Improving Service by Informing Customers About Anticipated Delays

Ward Whitt
AT&T Labs, Shannon Laboratory, 180 Park Avenue, Florham Park, New Jersey 07932-0971

This paper investigates the effect upon performance in a service system, such as a telephone call center, of giving waiting customers state information. In particular, the paper studies two $M/M/s/r$ queueing models with balking and reneging. For simplicity, it is assumed that each customer is willing to wait a fixed time before beginning service. However, customers differ, so the delay tolerances for successive customers are random. In particular, it is assumed that the delay tolerance of each customer is zero with probability β, and is exponentially distributed with mean α^{-1} conditional on the delay tolerance being positive. Let N be the number of customers found by an arrival. In Model 1, no state information is provided, so that if $N \geq s$, the customer balks with probability β; if the customer enters the system, he reneges after an exponentially distributed time with mean α^{-1} if he has not begun service by that time. In Model 2, if $N = s + k \geq s$, then the customer is told the system state k and the remaining service times of all customers in the system, so that he balks with probability $\beta + (1 - \beta)(1 - q_k)$, where $q_k = P(T > S_k)$, T is exponentially distributed with mean α^{-1}, S_k is the sum of $k + 1$ independent exponential random variables each with mean $(s\mu)^{-1}$, and μ^{-1} is the mean service time. In Model 2, all reneging is replaced by balking. The number of customers in the system for Model 1 is shown to be larger than that for Model 2 in the likelihood-ratio stochastic ordering. Thus, customers are more likely to be blocked in Model 1 and are more likely to be served without waiting in Model 2. Algorithms are also developed for computing important performance measures in these, and more general, birth-and-death models.

(Service Systems; Telephone Call Centers; Balking; Reneging; Abandonments; Retrials; Birth-and-Death Processes; Communicating Anticipated Delays)

1. Introduction

In this paper we investigate alternative ways to manage a service system. We have in mind a telephone call center staffed by a group of operators, but there are other possible applications, e.g., internet access. We introduce birth-and-death (BD) stochastic process models that can be used to study the performance of, and demonstrate the advantage of: (1) allowing waiting before beginning service, and (2) communicating anticipated delays to customers upon arrival (or providing state information to allow customers to predict delays).

A frame of reference is the classical loss system, in which there are s servers working in parallel and no extra waiting space. Allowing waiting helps to avoid blocking and thus helps to serve more customers. However, in many loss systems with telephone access, blocked customers can retry rapidly and easily because of automatic redialers. When customers can retry rapidly, and elect to do so, the system without a provision for waiting tends to behave like the system with a provision for waiting. In that situation, the service provider serves many customers and avoids the cost of maintaining a queue. However, we contend

0025-1909/99/4502/0192$05.00

that it is usually better for the service provider to allow for waiting and directly manage it. Retrying typically imposes costs on both the customers and the service provider. First, the customer must expend time and effort retrying. Second, even unsuccessful attempts often consume resources of the service provider. Typically, some resources are required to process each request for service, whether or not it is successful. Thus, the service provider's processing capacity may be reduced by having to handle many unsuccessful attempts. Moreover, with retrials, the first-come, first-served (FCFS) service discipline is lost. The FCFS discipline is often strongly preferred by customers because of its inherent fairness. The random order of service associated with retrials also makes the waiting time before beginning service more variable, which tends to be detrimental. Thus, there are several reasons motivating service providers to allow for waiting.

Given that the service provider does allow for waiting, there are two alternatives. The service provider may either communicate anticipated delays to customers upon arrival or not. We contend that, once the service provider has decided to allow for waiting, it is usually better to inform customers about anticipated delays, assuming that there is the capability of doing so, which is more and more becoming the case, e.g., see Rappaport (1996). The primary reason for informing customers about anticipated delays is to increase customer satisfaction and, thus, obtain more repeat business. In the BD models, this advantage would be reflected by a higher arrival rate. However, we do not attempt to model how the arrival rate increases.

We present BD models that enable us to study system performance in both situations. If the service provider does communicate anticipated delays, then the customers are more likely to balk when all servers are busy (leave immediately upon arrival) than renege (leave after waiting for some time). We show how the BD models can be analyzed to describe and compare these alternatives.

It is common practice to restrict attention to the special case of the $M/M/s/r$ model, which has s servers and r extra waiting spaces. Indeed, it is common to use only the Erlang B (loss) model ($r = 0$) or the Erlang C (delay) model ($r = \infty$), but none of these alternatives account for balking or reneging. However, it is actually not difficult to account for balking and reneging in a BD model, and it is often very important to do so. By having a BD model that incorporates both possibilities, it is possible to evaluate the alternatives.

The way to calculate the steady-state distribution of a general BD process is quite well known. We go beyond that initial step by showing how to compute the probability that a customer receives service, the probability that a customer reneges, and the distributions and first two moments of the conditional response time given that service is completed and the conditional time to renege given that the customer reneges. These descriptions of the conditional distributions are helpful because the conditional distributions can differ significantly from the unconditional distributions.

These general BD models also can be used to study complex *networks* of service facilities. As in Whitt (1985), Kelly (1991) and Ross (1995), the BD model can serve as the fundamental building block for a reduced-load approximation for a network of service facilities. Then the overflows from one facility due to blocking, reneging or balking can become part of the arrival rate to other facilities. The overall performance can be determined by iteratively solving a system of nonlinear equations. The computational method is essentially the same as for the previously studied pure-blocking systems, but now the approach can be used for systems with balking and reneging as well as blocking. We intend to discuss such reduced-load approximations in a subsequent paper.

The BD model simplicity makes it possible to describe performance in detail using an elementary algorithm, but the model requires Markov assumptions such as exponential service-time distributions that may well be seriously violated in practice. The analytical BD model nevertheless can provide important insight. However, to actually predict customer delays in system operation, it may be better not to use the BD model; then it is possible to exploit nonexponential service-time distributions to make better

predictions; see the companion paper Whitt (1999). For further discussion about queue management and what to tell customers, see Hui and Tse (1996), Katz et al. (1991), Taylor (1994) and references therein.

The rest of this paper is organized as follows. In §2 we present Model 1, which is intended to represent the case in which no state information is communicated to customers. Customers only learn whether or not they can enter service immediately upon arrival. We regard Model 1 as the traditional BD model to describe performance when some arrivals balk and waiting customers renege after an exponential time. Model 1 can represent both the loss model with rapid retrials and the delay model for the case in which the service provider allows waiting, but provides no additional state information. With retrials, we do not try to directly represent the retrials as in Chapter 7 of Wolff (1989). Instead, assuming that relatively rapid retries are possible, we consider retrying customers to be waiting customers. However, we assume the FCFS service discipline, so that our analysis in §2 does not capture the random order of service associated with retrials.

In §3 we introduce Model 2, an alternative BD model to describe the performance when the service provider informs customers about anticipated delays before beginning service or provides state information so that the arriving customers can make this prediction. We relate the state-dependent balking in this setting to the reneging rate in §2. The principal change from §2 to §3 is to replace reneging with balking. In §3 we also introduce a more general BD model that includes both balking and reneging. For the general BD model, we develop algorithms to compute the probabilities of receiving service, blocking, balking and reneging. We also develop algorithms to compute the conditional distributions of the time to receive service and the time to renege given each outcome.

In §4 we make stochastic comparisons between Models 1 and 2, assuming common arrival rates, showing that state-dependent balking instead of reneging (at comparable rates) leads to fewer customers in the system in steady state. In §5 we present some numerical examples giving explicit comparisons. We obtain our numerical results by numerically solving for the performance measures in the BD models. These examples show that the performance in the two scenarios is often remarkably similar (again, assuming common arrival rates). The major difference is that, with balking instead of reneging, customers who do not receive service do not waste time waiting. We also use the numerical examples to show the economies of scale (having fewer groups of larger numbers of servers instead of more groups of smaller numbers of servers).

In §§6 and 7 we discuss ways to estimate model parameters and validate the BD models. Finally, in §8 we briefly discuss other possible deviations from the model assumptions and ways to approximately cope with them; e.g., we discuss ways to approximately capture the performance impact of occasional extra long service times. We refer to Boxma and de Waal (1995) and Falin (1990) for accounts of the literature on queues with reneging and retrials.

2. When Customers Receive No Information

In this section we review a reasonably well known BD model for the case in which the system state is not communicated to arriving customers, which we call Model 1; e.g., see Chapter 2 of Gross and Harris (1985) and Chapter 4 of Heyman and Sobel (1982). If a server is not immediately available, then the arriving customer *balks* (leaves immediately) with probability β and waits with probability $1 - \beta$. If a server is not immediately available and the customer does not balk, then he *reneges* (abandons later) after an exponential time with mean α^{-1}, if he has not yet begun service. We assume that the system state is not known by customers, so that the parameters α and β cannot depend directly on the number of customers in the system (beyond whether the servers are all busy or not). Once a customer starts service, he stays until service is completed. (It is easy to modify the BD model if this assumption is not reasonable.)

It is useful to consider how individual customer decision making can lead to the balking and reneging model above. To support what we do, we make a simple assumption about customer behavior. We assume that each customer is willing to wait a fixed time

before beginning service, called the delay threshold. However, different customers may have different delay thresholds, just as different customers have different utility functions, so that we assume that successive customers are willing to wait random times that are independent and identically distributed. (Customers are making their decisions deterministically. The delay thresholds are random because the customers with different delay thresholds are appearing at random.) With probability β, a customer is unwilling to wait any amount of time, and so will balk. Conditional on the customer being willing to wait at all, we assume that the customer is willing to wait a random time T. We assume that T has an exponential distribution with mean α^{-1}. (We need to assume that T has an exponential distribution in order to obtain the Markov property in the BD model.) When no state information is provided, we assume that the customer waits if necessary until his delay threshold, and then reneges if service has not yet been provided.

We now specify the rest of the BD model. Let the arrival process be a Poisson process with constant rate λ. Let there be s servers, a waiting room of size r and the FCFS discipline. (The total system capacity is thus $s + r$.) An arrival finding $s + r$ customers present is *blocked* (lost without retrying). Let the service times be i.i.d. exponential random variables with mean μ^{-1}. Then the birth (arrival) and death (departure) rates defining the BD model are, respectively,

$$\lambda_k = \begin{cases} \lambda, & 0 \le k \le s-1, \\ \lambda(1-\beta), & s \le k \le s+r-1, \end{cases} \tag{2.1}$$

and

$$\mu_k = \begin{cases} k\mu, & 1 \le k \le s-1, \\ s\mu + (k-s)\alpha, & s \le k \le s+r. \end{cases} \tag{2.2}$$

When the system state is k, i.e., when there are k customers in the system, the time until the next event is exponentially distributed with mean $(\lambda_k + \mu_k)^{-1}$; the event is an arrival with probability $\lambda_k/(\lambda_k + \mu_k)$; otherwise it is a departure. When $k > s$, some departures are service completions, while others are abandonments (reneging) and blocked arrivals. The long-run constant arrival rate λ must be the sum of the rates of service completion, blocking, balking, and reneging. We do not discuss how to analyze this BD model here because it is a special case of the model introduced in the next section, which we do analyze.

3. When Customers Receive Additional Information

In this section we consider the case in which customers learn the system state upon arrival. The customers may also receive updates while they are waiting. The customers might be told the number of customers ahead of them in line at all times (e.g., by displays on a monitor with access through a personal computer) and/or they might receive periodic predictions of their remaining time to wait before beginning service (e.g., by telephone announcements with access through a telephone).

Assuming that customers know their preferences, it is natural that customers would respond to this additional information when all servers are busy by replacing reneging after waiting with state-dependent balking; i.e., customers should be able to decide immediately upon arrival whether or not they are willing to join the queue and wait to receive service. Having joined the queue, customers should be much more likely to remain until they begin service. Reneging is even less likely if the customer can see that the remaining time to wait is steadily declining.

Hence in this section we consider an alternative BD model to represent state-dependent balking instead of time-dependent reneging. Since there may still be some reneging in this new situation with additional state information (e.g., because customers change their minds or because progress in the line is slower than anticipated), we also include reneging in the model. However, we are especially interested in the comparison between Model 1 in §2 with reneging and the special case of the general BD model in this section in which the reneging is replaced entirely by state-dependent balking, which we call Model 2. We make stochastic comparisons between Models 1 and 2 in §4.

As before, there is a Poisson arrival process with rate λ and s servers, each with exponential service times having mean μ^{-1}. There is a waiting room of size r and the FCFS service discipline. An arrival encountering a full system is blocked.

We now want to specify how customers respond to the additional state information when all servers are busy. This is a difficult modeling step, because there are several different kinds of state information that might be provided and, for any one, we must discern the human response. We shall first consider the case in which the new customer learns upon arrival what will be his required waiting time. In that situation, we assume that the customer balks if the required waiting time exceeds his (random) threshold; otherwise he stays to receive service. As in §2, the customer is unwilling to wait with probability β and, given that he is willing to wait, has a delay threshold T which is exponentially distributed with mean α^{-1}. If the number seen by the arrival (not including the arrival) is less than or equal to $s - 1$, then the new arrival enters service immediately. If the number seen by the arrival is $s + k$ for $0 \le k \le r - 1$, then the arrival may elect to balk (leave immediately) or join the queue. Paralleling §2, each customer finding all servers busy balks with probability β. However, the customer may also elect to balk depending on the system state. To be concrete, we suppose that the arriving customer learns k and the remaining service time of each customer in the system, and that the arriving customer acts as if all these customers will remain until they receive service; i.e., no allowance is made for reneging of customers ahead of the new customer. Then, the customer will join if his random delay threshold T exceeds the, now known, time until he is scheduled to begin service, assuming no reneging ahead of him. Thus, we stipulate that the customer joins with the probability that a server becomes free before he would abandon. Let S_k be the time required from arrival until a server first becomes available for this customer, as a function of k, assuming that departures occur only by service completions (not considering reneging by customers in queue ahead of the current customer). Then the arrival finding $s + k$ customers in the system upon arrival (not counting himself) joins with probability

$$q_k \equiv P(T > S_k), \quad 0 \le k \le r - 1. \tag{3.1}$$

Even though the customer will learn the remaining service times, what the customer will learn is unknown in advance, so that S_k in (3.1) is random. To make (3.1) a feasible computation, we assume that the balking decisions of successive customers are independent. In fact, this is not correct if customers learn remaining service times. However, the independence seems to be a reasonable approximation.

Alternatively, we can assume that customers learn only the system state k. Then we can just directly assume that (3.1) represents the balking behavior. Since T is exponential with mean α^{-1}, the state-dependent balking closely parallels the reneging in Model 1.

Since (i) S_k and T are independent, (ii) S_k has the distribution of the sum of $k + 1$ exponentials each with mean $(s\mu)^{-1}$, and (iii) T has an exponential distribution with mean α^{-1}, we can exploit Laplace transforms to calculate q_k in (3.1) explicitly. For this purpose, let $g_k(t)$ denote the probability density of S_k. Then

$$q_k = \int_0^\infty e^{-\alpha t} g_k(t) dt = Ee^{-\alpha S_k} = \left(\frac{s\mu}{s\mu + \alpha}\right)^{k+1}. \tag{3.2}$$

We also indicate several alternatives to (3.1) and (3.2). The first alternative is intended to represent the case in which the service provider communicates the expected delay when there are $s + k$ customers in the system. Then we would replace S_k in (3.1) by its mean, i.e., we would use

$$\bar{q}_k \equiv P(T > ES_k) = e^{-\alpha(k+1)/s\mu}, \quad k \ge 0. \tag{3.3}$$

Formula (3.3) is exact if we stipulate that the customer acts as if the mean ES_k were the actual delay, which may be a reasonable approximation.

REMARK. Note that when k is large, S_k will tend to be relatively close to ES_k by the law of large numbers, so that (3.2) and (3.3) should not differ greatly. Directly, we can see that, if k and s are suitably large, then (3.2) will be close to (3.3), i.e., using the approximation $(1 - (x_n/n))^n \approx e^{-x_n}$ (which is asymptotically correct as $n \to \infty$ if $x_n \to x$ as $n \to \infty$),

$$\left(\frac{s\mu}{s\mu+\alpha}\right)^{k+1} = \left(1-\frac{\alpha}{s\mu+\alpha}\right)^{k+1}$$
$$= \left(1-\frac{(k+1)\alpha}{(k+1)(s\mu+\alpha)}\right)^{k+1}$$
$$\approx e^{-(k+1)\alpha/(s\mu+\alpha)} \approx e^{-(k+1)\alpha/s\mu}. \quad (3.4)$$

In general, $\bar{q}_k \geq q_k$. □

The analysis leading to (3.2) and (3.3) suggests that the probability a customer joins the queue (does not balk) when he finds $s + k$ in system should be of the general form $\xi\eta^k$ for parameters ξ and η with $0 \leq \xi \leq 1$ and $0 \leq \eta \leq 1$. In practice the balking probability as a function of k needs to be estimated. This balking probability should depend on the information supplied to the customer.

We now define a general BD model representing state-dependent balking. Since there may still be some reneging, we include state-dependent reneging as well. Let a customer with $j - 1$ customers ahead of him in queue renege at a rate δ'_j. Then the total reneging rate when there are $s + k$ customers in the system is

$$\delta_k = \sum_{j=1}^{k} \delta'_j. \quad (3.5)$$

The birth (arrival) and death (departure) rates for the general BD model are, respectively,

$$\lambda_k = \begin{cases} \lambda, & 0 \leq k \leq s-1, \\ \lambda(1-\beta)q_{k-s}, & s \leq k \leq s+r-1, \end{cases} \quad (3.6)$$

and

$$\mu_k = \begin{cases} k\mu, & 1 \leq k \leq s, \\ s\mu + \delta_{k-s}, & s+1 \leq k \leq s+r. \end{cases} \quad (3.7)$$

In (3.7) we have allowed general state-dependent reneging rate for each waiting customer, δ'_k, but we will usually consider the special case in which $\delta_k = k\delta'$. Then $\delta_{k-s} = (k-s)\delta'$. Model 2 is the special case of (3.6) and (3.7) with $\delta_k = 0$. In Model 2, the customer delay parameter α enters in via q_k as defined in (3.2). In contrast, Model 1 has $q_k = 1$ and $\delta_k = k\delta'$ with $\delta' = \alpha$ (the same α as for Model 2).

We now indicate how to numerically solve for the steady-state probabilities p_k associated with the general BD model. Since the larger probabilities should be near s (assuming that s is reasonably well chosen), it is convenient to solve for the steady-state distribution recursively starting at s. Let $x_s = 1$,

$$x_{s+k+1} = \frac{\lambda_{s+k}x_{s+k}}{\mu_{s+k+1}}$$
$$= \frac{\lambda(1-\beta)q_k x_{s+k}}{s\mu+\delta_{k+1}}, \quad 0 \leq k \leq r-1, \quad (3.8)$$

and

$$x_{k-1} = \frac{\mu_k x_k}{\lambda_{k-1}} = \frac{k\mu x_k}{\lambda}, \quad 1 \leq k \leq s. \quad (3.9)$$

Then, let

$$y = \sum_{k=0}^{s+r} x_k \quad (3.10)$$

and

$$p_k = x_k/y, \quad 0 \leq k \leq s+r. \quad (3.11)$$

So far the results have been quite standard, but now we go on to compute the probability of completing service and the mean, variance and full distribution of the conditional response time (time to complete service) given that service is completed. We also compute the probability that a customer reneges and the mean, variance, and full distribution of the conditional time to renege given that the customer reneges.

Our approach is to condition on the state seen by arrivals and then average over all the possibilities. Since the arrival process is Poisson, the state seen by arrivals is the same as at an arbitrary time by the Poisson-Arrivals-See-Time-Average (PASTA) property; see §5.16 of Wolff (1989). Conditional on the arrival seeing $s + k$ customers in the system upon arrival, it suffices to consider a pure-death process starting at level $s + k + 1$, ignoring all future arrivals. The times until successive deaths in the pure-death process are exponential with state-dependent parameters.

Let γ_k be the probability that the kth customer in line abandons in the next departure event and let m_k be the mean time to the next departure event, in both

cases considering only the first $s + k$ customers in the system; i.e.,

$$\gamma_k = \frac{\delta'_k}{s\mu + \delta_k} \quad \text{and} \quad m_k = \frac{1}{s\mu + \delta_k}. \tag{3.12}$$

Then the probability that customer $s + k$ eventually receives service is

$$\Gamma_k = (1 - \gamma_k)(1 - \gamma_{k-1}) \cdots (1 - \gamma_1) \tag{3.13}$$

for γ_k in (3.12). Then the probability that a new arrival eventually completes service, is

$$P(S) = \left(\sum_{k=0}^{s-1} p_k \right) + \sum_{k=0}^{r-1} p_{s+k}(1 - \beta) q_k \Gamma_{k+1}. \tag{3.14}$$

Let C be the response time. (We let C be 0 when service is not completed.) Then, using properties of the exponential distribution, we obtain

$$EC = \left(\sum_{k=0}^{s-1} p_k \right) \frac{1}{\mu} + \sum_{k=0}^{r-1} p_{s+k}(1 - \beta) q_k \Gamma_{k+1} \left(\frac{1}{\mu} + \sum_{j=1}^{k+1} m_j \right) \tag{3.15}$$

and

$$EC^2 = \left(\sum_{k=0}^{s-1} p_k \right) \frac{2}{\mu^2} + \sum_{k=0}^{r-1} p_{s+k}(1 - \beta) q_k \Gamma_{k+1} (V_{k+1} + M_{k+1}^2) \tag{3.16}$$

where

$$V_{k+1} = \frac{1}{\mu^2} + \sum_{j=1}^{k+1} m_j^2 \tag{3.17}$$

and

$$M_{k+1} = \frac{1}{\mu} + \sum_{j=1}^{k+1} m_j. \tag{3.18}$$

Then the first and second moments of the conditional time to complete service given that service is completed are

$$E(C|S) = EC/P(S) \quad \text{and} \quad E(C^2|S) = EC^2/P(S). \tag{3.19}$$

The conditional variance and standard deviation are then

$$\text{Var}(C|S) = E(C^2|S) - (E(C|S))^2 \tag{3.20}$$

and

$$\text{SD}(C|S) = \sqrt{\text{Var}(C|S)}. \tag{3.21}$$

Now let $\hat{c}(z) \equiv Ee^{-zC}$ be the Laplace transform of C (Laplace-Stieltjes Transform of its cdf). Paralleling (3.15), we have

$$(z) = \left(\sum_{k=0}^{s-1} p_k \right) \left(\frac{\mu}{\mu + z} \right) + \sum_{k=0}^{r-1} p_{s+k}(1 - \beta) q_k \Gamma_{k+1} \hat{d}_{k+1}(z), \tag{3.22}$$

where

$$\hat{d}_{k+1}(z) = \left(\frac{\mu}{\mu + z} \right) \prod_{j=1}^{k+1} \left(\frac{m_j^{-1}}{m_j^{-1} + z} \right). \tag{3.23}$$

We can now easily calculate $P(C > t)$ for any desired t by numerically inverting its Laplace transform $(1 - \hat{c}(z))/z$, e.g., by using the Fourier-series method described in Abate and Whitt (1995). The associated conditional response-time distribution is

$$P(C > t|S) = P(C > t)/P(S). \tag{3.24}$$

Let R be the event that an arrival eventually reneges and let A be the time to renege. (Let $A = 0$ when the customer does not renege.) Let A_k be the time to abandon for a customer who starts in position k in queue. Then, by essentially the same reasoning,

$$P(R) = \sum_{k=0}^{r-1} p_{s+k}(1 - \beta) q_k (1 - \Gamma_{k+1}), \tag{3.25}$$

$$EA = \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k EA_{k+1}, \tag{3.26}$$

and

$$EA^2 = \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k EA_{k+1}^2, \tag{3.27}$$

where

$$\begin{aligned} EA_k = {} & \gamma_k m_k + (1-\gamma_k)\gamma_{k-1}(m_k + m_{k-1}) \\ & + (1-\gamma_k)(1-\gamma_{k-1})\gamma_{k-2}(m_k + m_{k-1} + m_{k-2}) \\ & + \ldots + (1-\gamma_k)\ldots(1-\gamma_2)\gamma_1(m_k + \ldots + m_1) \end{aligned} \tag{3.28}$$

and

$$\begin{aligned} EA_k^2 = {} & \gamma_k 2m_k^2 \\ & + (1-\gamma_k)\gamma_{k-1}(m_k^2 + m_{k-1}^2 + (m_k + m_{k-1})^2) \\ & + \ldots + (1-\gamma_k)(1-\gamma_{k-1})\ldots(1-\gamma_2) \\ & \times \gamma_1(m_k^2 \ldots + m_1^2 + (m_k + \ldots + m_1)^2). \end{aligned} \tag{3.29}$$

The associated conditional moments are

$$E(A|R) = EA/P(R) \quad \text{and} \quad E(A^2|R) = EA^2/P(R). \tag{3.30}$$

for $P(R)$ in (3.25). Finally, the conditional variance and standard deviation are

$$\text{Var}(A|R) = E(A^2|R) - (E(A|R))^2 \tag{3.31}$$

and

$$\text{SD}(A|R) = \sqrt{\text{Var}(A|R)}. \tag{3.32}$$

Now let $\hat{a}(z) \equiv Ee^{-zA}$ be the Laplace transform of A. Paralleling (3.26), we have

$$\hat{a}(z) = \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k \hat{a}_{k+1}(z), \tag{3.33}$$

where

$$\begin{aligned} \hat{a}_k(z) = {} & \left(\frac{m_k^{-1}}{m_k^{-1}+z}\right) \sum_{j=0}^{k-1} \gamma_{k-j} \prod_{l=1}^{j} \\ & \cdot \left[(1-\gamma_{k-l+1})\left(\frac{m_{k-l}^{-1}}{m_{k-l}^{-1}+z}\right)\right]. \end{aligned} \tag{3.34}$$

Paralleling $P(C > t)$ above, we can compute $P(A > t)$ by numerically inverting its Laplace transform $(1 - \hat{a}(z))/z$. Then the conditional distribution of the time to renege given reneging is

$$P(A > t|R) = P(A > t)/P(R). \tag{3.35}$$

Finally, the probability of blocking is p_{s+r}, so that the probability of balking is

$$P(\text{balking}) = 1 - P(S) - P(R) - p_{s+r}. \tag{3.36}$$

4. Stochastic Comparisons

The consequences of informing customers about anticipated delays are not entirely clear. We believe that customers should prefer this additional information and that the greatest benefit will stem from improved customer satisfaction. The improved customer satisfaction should in turn benefit the service provider by producing an increased arrival rate; better service should mean more business. Alternatively, those service providers that displease customers may experience decreasing arrival rate and ultimately go out of business. An increased arrival rate may mean that the service provider should increase the number of servers. Overall, the response is complicated, depending on the nature of the service and the competition.

We do not attempt to predict how the basic parameters λ and s will change in response to improved customer satisfaction. Instead, to help place that issue in perspective, in this section we make comparisons assuming that the basic parameters remain unchanged. In particular, we consider Models 1 and 2 with λ, μ, s, r, β and α fixed.

Intuitively, it seems that balking upon arrival instead of joining the queue and later reneging should lead to fewer customers in the system, provided that the chance of balking relates appropriately to the

chance of reneging, as in the construction in §3, in particular, assuming (3.2). We now show that a strong comparison is possible. In particular, we establish *likelihood ratio* (LR) ordering. See Chapter 1 of Shaked and Shanthikumar (1994) for background on stochastic orderings.

Consider two random variables X_1 and X_2 with values in the state space $\{0, 1, \ldots, n\}$, $1 \le n \le \infty$, that have probability mass functions (pmf's) that are positive for all states. We say that X_1 is less than or equal to X_2 in the *likelihood ratio* (LR) ordering and write $X_1 \le_{lr} X_2$ if

$$\frac{P(X_1 = k+1)}{P(X_1 = k)} \le \frac{P(X_2 = k+1)}{P(X_2 = k)}, \quad 0 \le k \le n-1. \tag{4.1}$$

We say that X_1 is *stochastically less than or equal to* X_2 and write $X_1 \le_{st} X_2$ if

$$P(X_1 \ge k) \le P(X_2 \ge k), \quad 0 \le k \le n. \tag{4.2}$$

The LR order implies stochastic order. Indeed, the LR order is equivalent to stochastic order holding under conditioning for all intervals; i.e., $X_1 \le_{lr} X_2$ if and only if

$$(X_1 | a \le X_1 \le b) \le_{st} (X_2 | a \le X_2 \le b) \tag{4.3}$$

for all a and b with $a < b$; see p. 29 of Shaked and Shanthikumar (1994).

We now present a sufficient condition for the steady-state distributions of BD processes to be ordered in the LR ordering. This result is a special case of Theorem 5 of Smith and Whitt (1981) (which applies to more general processes).

THEOREM 4.1. *Consider two BD processes with common state space* $\{0, 1, \ldots, n\}$, *birth rates* $\lambda_k^{(i)}$, *death rates* $\mu_k^{(i)}$, *and steady-state random variables* N_i, $i = 1, 2$. *If*

$$\frac{\lambda_k^{(1)}}{\mu_{k+1}^{(1)}} \ge \frac{\lambda_k^{(2)}}{\mu_{k+1}^{(2)}} \quad \text{for } 0 \le k \le n-1, \tag{4.4}$$

then

$$N_1 \ge_{lr} N_2.$$

We now compare the processes in Models 1 and 2. Recall that Model 2 is the special case of the general BD model in §3 with no reneging; i.e., with $\delta_k = 0$. Let $\lambda_k^{(i)}$, $\mu_k^{(i)}$, N_i and θ_i denote the birth rates, death rates, steady-state number of customers present and throughput in Model i.

THEOREM 4.2. *Consider the BD processes in Models* 1 *and* 2 *with common parameters* λ, μ, α, β, s *and* r, *using* (3.2). *Then*

$$N_1 \ge_{lr} N_2 \quad \text{and} \quad \theta_1 \ge \theta_2.$$

PROOF. By Theorem 4.1, it suffices to establish (4.4). For $0 \le k \le s-1$, $\lambda_k^{(1)} = \lambda_k^{(2)} = \lambda$ and $\mu_{k+1}^{(1)} = \mu_{k+1}^{(2)} = (k+1)\mu$. For $k \ge 0$,

$$\frac{\lambda_{s+k}^{(1)}}{\mu_{s+k+1}^{(1)}} = \frac{\lambda(1-\beta)}{s\mu + (k+1)\alpha}, \quad \frac{\lambda_{s+k}^{(2)}}{\mu_{s+k+1}^{(2)}} = \frac{\lambda(1-\beta)q_k}{s\mu},$$

so that it suffices to show that

$$\frac{\lambda_{s+k}^{(2)}}{\lambda_{s+k}^{(1)}} = q_k \equiv \left(\frac{s\mu}{s\mu + \alpha}\right)^{k+1} \le \frac{s\mu}{s\mu + (k+1)\alpha} = \frac{\mu_{s+k+1}^{(2)}}{\mu_{s+k+1}^{(1)}} \tag{4.5}$$

for $0 \le k \le r-1$. However, (4.5) holds because, by the binomial theorem, $(1+x)^k \ge 1 + kx$ for all $x > 0$ and all positive integers k. The ordering for θ_i follows since

$$\theta_i = \sum_{k=1}^{s+r} P(N_i = k)(k \wedge s). \quad \square \tag{4.6}$$

Model 1 produces higher throughput but at the expense of having some customers wait without eventually receiving service. As a consequence of the stochastic order $N_1 \ge_{st} N_2$, Model 1 has higher blocking. Also, Model 2 has a higher probability of a customer receiving service without having to wait. Having higher throughput might make Model 1 preferable to Model 2 for the service provider, but recall that the arrival rates are assumed equal in Theorem 4.2. From the perspective of throughput, Theorem 4.2 shows that the benefit to the service provider from Model 2 (informing customers about anticipated delays) must stem from a subsequent increase in the arrival rate λ (or other basic parameter change).

The stochastic comparison we have made between the two models in Theorem 4.2 assumes that the basic

parameter tuple $(\lambda, \mu, \alpha, \beta, s, r)$ is the same for both models. However, if we change the way the system operates, then these parameters may change too, leading to more complex comparisons. We can describe how each system separately responds to changes in the parameters, though. For simplicity, let $\delta_k = k\delta'$ in §3, then the model there depends on the parameter tuple $(\lambda, \mu, \alpha, \beta, \delta', s, r)$.

THEOREM 4.3. *Consider the general BD model in §3 with two candidate parameter tuples* $(\lambda^{(i)}, \mu^{(i)}, \alpha^{(i)}, \beta^{(i)}, \delta'^{(i)}, s, r)$, $i = 1, 2$. *If* $\lambda^{(1)} \le \lambda^{(2)}$, $\mu^{(1)} \ge \mu^{(2)}$, $\alpha^{(1)} \ge \alpha^{(2)}$, $\beta^{(1)} \ge \beta^{(2)}$ *and* $\delta'^{(i)} \ge \delta'^{(2)}$, *then*

$$N^{(1)} \le_{lr} N^{(2)}.$$

PROOF. It is easy to see that $\lambda_k^{(1)} \le \lambda_k^{(2)}$ and $\mu_{k+1}^{(1)} \ge \mu_{k+1}^{(2)}$ for all k, $0 \le k \le s + r - 1$, so that (4.4) holds. Hence, we can apply Theorem 4.1. □

From Theorem 4.3 it is not evident how the long-run balking and reneging rates respond to increases in the parameters α, β and δ. It is intuitively clear that the long-run rates should increase, but if we increase β, then the steady-state distribution decreases, so that there is less opportunity for balking. Nevertheless, we can establish the desired comparison by exploiting a sample-path comparison.

THEOREM 4.4. *Consider Models* 1 *and* 2.

(a) *If* α *increases, then the long-run service-completion rate decreases and the long-run reneging rate* (*Model* 1) *or balking rate* (*Model* 2) *increases.*

(b) *If* β *increases, then the long-run service-completion rate decreases and the long-run balking rate increases.*

(c) *For Model* 2, *if* δ' *increases, then the long-run service-completion rate decreases and the long-run reneging rate increases.*

(d) *Under condition* (a), (b) *or* (c), *the steady-state distribution of* N *decreases in stochastic order.*

PROOF. We only consider part (a) for Model 1, because the reasoning is the same in the other cases. Let $N^{(i)}(t)$ be the number of customers in system i as a function of time. As in Whitt (1981), it is possible to construct the two systems on the same sample space so that the sample paths of $N^{(1)}(t)$ and $N^{(2)}(t)$ are ordered (a coupling). Let the two systems be indexed by i, where $\alpha^{(1)} < \alpha^{(2)}$. Let the two systems both start out empty. We can generate all events from a common Poisson process with a constant rate $\gamma \equiv \lambda + s\mu + r\alpha^{(2)}$. Then we determine the nature of the events according to the birth and death rates. For example, if the state is $k < s$, then with probability $k\mu/\gamma$, the event is a service completion, with probability λ/γ, the event is an external arrival; while with probability $(\gamma - \lambda - k\mu/\gamma$ the event is a fictitious event, leading to no state change. Whenever the two sample paths coincide with $s + k$ customers present for $k \ge 1$, let service completions be the same in both systems and let there be reneging in the system with parameter $\alpha^{(2)}$, where $\alpha^{(2)} > \alpha^{(1)}$, whenever there is reneging in the system with parameter $\alpha^{(1)}$. However, there may be additional reneging in system 2, making $N^{(2)}(t) \le N^{(1)}(t)$. Whenever $N^{(2)}(t) \le N^{(1)}(t)$, the service completion rates are greater for system 1. Hence, let there be a service completion in system 1 whenever there is one in system 2. This allows extra service completions in system 1. Also, let there be a balking or blocking event in system 1 whenever there is a balking event in system 2. With this construction, a gap $N^{(1)}(t) - N^{(2)}(t)$ can only be created and grow by excess reneging in system 2. This gap may be reduced in several ways, including by subsequent reneging in system 1, but the cumulative number of customers reneging always stays ahead for system 2. Whenever the gap closes to 0, couple $N^{(1)}(t)$ and $N^{(2)}(t)$ again so that the ordering $N^{(1)}(t) \ge N^{(2)}(t)$ and the ordering on cumulative numbers of customers to have reneged are maintained. Since the sample paths are ordered $N^{(1)}(t) \ge N^{(2)}(t)$ for all t with this special construction, first the finite-dimensional distributions and, second, the steady-state distributions are stochastically ordered. □

REMARK. For Model 1, the proof of Theorem 4.4 shows that the long-run service-completion and balking rates both decrease when α increases. When α and β both increase, we can deduce that the long-run service-completion rate decreases, but not how the long-run reneging and balking rates are affected, because customer loss can take several forms, namely, blocking, balking, or reneging.

5. Numerical Examples

We now illustrate how the BD models can be used by considering a few numerical examples. In Theorem 4.2, we established an ordering between Models 1 and 2 with common parameter tuples $(\lambda, \mu, \alpha, \beta, s, r)$. However in numerical examples we have found that in many respects the two systems with common parameter tuples behave very similarly. The main difference is that, for Model 1, some customers who do not eventually receive service spend time waiting before reneging. This wasted customer effort is eliminated by predicting delays, if the prediction leads to Model 2. Throughout this section we use Definition (3.2).

EXAMPLE 5.1. Economies of Scale. In addition to comparing Models 1 and 2 with common parameter tuples, our first example illustrates the economies of scale. In particular, we consider both systems with $s = 4 \times 10^k$ for $k = 0, 1, 2$ and 3. In each case, we let $\lambda = s$, $\mu = 1.0$, $\alpha = 1.0$ and $\beta = 0.2$. We choose r to be sufficiently large so that blocking is negligible. With this parameter choice, the system with $s = 4 \times 10^k$ corresponds to the combination of 10 identical systems with $s = 4 \times 10^{k-1}$. We have resource sharing in the sense of Smith and Whitt (1981).

Numerical results for these cases are presented in Table 1. Since $(x)^+ = \max\{x, 0\}$, the expression $E(N - s)^+$ in Table 1 represents the expected number of customers waiting. Table 1 shows that the two systems do not differ much, with the difference decreasing as s increases. In all cases, the values of the probability that an arrival is eventually served are very close for the two systems. Table 1 also shows that all measures of performance improve as s increases, thus quantifying the economies of scale.

It is interesting to contrast Models 1 and 2 with the pure-loss model, which otherwise has the same parameters. The probability of eventually being served in the associated $M/M/s/0$ loss model is 0.639, 0.884, 0.961 and 0.9875 for $s = 4 \times 10^k$ and $k = 0, 1, 2$ and 3. The difference is substantial for smaller s, but negligible for larger s. For larger s, the balking acts like blocking. When $\lambda = 4000$ and $\beta = 0.2$, the arrival rate drops to 3200 when all servers are busy. In that case, s acts much like an upper barrier. Indeed, in that case, the conditional mean queue length given all servers are busy is only $E(N - s)^+ / P(N \geq s) = 3.95$.

Table 1 **A Comparison Between Models 1 and 2 as a Function of System Size, $s = 4 \times 10^k$ for $k = 1, 2, 3$ and 4.**

Performance Measures	$s = 4$		$s = 40$	
	Model 1	Model 2	Model 1	Model 2
$P(N \geq s)$	0.501	0.493	0.335	0.333
$E(N - s)^+$	0.498	0.445	0.816	0.796
EN	3.60	3.53	37.3	37.3
$SD(N)$	1.74	1.67	4.86	4.83
P(renege)	0.124	0	0.020	0
P(served)	0.775	0.772	0.913	0.912
$E(C\|S)$	1.115	1.144	1.021	1.022
$SD(C\|S)$	1.026	1.046	1.001	1.001
$E(A\|R)$	0.282	—	0.069	—
	$s = 400$		$s = 4000$	
	Model 1	Model 2	Model 1	Model 2
$P(N \geq s)$	0.162	0.162	0.0593	0.0593
$E(N - s)^+$	0.589	0.589	0.234	0.234
EN	387.0	387.0	3953.	3953.
$SD(N)$	13.1	13.1	38.9	38.9
P(renege)	0.0015	0	0.00006	0
P(served)	0.966	0.966	0.9881	0.9881
$E(C\|S)$	1.0015	1.0015	1.0000	1.0000
$SD(C\|S)$	1.0000	1.0000	1.0000	1.0000
$E(A\|R)$	0.110	—	0.0012	—

Note: In all cases $\lambda = s$, $\mu = \alpha = 1$ and $\beta = 0.2$. The variable N is the steady-state number of customers in the system, C is the time to complete service and A is the time to abandon.

EXAMPLE 5.2. Heavy Loads. Models 1 and 2 do not differ when all servers are not busy. Thus, the difference should increase as the load increases. We next illustrate the larger differences that are possible with higher loads. For this example, we let $s = 10$, $\mu = 1.0$, $\alpha = 1.0$ and $r = 50$. We consider two cases: In the first case, we let $\lambda = 20$ and $\beta = 0.2$; in the second case we let $\lambda = 40$ and $\beta = 0.5$. With the conventional definition of traffic intensity $\rho \equiv \lambda/s\mu$, $\rho = 2.0$ and 4.0 in the two cases. Numerical results for these two cases are displayed in Table 2.

In Table 2 the differences between Models 1 and 2 are greater than in Table 1, but still not large. The

Table 2 **A Comparison Between Models 1 and 2 Under Heavy Loadings**

Performance Measures	$\lambda = 20, \mu = 1.0, \alpha = 1.0, \beta = 0.2, s = 10, r = 50$		$\lambda = 40, \mu = 1.0, \alpha = 1.0, \beta = 0.5, s = 10, r = 50$	
	Model 1	Model 2	Model 1	Model 2
$P(N \geq s)$	0.970	0.958	0.9981	0.9955
$E(N - s)^+$	6.17	4.66	10.04	6.84
EN	16.1	14.6	20.0	16.8
$SD(N)$	3.90	3.09	4.43	3.16
P(renege)	0.308	0	0.251	0
P(served)	0.498	0.497	0.250	0.250
$E(C\|S)$	1.44	1.47	1.65	1.68
$SD(C\|S)$	1.04	1.065	1.045	1.080
$E(A\|R)$	0.293	—	0.356	—

probability of being eventually served and the mean and standard deviation of the conditional time to be served are very close. The greatest differences are in EN and $SD(N)$, the mean and standard deviation of the steady-state number of customers in the system.

Note that the most serious detrimental effect of the heavy loads is the low proportion of customers served. The delays experienced by those customers served are not especially large. These results show that a focus on the delays experienced by served customers, while ignoring the customers lost to balking or reneging, can seriously overestimate the quality of service provided.

Also note that the performance is quite different from the $M/M/s/r$ model without balking or reneging. Then the steady-state number N is close to $s + r$, which in Table 2 would be 60. The probability of being served is about the same, however. In the setting of Table 2 the blocking is negligible. The high blocking in $M/M/s/r$ is replaced by balking and reneging in these cases. □

At first glance, it might be thought that in the setting of §2 the reneging rate α might be reasonably well estimated by the reciprocal $E(A|R)$, the expected time to renege given that reneging occurs. However, it can be much less. Note that we consistently have

$$E(A|R) \leq \alpha^{-1}. \tag{5.1}$$

This must occur because the sequence of reneging times is censored. (Many customers are served before they have a chance to renege.)

6. Estimating the Balking Parameters

In this section we consider how to estimate the balking parameters α and β in §3 assuming (3.2), and how to validate the model. For background on standard procedures for estimating parameters in BD models, see Basawa and Prakasa Rao (1980) and references cited there.

For $0 \leq k \leq r - 1$, let $A_k(t)$ be the number of arrivals finding $s + k$ customers in the system upon arrival and let $J_k(t)$ be the number of these arrivals to join the queue in an operation of the system over a time interval $[0, t]$. (The number balking is thus $A_k(t) - J_k(t)$.) Under the BD model assumptions, it is possible to establish laws of large numbers and central limit theorems for these estimators. As a consequence, as $t \to \infty$, the ratio will converge as the sampling period grows, i.e.,

$$R_k(t) \equiv \frac{J_k(t)}{A_k(t)} \to \eta \equiv (1 - \beta)\left(\frac{s\mu}{s\mu + \alpha}\right)^{k+1} \quad \text{as } t \to \infty, \tag{6.1}$$

so that

$$-\log R_{k-1}(t) \to -\log(1 - \beta) - k \log\left(\frac{s\mu}{s\mu + \alpha}\right) \quad \text{as } t \to \infty. \tag{6.2}$$

Moreover, under the model assumptions, conditional on $A_k(t)$, $J_k(t)$ has a binomial distribution with parameters $n = A_k(t)$ and $p = \eta$ in (6.1). Hence, we propose estimating the parameters α and β by performing a linear regression with the variables $-\log R_{k-1}(t)$, $k \geq 1$, i.e., we find the best linear fit

$$-\log R_{k-1}(t) = \hat{a}_1 + \hat{a}_2 k. \tag{6.3}$$

We then estimate α and β by $\hat{\alpha}$ and $\hat{\beta}$, where

$$-\log(1 - \hat{\beta}) = \hat{a}_1 \quad \text{and} \quad -\log\left(\frac{s\mu}{s\mu + \alpha}\right) = \hat{a}_2, \tag{6.4}$$

so that

$$\hat{\beta} = 1 - e^{-\hat{a}_1} \tag{6.5}$$

and

$$\hat{\alpha} = s\mu(e^{\hat{a}_2} - 1). \tag{6.6}$$

By (6.2), these estimators of α and β are consistent (converge as $t \to \infty$). The degree to which a linear fit in (6.3) is appropriate also indicates the quality of the model fit. If we were to assume (3.3) instead of (3.2), then we would obtain $\hat{\alpha} = s\mu\hat{a}_2$ instead of (6.6).

When the fit in (6.3) is not good, we should question whether (3.2) and (6.1) hold, i.e., whether T has an exponential cdf. Instead of (3.1) or (3.3) we could work with the probability $q_k^* \equiv P(T > ES_k) \equiv 1 - H(ES_k)$, where H is the (now general) cdf of T, and then estimate it by

$$\hat{q}_k^* = R_k(t), \quad 0 \le k \le r - 1. \tag{6.7}$$

A disadvantage of (6.7) for prediction is that it yields r parameters instead of only 2. However, from (6.7) we obtain an estimate of the cdf H at r points, because $q_k^* = H((k + 1)/s\mu)$, $0 \le k \le r - 1$.

However, if the cdf of T is not nearly exponential, then the overall BD model is not valid. Another approach is to test for the BD model and estimate the parameters λ_k and μ_k directly. These estimates in turn can be used to estimate the parameters β, q_k and δ_k in (3.6) and (3.7). The successive holding times in state k should be i.i.d. exponential random variables with mean $(\lambda_k + \mu_k)^{-1}$. The successive transitions from level k should be i.i.d. Bernoulli random variables, up 1 to level $k + 1$ with probability $\lambda_k/(\lambda_k + \mu_k)$ and down 1 to level $k - 1$ with probability $\mu_k/(\lambda_k + \mu_k)$.

It is also natural to consider other two-parameter or three-parameter models for nonbalking. For example, instead of $(1 - \beta)\gamma^{k+1}$ in (6.1), we might consider $(1 - \beta)(k + 1)^{-\gamma}$. If we do estimate the balking probabilities in state $s + k$ for each k, then it is natural to impose a monotonicity condition, exploiting the condition that the balking probability should be increasing in k. See Barlow et al. (1972) for appropriate statistical methods.

7. Estimating the Reneging Rate

In this section we consider how to estimate the reneging rate α in §2 or δ' (assuming $\delta_k = k\delta'$) in §3. As noted at the end of §5, the average conditional time to abandon $E(A|R)$ for the model in §2 is often substantially less than α^{-1}, the reciprocal of the reneging rate. As an estimator $\hat{\alpha}$ for α, we propose that value of α, with the other elements of the parameter tuple $(\lambda, \mu, \alpha, \beta, s, r)$ that yields the observed estimate for the mean $E(A|R)$; i.e., we directly estimate $E(A|R)$ by looking at the sample mean of the reneging times and then we apply the BD model to find that value of α that yields the estimate. The most important point is not to confuse $E(A|R)$ with α^{-1}.

By Theorem 4.4, the long-run reneging rate is always increasing in α, so that the search is not difficult to perform, e.g., by bisection search. This estimation procedure can also be used when there is reneging even when delays are predicted.

When the service provider announces delay predictions to each arrival, it is possible that the reneging behavior depends on the initial state. To confirm the delay predictions and to understand the reneging behavior, it is good to monitor the outcomes starting with each initial state $s + k$ for $k \ge 0$. Reneging events well before the anticipated waiting time $(k + 1)/s\mu$ represent an unwillingness to wait for the predicted time. Reneging events after the anticipated waiting time $(k + 1)/s\mu$ represent a failure to accurately predict the delay and associated customer dissatisfaction.

8. Coping with Other Model Deviations

We conclude by briefly discussing possible deviations from the basic BD model and how they might be coped with. Serious investigations of these procedures represent topics for future research.

Time Dependence. Perhaps the most common difficulty is that the arrival process can be nonstationary. In many applications a reasonable model for the arrival process is a nonhomogeneous Poisson process with deterministic arrival-rate function $\lambda(t)$ that varies over time; e.g., see Chapter 6 of Hall (1991). The

service-time distribution may be time-dependent as well. One approach to this complication is to apply numerical methods to solve the time-dependent BD process, obtained by working with $\lambda(t)$ and $\mu(t)$ instead of λ and μ. A specific algorithm based on a discrete-time approximation is given in Davis et al. (1995). References are also cited there to sources applying the related Runge-Kutta methods to numerically solve the ordinary differential equations.

A simple approximation for the time-dependent distribution of the time-dependent BD process is the pointwise stationary approximation (PSA), which is the steady-state distribution of the BD process calculated in terms of the arrival-rate and service-rate functions $\lambda(t)$ and $\mu(t)$ as a function of time t. If $\lambda(t)$ varies significantly over time, then the PSA is often a far better description than the BD model with the long-run average arrival and service rates; e.g., see Green and Kolesar (1991). The PSA is also asymptotically correct as the arrival and service rates increase which corresponds to the rates changing more slowly; see Whitt (1991). In other words, the steady-state analysis here is directly applicable as a reasonable approximation when the arrival and service rates fluctuate if it is applied over suitable subintervals over which these functions do not change much. The estimated rates are then averages over these subintervals.

A complication when time-dependence is recognized is that it becomes necessary to estimate the functions $\lambda(t)$ and $\mu(t)$ instead of the single parameters λ and μ. Appropriate data smoothing is thus often required.

Nonexponential Service-Time Distributions. Nonexponential service-time distributions will tend to invalidate the BD model predictions. The congestion is likely to be greater (less) if the service-time distribution is more (less) variable than exponential. The impact of a nonexponential service-time distribution can be at least roughly estimated by examining its impact on the related $M/G/s/\infty$ pure-delay model; e.g., see Whitt (1993) and references cited there for simple approximations.

The impact of a nonexponential service-time distribution should be negligible if the arrival process is Poisson and the probability that all servers are busy is small, because the $M/G/s/0$ and $M/G/\infty$ models have the insensitivity property. However, the steady-state behavior conditional on all servers being busy should be significantly affected by the service-time distribution beyond its mean.

Below we propose a way to study the impact of a few exceptionally long service times. If the service-time distribution can be regarded as approximately exponential after removing such exceptionally long service times, then the modified BD analysis should be successful.

Similarly, if there is an excess of customers with very short service times, then they could be ignored. The resulting lower arrival rate and higher mean service time of the remaining customers may yield more accurate descriptions, assuming a BD model based on the approximate exponential distribution.

Occasional Extra Long Service Times. The BD model requires that the service times have an exponential distribution. This property might fail badly because some customers have exceptionally long service times; i.e., the service-time distribution might have a heavy tail. Thus, we now propose some simple methods to describe the impact of occasional extra long service times. We are primarily concerned with modifications to the model in §3 to produce appropriate approximate modified performance predictions. Our idea is to represent the special service times as server vacations or server interruptions. Since these service times are postulated to be unusually long, we can consider them to occur in a longer time scale. Thus, it is natural to represent these service times as special high-priority customers that occasionally require servers. Moreover, since the special service times are unusually long, it should be reasonable, at least as a rough approximation, to first determine the distribution of the number of long-service-time customers and then reduce the number of servers available to the regular customers by this random number; i.e., we use the approach of nearly decomposable Markov chains, as in Courtois (1977).

Hence, we first model the long service times by an $M/G/\infty$ model. The steady-state number of servers occupied with these special customers thus has a Poisson distribution with mean equal to $m_L \equiv \lambda_L/\mu_L$,

where λ_L is the arrival rate and μ_L^{-1} is the mean of these special long service times. We are assuming that the total offered load of these special customers, m_L, is sufficiently small that the chance that all servers are busy serving only them is negligible. Because of the insensitivity of the $M/G/\infty$ model, the service-time distribution beyond the mean plays no role at this point.

We can then consider the original model, where the number of servers is random (but fixed for all time) having the value $s - N_L$, where N_L has a Poisson distribution with mean m_L. That is, we consider the BD model in §3, where the number of servers is $s - k$ for various values of k. (The arrival rate λ and mean service time μ^{-1} in this new model must be appropriately reduced to account for the removal of the especially long service times.) For each reasonably likely k, we compute the steady-state distributions for the BD model with $s - k$ servers. The upper limit might be the mean plus a few standard deviations, i.e., $k \leq m_L + 3\sqrt{m_L}$. The performance measures for the models with $s - k$ servers can then be averaged with regard to the Poisson probabilities of k servers being busy serving the long service times. However, it may be more revealing to look at the conditional performance measures for fixed k, given those k whose likelihood is considered sufficiently large. Tables and plots of *both* the probability of k servers being used by the long-service-time customers and the conditional performance measures for the remaining customers given $s - k$ servers, as a function of k, should provide useful insight.

Non-Poisson Arrival Processes. In many settings, the Poisson arrival process (possibly nonhomogeneous) is natural, representing the result of many different customers making independent decisions. However, if the Poisson property is not nearly realistic, then the BD predictions can be far off. Non-Poisson processes arise naturally when the arrival process is itself an overflow process from another group of servers.

One way to approximately cope with non-Poisson stationary arrival processes is to substitute time-dependence or state-dependence for the stochastic dependence in the actual arrival process. The use of time-dependence is to reverse the approximation procedure discussed in Massey and Whitt (1996). In our setting with balking and reneging, the time-dependent birth-and-death process may be substantially easier to analyze than the stationary model with a non-Poisson arrival process.

Alternatively, we can try to approximately represent stochastic variability by a state-dependent arrival rate. In particular, we could use the Bernoulli-Poisson-Pascal (BPP) model in which the arrival rate λ is replaced by the linear function $\lambda_k = \alpha + \beta k$ for $k \geq 0$; see Delbrouck (1981) and Choudhury et al. (1995). The less bursty binomial case corresponds to $\beta < 0$, while the more busty Pascal case corresponds to $\beta > 0$.

All these analytical approximations can be substantiated by computer simulation.[1]

[1] I thank Avishai Mandelbaum for helpful pointers to the literature and Rhonda Righter for helpful comments.

References

Abate, J., W. Whitt. 1995. Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Computing* 7 36–43.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, H. D. Brunk. 1972. *Statistical Inference Under Order Restrictions*. Wiley, New York.

Basawa, I. V., B. L. S. Prakasa Rao, 1980. *Statistical Inference for Stochastic Processes*. Academic Press, New York.

Boxma, O. J., P. R. de Waal, 1995. Multiserver queues with impatient customers. J. Labetoulle, J. W. Roberts, Eds., *Proceedings ITC 14*. North-Holland, Amsterdam. 743–756.

Choudhury, G. L., K. K. Leung, W. Whitt. 1995. An inversion algorithm to compute blocking probabilities in loss networks with state-dependent rates. *IEEE/ACM Trans. Networking* 3 585–601.

Courtois, P. J. 1977. *Decomposability*. Academic Press, New York.

Davis, J. L., W. A. Massey, W. Whitt. 1995. Sensitivity of the service-time distribution in the nonstationary Erlang loss model. *Management Sci.* 41 1107–1116.

Delbrouck, L. E. N. 1981. A unified approximate evaluation of congestion functions for smooth and peaky traffic. *IEEE Trans. Commun.* COM29, 85–91.

Falin, G. 1990. A survey of retrial queues. *Queueing Systems* 7 127–167.

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

Green, L., P. Kolesar. 1991. The pointwise stationary approximating for queues with nonstationary arrivals. *Management Sci.* 37 84–97.

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

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

Heyman, D. P., M. J. Sobel. 1982. *Stochastic Models in Operations Research*, Vol. I McGraw-Hill, New York.

Hui, M. K., D. K. Tse. 1986. 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. *Sloane Management Rev.* **32** 44–53.

Kelly, F. P. 1991. Loss networks. *Ann. Appl. Prob.* **1** 319–378.

Massey, W. A., W. Whitt. 1996. Stationary-process approximations for the nonstationary Erlang loss model. *Op. Res.* **44** 976–983.

Rappaport, D. M. 1996. Key role of integration in call centers. *Business Comm. Rev.* (July) 44–48.

Ross, K. W. 1995. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer, New York.

Shaked, M., J. G. Shanthikumar. 1994. *Stochastic Orders and Their Applications*. Academic Press, New York.

Smith, D. R., W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Tech. J.* **60** 39–55.

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

Whitt, W. 1981. Comparing counting processes and queues. *Adv. Appl. Prob.* **13** 207–220.

——. 1985. Blocking when service is required from several facilities simultaneously. *AT&T Tech. J.* **64** 1807–1856.

——. 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Sci.* **37** 307–314.

——. 1993. Approximations for the $GI/G/m$ queue. *Production Oper. Management*, **2** 114–161.

——. 1999. Predicting queueing delays. To appear in *Management Sci.*

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

Accepted by Linda V. Green; received November 6, 1997. This paper has been with the author 1½ months for 2 revisions.