED(t) \approx \sum_{i=1}^{s} G_i(t \mid x_i) + \sum_{j=s+1}^{s+k} F_j(t), \tag{6.13}$$

where t_j is in (6.7) and

$$F_j(t) = H_j(t_j) + [1 - H_j(t_j)]G_j(t - t_j), \tag{6.14}$$

with $G_j(t) = 0$ and $F_j(t) = 0$ for $t < 0$. Motivated by (6.13), we suggest approximating the distribution of $D(t)$ by the distribution of the sum of $s + k$ independent random variables with cdf's $G_i(t \mid x_i)$, $1 \le i \le s$, and $F_j(t)$, $s + 1 \le j \le s + k$. Thus, we estimate the variance of $D(t)$ by

$$\text{Var } D(t) \approx \sum_{i=1}^{s} G_i(t \mid x_i)(1 - G_i(t \mid x_i))$$
$$+ \sum_{j=s+1}^{s+k} F_j(t)(1 - F_j(t)). \tag{6.15}$$

We anticipate that (6.15) will underestimate the true variance of $D(t)$ because the customers in queue are treated as if they enter service at the fixed times t_j. We apply the central limit theorem for independent nonidentically distributed random variables, p. 262 of Feller (1971), to justify regarding $D(t)$ as approximately normally distributed with mean in (6.13) and variance in (6.15). Supporting theoretical results appear in Duffield and Whitt (1997).

Given the normal approximation for $D(t)$, (6.13) and (6.15), we can estimate the full waiting-time distribution (approximately). Let $N(0, 1)$ denote a standard (mean 0, variance 1) normal random variable and let Φ be its cdf. By (6.3),

$$P(W > t) = P(D(t) < k + 1)$$
$$= P\left(\frac{D(t) - ED(t)}{SD(D(t))} < \frac{k + 1 - ED(t)}{SD(D(t))}\right)$$
$$\approx \Phi([k + 1 - ED(t)]/SD(D(t))). \tag{6.16}$$

To estimate the $(1 - \alpha)$th percentile of the cdf of W, i.e., to find w_{x_α} such that $P(W > w_{x_\alpha}) \approx \alpha$, let x_α be such that $\Phi(x_\alpha) = \alpha$; i.e., $x_\alpha = \Phi^{-1}(\alpha)$. Then let

$$w_x = \min\{t \ge 0 : ED(t) + xSD(D(t)) = k + 1\}. \tag{6.17}$$

Combining (6.16) and (6.17), we see that w_{x_α} is the approximate $(1 - \alpha)$-percentile of the distribution of W, i.e.,

$$P(W \ge w_{x_\alpha}) \approx \alpha. \tag{6.18}$$

From (6.16) or (6.18), we can obtain the estimated

complementary cdf $P(W > w)$ and then compute any desired summary characteristic. We also suggest using the estimated median, either directly or as an estimate of the mean, which leads to (6.4) with $ED(t)$ estimated in (6.13), which is equivalent to (6.8). A conservative point estimate of the waiting time might be the estimated 90th percentile.

REMARK. Note that the approximations in (6.9) extend to the variance $\text{Var}(D(t))$ in (6.15), yielding alternative approximations for the full cdf of W. The first approximation for $\text{Var}(D(t))$ is the first term in (6.15), while the second approximation has $t_j = 0$ for all j in the second term of (6.15).

EXAMPLE 6.2. Suppose that we consider the limiting two-class case in Example 3.1 in which class 2 has 0 service times. Suppose also that we learn the identity of all customers upon arrival, including the ones in queue. Then we can ignore class 2 customers in queue, so that we are faced with the $A/M/s/r$ problem in Example 6.1. Suppose that there is no reneging. As in Example 6.1, when we apply (6.7), we get $t_j = -j\mu_1^{-1}\log(1 - s^{-1})$. Then, by (6.13) and (6.15), when there are $s + k$ class-1 customers in the system, the mean and variance of $D(t)$ are approximately

$$ED(t) \approx s(1 - e^{-\mu_1 t}) + k - e^{-\mu_1 t}\left(\frac{a(a^k - 1)}{a - 1}\right) \quad (6.19)$$

and

$$\text{Var } D(t) \approx e^{-\mu_1 t}\left(s + \frac{a(a^k - 1)}{a - 1}\right) - e^{-2\mu_1 t}\left(s + \frac{a^2(a^{2k} - 1)}{a^2 - 1}\right), \quad (6.20)$$

where μ_1^{-1} is the mean class-1 service time and $a = s/(s - 1)$. Suppose that the number of class-1 customers in the system is initially $s + k$ with $s = 100$, $k = 80$ and $\mu_1 = 1$. At time $t = 0.9$, from (6.19) and (6.20) we get the approximations $ED(0.9) \approx 89.2$ and $SD(D(0.9)) \approx 6.45$. However, when all servers are busy, $D(t)$ actually has a Poisson distribution with mean and variance $s\mu_1 t = 100(1)(0.9) = 90$. Thus the approximation for the mean is very accurate, but the approximate standard deviation underestimates the true value $\sqrt{90} \approx 9.5$. By (6.16), we obtain the approximation

$$P(W > 0.9) \approx \Phi((81 - 89.2)/6.45) = \Phi(-1.27) \approx 0.10. \quad (6.21)$$

By the normal approximation using the exact mean and variance in (2.1),

$$P(W > 0.9) \approx \Phi(-1.0) = 0.16. \quad (6.22)$$

We regard (6.21) as a reasonable rough approximation for (6.22). The approximations look better if we focus on percentiles. The actual 50th and 90th percentiles are 0.81 and 0.92, while the approximations are 0.81 and 0.89. In the next section we will see that the approximation in this section tends to be substantially more accurate when the remaining service times have less variability than the exponential distribution considered here.

7. Known Remaining Service Times Without Reneging

Suppose that the remaining service times of all customers in the system are known at each instant, and that there is no reneging, so that there is no uncertainty about the remaining service times when the prediction is to be made. This case includes many subcases. First, the original service times may have been deterministic, and may or may not have had a common value. In either case, the remaining service times of the customers in service will typically not have a common value, and we need to exploit the elapsed service times to know the remaining service times. Second, the original service times may have been random, e.g., as in §§2–4, but the predictor may learn these service requirements after the customers arrive.

It is significant that the approximation (6.8) in §6 is *exact* when the remaining service times are known and thus is no reneging. Then the cdf's $G_i(t \mid x_i)$ and $G_{s+i}(t)$ are all step functions, i.e., $G_i(t \mid x_i) = 0$ for $t < S_i$ and $G_i(t \mid x_j) = 1$ for $t \geq S_i$, where S_i is the remaining service time. Hence $D(t) = ED(t)$, so that the waiting time W in (6.3) is deterministic and coincides with the approximation for EW in (6.8).

Indeed, with deterministic remaining service times, the approximation in §6 can be seen to coincide with a familiar recursion for the waiting time or departure time.

The desired recursion is a variant of the classical recursion for the successive waiting times in a general infinite-capacity s-server FCFS queue; e.g., p. 81 of Baccelli and Bremaud (1994). In our setting we have no interarrival times to consider. Let $V_n = (V_{n1}, \ldots, V_{ns})$ be an s-dimensional workload vector for each n. Let S_n be the nth remaining service time, let e be the s-dimensional unit vector $e = (1, 0, \ldots, 0)$, and let $\mathcal{R}$ be the operator on vectors which rearranges them in increasing order. Then let $V_0 = (0, \ldots, 0)$,

$$V_{n+1} = \mathcal{R}(V_n + S_n e), \quad n \geq 1, \tag{7.1}$$

and the desired waiting time is

$$W_{s+k} = V_{s+k,1}. \tag{7.2}$$

This is an important theoretical reference point for the approximation procedure proposed in §6. Since the approximation is exact for known deterministic service times, it is evident that the approximation in §6 should perform well when the remaining service times have low variability.

EXAMPLE 7.1. We now want to show that it can be very important to focus on the ages and remaining-service-time cdf's. To dramatically make this point, we consider a very idealized model, in particular, an $A/D/100$ model with constant service times of length 100 and batch arrivals of size 120 every 400 time units. Thus, the first 100 customers in each batch go into service immediately, while the remaining 20 customers wait exactly 100, after which the servers are idle for an interval of length 200, and the process repeats. In contrast, if we ignore the remaining service times and apply the approximation in §5, we would estimate the waiting time to be k for customer number $100 + k$ in the batch (assuming he counts the arrivals ahead of him in the same batch). Since the method of §6 coincides with the deterministic recursion, it yields the exact waiting times for this example, avoiding this big error. Similar behavior will hold for more general low-variability service times and bursty arrival processes.

We may even elect to use the deterministic recursion as an approximation. When the remaining service times have low variability, we can act as if each remaining service time is in fact deterministic with a value equal to its mean. Then it suffices to apply the recursion in (7.1)–(7.2), and we do not need to determine the cdf's of the remaining service times.

EXAMPLE 7.2. However, the deterministic recursion using mean remaining service times can be a poor approximation when the remaining service times have substantial variability. To illustrate the problem, consider the $A/M/s/r$ model with i.i.d. exponential service times having mean μ^{-1}. If we use the deterministic recursion with the means, then all service times are treated as deterministic with mean μ^{-1} (clearly not a good approximation). The deterministic recursion thus predicts that the first s departures occur together at time μ^{-1}. If $k < s$, then the deterministic recursion leads us to approximate the mean wait by $EW \approx \mu^{-1}$, when the actual value in (2.1) is $(k + 1)/s\mu$. If $k \ll s$, then there is a big error in the prediction.

A more refined approximation when the remaining service times have greater variability is to perform simulations in real time to estimate the waiting time cdf. To do so, it suffices to generate independent random variates with the remaining-service-time cdf's and apply the deterministic recursion in (7.1)–(7.2). Since the simulation reduces to generating random variates and applying the recursion (7.1) and (7.2), the simulation can be performed very quickly for real-time estimates. Multiple independent replications provide an estimate of the waiting-time cdf. From a practical point of view, even a modest number of replications (e.g., 20 or fewer) may provide a satisfactory estimate of the variability of the waiting time as well as the mean. As discussed in §§2 and 6, given that we exploit system state information, we anticipate that the waiting time should be approximately normally distributed. Thus, with simulation, we should be approximately in the setting of estimating the mean and variance of a normal distribution from an i.i.d. sample. In that setting, the sample variance has a chi-square distribution with $n - 1$ degrees of freedom. For n not too small, the chi-squared distribution

is approximately normal. Moreover, the sample variance has mean σ^2 and standard deviation $\sigma^2(\sqrt{2/(n-1)})$. For $n = 20$, the standard deviation is 0.32 σ^2, so that we are likely to know σ^2 to within a factor of 2 from a sample of size 20.

8. Special Methods for Customers Near the Front of the Queue

In some applications we may be especially interested in the waiting times of the first few customers in queue. Let W_k be the waiting time of the kth customer in line. Then the complementary cdf of the waiting time of the first customer in queue is exactly

$$P(W_1 > t) = \prod_{i=1}^{s} (1 - G_i(t \mid x_i)) \tag{8.1}$$

which is easily calculated via

$$\log P(W_1 > t) = \sum_{i=1}^{s} \log(1 - G_i(t \mid x_i)). \tag{8.2}$$

We can approximate W_1 by an exponential distribution

$$P(W_1 > t) \approx e^{-\alpha_1 t}, \quad t \geq 0, \tag{8.3}$$

where α_1 is obtained from (8.1) via

$$\alpha_1 \approx \frac{\log P(W_1 > t_0)}{t_0} \tag{8.4}$$

for some appropriate t_0. To be specific, we might choose t_0 to be a rough estimate of EW_1, i.e., $EW_1 \approx 1/\sum_{i=1}^{s} (1/m_i)$ where m_i is the mean of $G_i(t \mid x_i)$. Approximation (8.3) and the approximation for EW_1 are exact when the cdf's $G_i(t \mid x_i)$ are all exponential; then (8.4) is independent of t_0. For large s, approximation (8.3) is supported by extreme-value limits in the i.i.d. case; see Leadbetter et al. (1983) and Resnick (1987). As a supporting regularity condition, we assume that $G_i(t)$ and thus $G_i(t \mid x_i)$ has a positive density on the entire half line.

Now we consider how to extend the approximation to other customers in queue when s is large compared to k. We propose acting as if the initial departure process is a Poisson process with rate α_1 for α_1 in (8.4), so that W_k has approximately a gamma distribution with mean and variance

$$EW_k = \frac{k}{\alpha_1} \quad \text{and} \quad \text{Var } W_k = \frac{k}{\alpha_1^2}. \tag{8.5}$$

Theoretical support for approximating the departure process by a Poisson process appears in Whitt (1984).

What we have just done in this section is equivalent to approximating the model by the $A/M/s/r$ model of §2 by defining an appropriate service rate μ, e.g., via (8.4). Alternatively, we could define μ by $\mu \equiv s^{-1} \sum_{i=1}^{s} (1/m_i)$, where m_i is the mean remaining service time of customer i. Having determined μ, we can also account for reneging just as in §2. If reneging is significant, then this method may be superior to the methods in §§6 and 7, which do not account for reneging.

9. Simulation Experiments

We validated the approximation methods by indicating scenarios for which they are exact. Since the approximations in §§4, 5, and 8 are exact for the $A/M/s/r$ model, but not for deterministic remaining service times, while the approximations in §§6 and 7 are exact for deterministic service times, but not for the $A/M/s/r$ model, it is evident that there is not one universally best method. We thus propose using computer simulation to evaluate the alternative methods in any desired application context. However, the prediction schemes are somewhat difficult to evaluate because the predictions depend on information conditions that vary. In this section we describe two ways to validate the predictions using simulation experiments.

It is relatively straightforward to evaluate point estimates, i.e., predictions of the conditional mean. We can simulate the queue and generate a predicted waiting time $\hat{W}_n$ for the nth arrival for each n, $1 \leq n \leq N$, using whatever information is to be used upon arrival. We can then subsequently observe the actual waiting time W_n of each of these customers and compare them. For example, we can look at the standard error (square root of the mean squared error)

$$\text{SDE} = \left[N^{-1} \sum_{n=1}^{N} (\hat{W}_n - W_n)^2 \right]^{1/2}. \qquad (9.1)$$

We can conclude that one scheme is better than another if it has smaller SDE. A further frame of reference is the steady-state mean EW_∞. If the system is simulated in steady state (e.g., by deleting an initial segment of the run), then the sde of EW_∞ is just the sample standard deviation, which converges to the steady-state standard deviation $SD(W_\infty)$ as $N \to \infty$. We can estimate the mean EW_∞ and standard deviation $SD(W_\infty)$ by the sample mean and standard deviation, i.e.,

$$\bar{W}_N \equiv N^{-1} \sum_{n=1}^{N} W_n \quad \text{and}$$

$$s_N = \left[N^{-1} \sum_{n=1}^{N} (W_n - \bar{W}_N)^2 \right]^{1/2}. \qquad (9.2)$$

We want the sample mean $N^{-1} \sum_{n=1}^{N} \hat{W}_n$ of the estimates to be close to $\bar{W}_N$ and the sde in (9.1) to be substantially smaller than s_N.

It is not clear, however, that (9.1) will always be the desired measure of error or loss. For example, we may be primarily concerned about prediction only when W_n is large. Moreover, we may be more concerned about relative error than absolute error when W_n is large.

A second validation approach starts from some specified initial system state with some number of customers in the system and the remaining-service-time cdf of each of these customers. Such initial system states can be generated by simulation. Given any initial system state, we estimate the conditional waiting-time cdf or summary statistics by any of the proposed prediction methods and compare the predictions to the actual values. We obtain an estimate of the actual waiting-time cdf or any desired summary statistic by generating random remaining service times according to their conditional cdf's and then applying the deterministic recursion in §6. If $F(t) \equiv P(W \le t)$ is the true cdf value, the estimated value obtained as the average from n independent replications has mean $F(t)$ and variance $F(t)(1 - F(t))/n$.

10. Conclusions

We have shown that it is often possible to reliably predict the waiting-time of a new arrival or a customer already in a multiserver FCFS queue, given the number of customers ahead of the designated customer and other state information. If we know all the remaining-service-time cdf's, then the waiting-time cdf does not depend on the arrival process or balking and blocking that might occur at arrival epochs. We gave algorithms to compute the exact waiting-time cdf's for the cases of: i.i.d. exponential service times with state-dependent reneging (§2), two classes of customers with two different exponential service-time distributions (§3) and fully known remaining service times (§7). We developed algorithms for approximately estimating the waiting-time cdf given remaining-service-time cdf's by exploiting approximations for the departure process $D(t)$ (§§6 and 8) and simulations based on the deterministic recursion (§7). The approximations were supported by indicating settings in which they are exact. The approximation in §6 is exact for deterministic remaining service times, while the approximations in §§4, 5, and 8 are exact for i.i.d. exponential service times. We also investigated how the approximations perform in certain cases; e.g., we applied the method in §6 to the $A/M/s/r$ model in §2. We conclude that there is not one universally best approximation. The approximation in §6 is our leading candidate, but it requires more information than the others (the service-time cdf's and ages). We showed how to evaluate the approximations in specific situations by performing computer simulations (§9). However, further study is needed to better understand the performance of the approximations.

We focused on the waiting time, i.e., the time to start service, but interest might instead be focused on the response time, i.e., the time to complete service. Assuming that the service time of each customer in queue is independent of his waiting time to begin service, the distribution of the time to complete service is naturally estimated by the convolution of the two estimated component distributions, which is a straightforward extension. In particular, the mean and variance of the conditional response time are just the sums of the means and variances. However, if there is

a large number of servers and considerable uncertainty about service times, then the variance of a customer's service time may be much larger than the variance of the customer's waiting time. (For example, in the setting of (2.1), the standard deviation of a waiting time is $\sqrt{k+1}/s$ times μ^{-1}, the standard deviation of a service time.) Thus, uncertainty about a customer's service time may make it substantially more difficult to reliably predict response times than waiting times.[1]

[1] The author thanks Robert L. Hails, Jr., and the referees for helping him improve the paper.

References

Abate, J., W. Whitt. 1992. The Fourier-Series Method for inverting transforms of probability distributions. *Queueing Systems* **10** 5–88.

——, ——. 1995. Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comput.* **7** 36–43.

Abramowitz, M., I. Stegun. 1972. *Handbook of Mathematical Functions*. National Bureau of Standards, Washington, DC.

Baccelli, F., P. Bremaud. 1994. *Elements of Queueing Theory*. Springer-Verlag, New York.

Berger, A. W., W. Whitt. 1992. The impact of a job buffer in a token-bank rate-control throttle. *Stochastic Models* **8** 685–717.

Coffman, Jr., E. G., L. Flatto, W. Whitt. 1996. Stochastic limit laws for schedule makespans. *Stochastic Models* **12** 215–243.

Duffield, N. G., W. Whitt. 1997. Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems* **26** 69–104.

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*. Vol. II, 2nd ed. Wiley, New York.

Gross, D., C. M. Harris. 1985. *Fundamentals of Queueing Theory*, 2nd ed. Wiley, New York.

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

Hui, M. K., D. K. Tse. 1996. What to tell customers in waits of different lengths: an integrative model of service evaluation. *J. Marketing* **60** 81–90.

Katz, K. L., B. M. Larson, R. C. Larson. 1991. Prescription for the waiting-in-line blues: entertain, enlighten and engage. *Sloan Management Rev.* **32** 44–53.

Leadbetter, M. R., G. Lindgren, H. Rootzén. 1983. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York.

Resnick, S. I. 1987. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York.

Stanford, D. A., B. Pagurek, and C. M. Woodside. 1983. Optimal prediction of times and queue lengths in the $GI/M/1$ queue. *Oper. Res.* **31** 322–337.

Taylor, S. 1994. Waiting for service: the relationship between delays and evaluations of service. *J. Marketing* **58** 56–69.

Whitt, W. 1984a. Departures from a queue with many busy servers. *Math. Oper. Res.* **9** 534–544.

——. 1984b. Approximations for departure processes and queues in series. *Naval. Res. Logist. Quart.* **31** 499–521.

——. 1999. Improving service by informing customers about anticipated delays. *Management Sci.* **45** 192–207.

Woodside, C. M., D. A. Stanford, B. Pagurek. 1984. Optimal prediction of queue lengths and delays in $GI/M/m$ multiserver queues. *Oper. Res.* **32** 808–817.

Accepted by Linda V. Green; received December 15, 1997. This paper has been with the author 2 months for 3 revisions.