Minimizing Delays in the $GI/G/1$ Queue

WARD WHITT

AT&T Bell Laboratories, Holmdel, New Jersey

(Received November 1981; accepted October 1983)

What service-time distribution minimizes delays in a $GI/G/1$ queue with given renewal arrival process and given mean service time? It is natural to conjecture that the deterministic service-time distribution with unit mass on the mean is optimal for all objective functions of the form $Ef(W)$ where f is a nondecreasing function of the steady-state delay W. However, we show that this conjecture is false. In fact, for hyperexponential interarrival-time distributions (mixtures of two exponential distributions), the service-time distribution minimizing the average delay maximizes the proportion of customers delayed.

SUPPOSE that we have the opportunity to choose the service-time distribution in a standard $GI/G/1$ queueing model with a given renewal arrival process and a given mean service time. What service-time distribution should we use to minimize delays? Since the mean service time is fixed, the issue is how the probability mass should be spread around the mean. Since variability usually causes greater congestion, we would expect the deterministic service-time distribution with unit mass on the mean to be optimal, and it often is. However, we have not yet specified the objective function. We might want to minimize the average delay or the proportion of customers that have been delayed by more than x for some $x \geq 0$. For example, there might be a penalty for each customer that is delayed, regardless of the length of the delay (the case $x = 0$). But does the specific objective function matter? It is natural to conjecture that the deterministic service-time distribution is optimal for all these objective functions. This would occur with stochastic order, i.e., if

$$Ef(W_D) \leq Ef(W_G) \tag{1}$$

for all nondecreasing real-valued functions f for which the expectations are defined, where W_G (W_D) is the equilibrium waiting-time with a general (deterministic) service-time distribution having the given mean.

This paper shows that the desired stochastic ordering of the steady-state waiting-time distributions in (1) *does not always hold.* Moreover, there are renewal arrival processes for which the service-time distribution

Subject classification: 696 minimizing delays.

0030-364X/84/3201-0041 $01.25

minimizing (maximizing) the average delay maximizes (minimizes) the proportion of customers delayed. In fact, this result is not a rare pathology. It occurs for all renewal processes with H_2 interarrival-time distribution (hyperexponential: mixture of two exponential distributions) and can be anticipated to occur for other highly variable "bursty" arrival processes. With such arrival processes, there is necessarily a tradeoff between average delay and proportion of customers delayed. However, limited experience indicates that the average delay is affected more seriously by this choice. It is well known that average delay in the $GI/G/1$ queue (and more general models) is minimized by the deterministic service-time distribution; see Rogozin [1966] and more recently Hajek [1981], Humblet [1982] and Stoyan [1983]; see Whitt [1980] for additional citations to the literature.

We actually consider more general questions. We look not only at the service-time distribution yielding the maximum or minimum value of the objective function, we consider comparisons of two $GI/G/1$ systems with any two service-time distributions or any two interarrival-time distributions. This approach yields many more comparisons. For example, we show that the probability of delay in an $H_2/E_k/1$ queue increases in k, where E_k is an Erlang distribution with mean independent of k. As k increases, the variance decreases, and as $k \to \infty$, E_k approaches the deterministic distribution.

In the comparison of two $GI/G/1$ queueing systems, it is well known that the waiting times W_n inherit basic stochastic order properties from the interarrival times u_n and service times v_n. In particular, if $Ef(-u_1^1) \leq Ef(-u_1^2)$ and $Ef(v_1^1) \leq Ef(v_1^2)$ for all nondecreasing (nondecreasing convex) real-valued functions, then $Ef(W_n^1) \leq Ef(W_n^2)$ for all n and the same set of functions (the superscript indexes the system); see Gaede [1965], Daley and Moran [1968], Stoyan and Stoyan [1969], Borovkov [1976, §24], Stoyan [1983, §5.2] and Whitt [1981a].

For the special systems $M/G/1$ and $GI/M/1$, the stochastic ordering based on convex functions for the interarrival times and service times implies the stochastic ordering based on all nondecreasing functions for the *stationary* waiting times. For $M/G/1$ systems with arrival rates λ_i, if $\lambda_1 \leq \lambda_2$ and $Ef(v_1^1) \leq Ef(v_1^2)$ for all convex f, then $Ef(W^1) \leq Ef(W^2)$ for all nondecreasing f. (W^i is the stationary waiting time in the ith system.) (Daley and Rolski [1983] have shown that it suffices to have $Ef(v_1^1) \leq Ef(v_1^2)$ for all nondecreasing convex f.) For $GI/M/1$ systems with service rates μ_i, if $Ee^{-su_1^1} \leq Ee^{-su_1^2}$ for all $s \geq 0$ and $\mu_1 \geq \mu_2$, then $Ef(W^1) \leq Ef(W^2)$ for all nondecreasing f. See Rolski [1972, 1976], Rolski and Stoyan [1976], Stoyan [1983; Theorem 5.2.3], and Harrison [1977]. Note that $Ee^{-su_1} \leq Ee^{-su_2}$ if $Ef(-u_1) \leq Ef(-u_2)$ for all nondecreasing convex f because e^{sx} is increasing and convex.

Here we show that these special properties of the $M/G/1$ and $GI/M/1$ systems do not hold for all $GI/G/1$ systems. It is not always possible to obtain the stronger stochastic ordering for the stationary waiting times from the convex stochastic ordering for the interarrival and service times. Avis [1977, p. 231] conjectured such an ordering for multiserver queues. Open problem 2 at the end of §5 in Stoyan [1983] also asks what $GI/G/1$ systems have this stronger stochastic ordering of the stationary waiting time distributions. We identify $GI/G/1$ systems that fail to have this property.

First, we look at the lower end of the waiting-time distribution, e.g., the probability of delay. Since convex ordering does hold for all $GI/G/1$ systems, lack of stochastic order is likely to show up at the lower end of the distribution rather than in the tail. Whitt [1981b] exploited this idea for similar counterexamples in s-server queues. Second, we use light-traffic asymptotics to see what is going on. We make $P(u_1^1 > v_1^1)$ near one so that the distribution of W^i is close to the distribution of $W_1^i = \max\{0, v_1^i - u_1^i\}$, then let v_1^1 (u_1^1) be more variable, but not stochastically smaller (larger) than, the corresponding variable v_1^2 (u_1^2). Examples based on this approach are given in Section 1.

Upon reflection, it is apparent that this phenomenon is not a rare pathology. In Section 2 we prove (Theorem 2) that the probability of delay decreases when the service-time distribution becomes more variable in all $H_2/G/1$ queues. On the other hand, we also show (Theorem 1) that the probability of delay increases when the service-time distribution becomes more variable in all $GE_2/G/1$ queues having interarrival-time distributions that are convolutions of two exponential distributions. These qualitative properties have important implications for approximations; e.g., Theorems 1 and 2 are consistent with an approximation for the probability of delay developed by Kraemer and Langenbach-Belz [1976].

We conclude in Section 3 with further discussion.

1. LIGHT-TRAFFIC ASYMPTOTICS

We begin by constructing examples in which we increase the variability of the interarrival times and service times while keeping their means fixed, and show that the stationary waiting-time distribution does not increase stochastically.

Example 1. We construct two $GI/G/1$ systems with $P(u_n^1 = 2) = P(u_n^2 = 2) = 1$, $P(v_n^1 = 0) = P(v_n^2 = 0) = 1 - \epsilon$, and $P(v_n^1 = 2.5) = 2P(v_n^2 = 2.4) = 2P(v_n^2 = 2.6) = \epsilon$, so that the interarrival times have the same distribution and $Ef(v_1^1) \leq Ef(v_1^2)$ for all convex f. To obtain our counterexample, we shall show that W^i is asymptotically equivalent to W_1^i as

$\epsilon \rightarrow 0$; see (7) below. For this purpose, let $C^i \equiv C^i(\epsilon, W_0)$ be the number of customers served in a busy period of the ith system as a function of the initial workload W_0 and ϵ:

$$C^i = \min\{j \geq 1: W_j^i = 0\}.$$

We use the regenerative structure to calculate $P(W^i > x)$; see Ross [1970, §5.4]:

$$P(W^i > x) = E \textstyle\sum_{k=1}^{C^i(\epsilon,0)} 1_{(W_k^i > x)}/EC^i(\epsilon, 0), \tag{2}$$

where $W_0^i = 0$ and 1_A is the indicator function of the set A.

Consider the case $i = 1$. Since

$$\begin{aligned} &P(C^1(\epsilon, 0) = 1,\ W_1^1 = 0) = 1 - \epsilon, \\ &P(C^1(\epsilon, 0) = 2,\ W_1^1 = 0.5,\ W_2^1 = 0) = \epsilon(1 - \epsilon), \\ \text{and}\quad &P(C^1(\epsilon, 0) > 2,\ W_1^1 = 0.5,\ W_2^1 = 1) = \epsilon^2, \end{aligned} \tag{3}$$

$$\begin{aligned} EC^1(\epsilon, 0) &= 1(1 - \epsilon) + 2\epsilon(1 - \epsilon) + [2 + EC^1(\epsilon, 1)]\epsilon^2 \\ &= 1 + \epsilon + 0(\epsilon^2); \end{aligned} \tag{4}$$

in (4) we have used the fact that $EC^1(\epsilon, 1) < \infty$ and $EC^1(\epsilon, 1)$ is decreasing in ϵ; in fact, $C^1(\epsilon, 1)$ is stochastically decreasing in ϵ by virtue of sample path orderings for $\{W_n^i\}$; see Theorem 4 of Whitt [1981a] and references cited in this paper. Also, from (3) we have

$$\begin{aligned} &E \textstyle\sum_{k=1}^{C^1(\epsilon,0)} 1_{\{W_k^1 > 0.45\}} - \epsilon(1 - \epsilon) \\ &\quad = E \textstyle\sum_{k=3}^{\max\{C^1(\epsilon,0),3\}} 1_{\{W_k^1 > 0.45\}} \leq \epsilon^2[2 + EC^1(\epsilon, 1)] = 0(\epsilon^2). \end{aligned} \tag{5}$$

From (2), (4) and (5), we have

$$P(W^1 > 0.45) = [\epsilon + 0(\epsilon^2)]/[1 + \epsilon + 0(\epsilon^2)] = \epsilon + 0(\epsilon^2). \tag{6}$$

Formula (6) can also be expressed as

$$\lim_{\epsilon \to 0} P(W^1 > x)/P(W_1^1 > x) = 1 \tag{7}$$

for x such that $P(W_1^1 > x) > 0$. By similar reasoning,

$$P(W^2 > 0.45) = [\epsilon/2 + 0(\epsilon^2)]/[1 + \epsilon + 0(\epsilon^2)] = \epsilon/2 + 0(\epsilon^2).$$

Hence, for sufficiently small ϵ, $Ef(W^1) > Ef(W^2)$ for the nondecreasing function $f(x) = 1_{(0.45,\infty)}(x)$.

Example 2. Now we let the two systems have identical service-time distributions and change the interarrival-time distributions. Let $P(u_n^1 = 2) = 1$, $P(u_n^2 = 1.9) = P(u_n^2 = 2.1) = 1/2$, $P(v_n^i = 0) = 1 - P(v_n^i = 2.5) = 1 - \epsilon$ $(i = 1, 2)$, $n \geq 1$. Now $Ef(u_n^1) \leq Ef(u_n^2)$ for all convex f and, by similar reasoning, $P(W^1 > 0.45) = \epsilon + 0(\epsilon^2)$ and $P(W^2 > 0.45) = \epsilon/2 + 0(\epsilon^2)$.

It is also easy to see that the standard stochastic ordering that holds for the stationary waiting times in $M/G/1$ and $GI/M/1$ systems need not hold for the transient waiting times under the same ordering of the interarrival times and service times.

Example 3. For $M/G/1$ systems, let $P(u_n{}^i > x) = e^{-x}$, $x \geq 0$, and $W_0{}^i = 0$, $i = 1, 2$. Let $v_n{}^i$ be as in Example 1. Then $Ef(W_1{}^1) \leq Ef(W_1{}^2)$ for all nondecreasing convex f by existing theory, but

$$P(W_1{}^1 > 1.4) = P(u_1{}^1 < 0.1) = 1 - e^{-0.1} = 0.0952 > 0.09063$$

$$= (½)(1 - e^{-0.2}) = (½)P(u_1{}^1 < 0.2) = P(W_1{}^2 > 1.4).$$

A similar counterexample holds for $W_1{}^i$ in $GI/M/1$ systems.

2. PROBABILITY OF DELAY IN $K_2/G/1$ QUEUES

The probability of delay (in fact, also the mean waiting time) has a very simple expression for K_2 interarrival-time distributions, which have a rational Laplace-Stieltjes transform with denominator or degree 2; see p. 329 of Cohen [1969]. In this special case, the probability of delay depends on a single real root, say δ, of an equation involving the transforms $\alpha(s)$ of the interarrival times and $\beta(s)$ of the service times, namely,

$$\beta(s) = \alpha(-s)^{-1}. \tag{8}$$

Cohen shows for $K_m/G/1$ queues that (8) has exactly m complex roots in the positive half plane, one of which is 0. Hence, for $m = 2$, there is a single positive real root. Moreover, the probability of delay satisfies

$$P(W > 0) = 1 - f(\alpha)/\delta, \tag{9}$$

where $f(\alpha)$ is a constant depending on the interarrival-time distribution. Hence, the qualitative effect of changing variability is easy to see. If the variability of the service-time distribution increases in the sense of the expected value of all convex functions, then $\beta(s)$ increases for all s. (This service-time assumption is a weaker sufficient condition for the ordering.) What happens to the root δ in response to this change in $\beta(s)$ depends on whether $\alpha(-s)^{-1}$ hits $\beta(s)$ from above or below at $s = \delta$. If $\beta(s) > \alpha(-s)^{-1}$ for all s, $0 < s < \delta$, then δ increases when $\beta(s)$ increases. Note that the derivatives of $\alpha(-s)^{-1}$ and $\beta(s)$ at 0 are $-Eu$ and $-Ev$, respectively, so the two functions are equal at 0 and $\alpha(-\mathrm{s})^{-1}$ starts off below $\beta(s)$. When $\alpha(-\mathrm{s})^{-1}$ is continuous (has no singularity), then everything is as expected, as is the case for the convolution of two exponential distributions, e.g., E_2. Let GE_2 denote the convolution of two exponential distributions, with possibly different parameters λ_1 and λ_2, having transform

$$\alpha(s) = (\lambda_1/(\lambda_1 + s))(\lambda_2/(\lambda_2 + s)).$$

THEOREM 1. *Consider two stable $GE_2/G/1$ queues with a common interarrival-time distribution. If $Ef(v_n^1) \leq Ef(v_n^2)$ for all convex real-valued f (or only if $\beta_1(s) < \beta_2(s)$ for all s), then $P(W^1 > 0) \leq P(W^2 > 0)$.*

Proof. As noted above, $\alpha(-s)^{-1}$ is initially less than $\beta(s)$. Moreover, in the case of the convolution of two exponentials,

$$\alpha(-s)^{-1} = [(\lambda_1 - s)(\lambda_2 - s)]/\lambda_1\lambda_2, \tag{10}$$

so that $\alpha(-s)^{-1}$ is continuous. Hence, $\alpha(-s)^{-1}$ must hit $\beta(s)$ from below, so that δ increases as $\beta(s)$ increases.

Here is a numerical example:

Example 4. The probability of delay in $E_2/D/1$, $E_2/E_2/1$ and $E_2/M/1$ systems with $\rho = 0.5$ is, respectively, 0.323, 0.360, 0.382.

It is also possible to calculate the range of possible values for $P(W > 0)$ for $GE_2/G/1$.

COROLLARY 1. *The probability of delay in a stable $GE_2/G/1$ queue with fixed traffic intensity ρ has a supremum ρ and a minimum attained for $G = D$.*

Proof. Obviously $\beta(s)$ is minimized for all s over all service-time distributions with mean μ^{-1} by $e^{-s/\mu}$, so the minimum is clear. For the supremum, note that $\beta(s)$ can be made arbitrarily close to 1 for any s and any mean by choosing a distribution with almost all the mass at 0 and a small bit at a very large value. So, for the supremum, it suffices to apply (9) and (10) and solve $\alpha(-s)^{-1} = 1$, which yields $P(W > 0) = \rho$ in this case.

There is also another possibility for the behavior of $\alpha(-s)^{-1}$: it can have a singularity, at s_0 say, so that $\alpha(s)$ approaches $-\infty(+\infty)$ as s approaches s_0 from the left (right). In fact, for the $K_2/G/1$ queue, $\alpha(-s)^{-1}$ can have at most one such singularity (see Cohen), which is the full extent of the complexity. Now if $\delta > s_0$, then $\alpha(-s)^{-1}$ hits $\beta(s)$ from above instead of below at $s = \delta$, which will always be the case as $\beta(s)$ is decreased since $\alpha(-(s_0+))^{-1} = +\infty$ while $\beta(s_0) \leq 1$. This case applies for the $H_2/G/1$ queue.

THEOREM 2. *Consider two stable $H_2/G/1$ queues with common interarrival-time distributions. If $Ef(v_n^1) \leq Ef(v_n^2)$ for all convex real-valued f, then $P(W^1 > 0) \geq P(W^2 > 0)$.*

Proof. Here

$$\alpha(-s)^{-1} = [(\lambda_1 - s)(\lambda_2 - s)]/[\lambda_1\lambda_2 - s(p\lambda_1 + (1 - p)\lambda_2)], \tag{11}$$

so that there is a singularity at

$$s_0 = \lambda_1\lambda_2/(p\lambda_1 + (1-p)\lambda_2).$$

Hence, just to the right of s_0, $\alpha(-s)^{-1}$ is near $+\infty$. Also, for very large s, $\alpha(-s)^{-1}$ is negative, so that the single root δ must be larger than s_0. Consequently, $\alpha(-s)^{-1}$ hits $\beta(s)$ from above at δ.

Here is a numerical example.

Example 5. Consider an H_2 interarrival-time density

$$f(x) = q\lambda_1 e^{-\lambda_1 x} + (1-q)\lambda_2 e^{-\lambda_2 x},\ x \geq 0,$$

with balanced means $(q/\lambda_1 = (1-q)/\lambda_2)$, overall mean 2, and squared coefficient of variation (the variance divided by the square of the mean) $c_a^2 = 2$. Let the service rate be $\mu = 1$, so that $\rho = 0.5$. Then $P(W > 0) = 0.59175$ for $H_2/M/1$ and $P(W > 0) = 0.61395$ for $H_2/D/1$.

Paralleling Corollary 1, we have:

COROLLARY 2. *The probability of delay in stable $H_2/G/1$ queues with common interarrival-time distribution and common traffic intensity ρ has an infimum of ρ and a maximum attained for $G = D$.*

Proof. As in Corollary 1.

It is also interesting to vary the interarrival-time distribution instead of the service-time distribution. Whitt [1982b, 1984b] investigates such changes for $H_2/G/1$ queues. Now we can add

THEOREM 3. *The probability of delay in stable $H_2/G/1$ queues with common service-time distribution, common traffic intensity ρ, and common squared coefficient of variation c^2 of the interarrival-time distribution, has an infimum of ρ attained at the limiting $M/G/1$ system and a supremum of $1 - 2(1-\rho)/(1+c^2)$ attained at the $M^B/G/1$ system (having an arrival process that is batch Poisson with geometrically distributed batches having mean batch size $EB = (c^2+1)/2$).*

Proof. Theorem 1 of Whitt [1984b] treats the case of exponential service times. Apply Corollary 2 with the fact that the bounds in Theorem 3 depend on the service-time distribution only through its mean.

3. DISCUSSION

(1) Theorem 2 shows that Conjectures 5.1 and 5.2 of Avis are both false.

(2) Example 1 shows that the desired stochastic ordering of the steady-state waiting time is not true in $D/G/1$ systems. Since $E_k \to D$ as $k \to \infty$, by continuity (§11 of Borovkov), the desired stochastic order is not true for $E_k/G/1$ systems as well for sufficiently large k. We *conjecture* that

the steady-state delay distribution is stochastically decreasing in k and l in $E_k/E_l/1$ systems though. This conjecture is consistent with all available tables.

(3) Corollaries 1 and 2 can be extended to indicate the range of possible delay probability values given additional constraints on the service-time distribution such as additional moments. For example, if the service-time variance, σ^2, is given, too, then the maximum delay probability for $E_2/G/1$ and the minimum for $H_2/G/1$ is attained at the two-point service-time distribution with mass $\sigma^2/(\sigma^2 + \mu^{-2})$ on 0 and mass $\mu^{-2}/(\sigma^2 + \mu^{-2})$ on $\mu^{-1} + \mu\sigma^2$; see (12) of Eckberg [1977]. For the $H_2/G/1$ queue, the maximum mean delay given two moments of the service time is also attained at this two-point distribution, which contradicts a conjecture in Daley and Trengove [1977]; see open problem 5.2.4 of Stoyan [1983]. Hence, the singularity in (11) has other important consequences. Further discussion appears in §4 of Whitt [1983a].

(4) The situation for $K_m/G/1$ queues is more complicated. For $K_m/G/1$ queues,

$$P(W > 0) = f(\alpha)/\prod_{i=1}^{m-1} \delta_i$$

where $\prod_{i=1}^{m-1} \delta_i$ is the product of the $m - 1$ roots. For example, if $m = 3$, then there are two roots. In the $GE_3/G/1$ case, one root increases and the other decreases when the transform $\beta(s)$ increases. Hence, no simple ordering for the probability of delay can be expected. However, Daryl Daley has shown (personal communication) that all the roots behave the same way for $H_m/G/1$ queues, so that H_m results do hold for $m \geq 2$.

(5) For $GI/K_m/1$ queues, the probability of delay can also be obtained from the roots of a transform equation like (8) (see p. 321 of Cohen). However, for $m = 2$, there are two relevant nonzero roots, so the simple argument of §2 is not available.

(6) Decreasing probability of delay in response to greater variability in the interarrival times and service times is actually not so counterintuitive. The basic idea is that greater variability in the interarrival times and service times tends to make the waiting times both larger and more variable. Not only will the mean delay tend to increase, but also the greater variability will cause the tails of the steady-state distribution to be fatter. Since the lower tail is just the probability of no delay, the probability of no delay might increase in response to greater variability. The greater variability can offset the effect of the greater mean.

(7) A very interesting discussion of the effect of variability on queues is given by Wolff [1977]. For example, he notes that the effect of greater service-time variability in a heavily loaded $GI/G/\infty$ system depends on the squared coefficient of variation of the interarrival times, c_a^2. If $c_a^2 <$ 1 (less variable), then increased service-time variability means increased

variability for the number of busy servers. On the other hand, if $c_a^2 > 1$ (more variable), increased service-time variability means decreased variability for the number of busy servers. It is interesting that $c_a^2 = 1$ tends to be a boundary for the qualitative behavior in our development in §2, too.

In the $GI/G/\infty$ setting, under heavy load the number busy servers is approximately normally distributed with mean λ/μ (independent of variability) and variance-to-mean ratio (peakedness)

$$z = 1 + (c_a^2 - 1)y$$

where

$$y = \mu \int_0^\infty [P(v > x)]^2 dx;$$

see Whitt [1982a] and references he cites. The parameter y tends to increase with decreasing service-time variability, achieving its maximum for a given mean at D. Since the $GI/G/n$ system behaves like a $GI/G/\infty$ system for fixed ρ as $n \to \infty$, the number of customers in a $GI/G/n$ system for very large n will respond to increasing service-time variability in the same way as the number of busy servers in $GI/G/\infty$ systems. In particular, when there is more than one server, even the mean delay can decrease when the service-time variability increases. (Use Little's formula.) As in comment (6) above, this situation provides an example of the effect of changing variability in a congestion measure. The mean number of busy servers in a $GI/G/\infty$ system is always λ/μ. It is the variance that is changing in response to service-time variability.

(8) The results here are consistent with the approximation for the probability of delay in $GI/G/1$ queues given by Kraemer and Langenbach-Belz, namely,

$$P(W > 0) \approx \rho + (c_a^2 - 1)\rho(1 - \rho)g(c_a^2, c_s^2, \rho)$$

where

$$g(c_a^2, c_s^2, \rho) = \begin{cases} [1 + c_a^2 + \rho c_s^2]/[1 + \rho(c_s^2 + 1) + \rho^2(4c_a^2 + c_s^2)], & c_a \leq 1 \\ 4\rho/[c_a^2 + \rho^2(4c_a^2 + c_s^2)], & c_a \geq 1, \end{cases}$$

and c_a^2 and c_s^2 are the squared coefficients of variation of the interarrival times and service times, respectively. By differentiating in the case $c_a \leq 1$, we see that the KL approximate probability of delay is decreasing in c_s^2 for all c_a^2 satisfying

$$c_a^2 \geq [\sqrt{16(1 + \rho)\rho^2 + (1 - \rho)^2} - (1 - \rho)]/8\rho^2.$$

For $E_2/G/1$ the KL approximation says that the probability of delay should be increasing in c_s^2. This qualitative behavior appears in Table 2.8–4 of Kühn [1976] for the $GI/G/1$ queue with $c_a = 0.75$ (which is based on KL).

(9) The variability in the service times considered here is variability in the service-time distribution. For arrival processes, one can also consider variability of the arrival rate over time such as occurs in a nonstationary Poisson process. Comparison results in this setting have been obtained by Ross [1978] and Rolski [1981]. Heyman [1982] has given counterexamples in the spirit of the ones developed here.

ACKNOWLEDGMENTS

I am grateful to Shaler Stidham, Jr., for getting me started, Richard Stubing for writing programs to compute the probability of delay in $K_2/G/1$ queues, and the referee for helpful comments.

REFERENCES

AVIS, D. M. 1977. Computing Waiting Times in $GI/E_k/c$ Queueing Systems. In *Algorithmic Methods in Probability, TIMS Studies in Management Science* **7,** M. F. Neuts (ed.), pp. 215–232.

BOROVKOV, A. A. 1976. *Stochastic Processes in Queueing Theory.* Springer-Verlag, New York.

COHEN, J. W. 1969. *The Single Server Queue.* North-Holland, Amsterdam.

DALEY, D. J., AND P. A. P. MORAN. 1968. Two-Sided Inequalities for Waiting Time and Queue Size Distribution in $GI/G/1$. *Theor. Prob. Appl.* **13,** 338–341.

DALEY, D. J., AND T. ROLSKI. 1983. Some Comparability Results for Waiting Times in Single- and Many-Server Queues, Statistics Dept. (I. A. S.), Australian National University.

DALEY, D. J., AND C. D. TRENGOVE. 1977. Bounds for Mean Waiting Times in Single-Server Queues: A Survey, Statistics Dept. (I. A. S.), Australian National University.

ECKBERG, A. E., JR. 1977. Sharp Bounds on Laplace-Stieltjes Transforms, with Applications to Various Queueing Problems. *Math. Opns. Res.* **2,** 135–142.

GAEDE, K.-W. 1965. Konfidenzgrengen bei Warteschlangen- und Lagerhaltungs-Problemen. *Zeit. Angew. Math. Mech.* **45,** T91–92.

HAJEK, B. 1983. The Proof of a Folk Theorem on Queueing Delay with Applications to Routing in Networks. *J. Assoc. Comp. Mach.* **30,** 834–851.

HARRISON, J. M. 1977. Some Stochastic Bounds for Dams and Queues. *Math Opns. Res.* **2,** 54–63.

HEYMAN, D. P. 1982. On Ross's Conjectures about Queues with Non-stationary Poisson Arrivals. *J. Appl. Prob.* **19,** 245–249.

HUMBLET, P. A. 1982. Determinism Minimizes Waiting Time in Queues, Department of Electrical Engineering and Computer Science, M.I.T.

KRAEMER, W., AND M. LANGENBACH-BELZ. 1976. Approximate Formulae for the Delay in the Queueing System $GI/G/1$. In *Congressbook, 8th International Teletraffic Congress,* Melbourne, 235-1/8.

KÜHN, P. 1976. Tables on Delay Systems, Institute for Switching and Data Technics, University of Stuttgart.

ROGOZIN, B. A. 1966. Some Extremal Problems in Queueing Theory. *Theor. Prob. Appl.* **11,** 144–151.

ROLSKI, T. 1972. On Some Inequalities for *GI/M/n* Queues. *Zastosowania Mat.* **13,** 43–47.

ROLSKI, T. 1976. Order Relations in the Set of Probability Distributions and Their Application in Queueing Theory. *Dissertationes Math.* **132,** 1–47.

ROLSKI, T. 1981. Queues with Non-stationary Input Stream: Ross's Conjecture. *Adv. Appl. Prob.* **13,** 603–618.

ROLSKI, T., AND D. STOYAN. 1976. On the Comparison of Waiting Times in *GI/G/*1 queues. *Opns. Res.* **24,** 197–200.

ROSS, S. M. 1970. *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco.

ROSS, S. M. 1978. Average Delay in Queues with Non-stationary Poisson Arrivals. *J. Appl. Prob.* **15,** 602–609.

STOYAN, D. 1983. *Comparison Methods for Queues and Other Stochastic Models.* John Wiley & Sons (to appear). (English translation and revision of *Qualitative Eigenschaften and Abschätzungen Stochastischer Modelle,* 1977, D. J. Daley (ed.).)

STOYAN, D., AND H. STOYAN. 1969. Montonieeigenschaften der Kundenwartezeiten im Model *GI/G/1. Zeit. Angew. Math. Mech.* **49,** 729–734.

WHITT, W. 1980. The Effect of Variability in the *GI/G/s* Queue. *J. Appl. Prob.* **17,** 1062–1071.

WHITT, W. 1981a. Comparing Counting Processes and Queues. *Adv. Appl. Prob.* **13,** 207–220.

WHITT, W. 1981b. On Stochastic Bounds for the Delay Distribution in the *GI/G/s* Queue. *Opns. Res.* **29,** 604–608.

WHITT, W. 1982a. On the Heavy-Traffic Limit Theorem for *GI/G/*∞ Queues. *Adv. Appl. Prob.* **14,** 171–190.

WHITT, W. 1982b. The Marshall and Stoyan Bounds for *IMRL/G/*1 Queues are Tight. *Opns. Res. Lett.* **1,** 209–213.

WHITT, W. 1984a. On Approximations for Queues; I. Extremal Distributions. *AT&T Bell Lab. Tech. J.* **63,** 000–000.

WHITT, W. 1984b. On Approximations for Queues; III. Mixtures of Exponential Distributions. *AT&T Bell Lab. Tech. J.* **63,** 000–000.

WOLFF, R. W. 1977. The Effect of Service Time Regularity on System Performance. In *Computer Performance,* pp. 297–304, K. M. Chandy and M. Reiser (eds.). North-Holland, Amsterdam.