THE MARSHALL AND STOYAN BOUNDS FOR IMRL/G/1 QUEUES ARE TIGHT

Ward WHITT

Bell Laboratories, WB-1A350, Crawfords Corner Road, Holmdel, NJ 07733, U.S.A.

Received September 1982
Revised November 1982

Previously established upper and lower bounds for the mean waiting time in a GI/G/1 queue given an interarrival-time distribution with increasing mean residual life are shown to be tight. Distributions for which the inequalities become equalities are displayed. The corresponding bounds for DMRL distributions are not tight.

Queues, bounds, approximations, monotone mean residual life

1. Introduction and summary

Consider the GI/G/1 queue partially specified by the first two moments of the interarrival times and service times. Let λ be the arrival rate, ρ the traffic intensity, c_a^2 the squared coefficient of variation of the interarrival-time distribution (variance divided by the square of its mean), the c_s^2 the squared coefficient of variation of the service-time distribution. Kingman [11] showed that the mean steady-state waiting time, EW, is bounded above as follows:

$$EW \leqslant B \equiv \frac{\rho^{-1}c_a^2 + \rho c_s^2}{2\mu(1-\rho)}. \tag{1}$$

Marshall [13] showed that if the interarrival-time distribution is also DFR (has decreasing failure rate), then

$$EW \leqslant B - \frac{\rho^{-1}c_a^2 + 1}{2\mu} = \frac{(c_a^2 - 1) + \rho(c_s^2 + 1)}{2\mu(1-\rho)}. \tag{2}$$

(See the Appendix for definitions of DFR and related concepts.) In fact, Daley has shown that the inequality (2) still holds if the interarrival-time distribution is only IMRL (has increasing mean residual life); see p. 26 of Daley and Trengove [7] or (5.6.5) of [16]. Stoyan and Stoyan [17] also established a lower bound for EW which holds when the interarrival-time distribution is only NWUE (new worse than used in expectation, which is implied by IMRL):

$$EW \geqslant \frac{\rho(1 + c_s^2)}{2\mu(1-\rho)}, \tag{3}$$

which is just the bound (2) with $c_a^2 = 1$, i.e., the exact M/G/1 formula; see §1.5 and 5.4 of [16] or Whitt [18]. (For NWUE distributions, $c^2 \geqslant 1$.)

Kingman's upper bound (1) is asymptotically tight in heavy traffic, but not in general. In fact, Daley [6] obtained a better upper bound, namely,

$$EW \leqslant \frac{(2-\rho)c_a^2 + \rho c_s^2}{2\mu(1-\rho)}. \tag{4}$$

The main purpose of this note is to show that the bounds (2) and (3) for IMRL and NWUE interarrival-time distributions, respectively, are tight by exhibiting interarrival-time distributions that make (2) and (3) equalities for all service-time distributions with the given moments. Thus, for IMRL/G/1 queues with the first two moments of the interarrival times and service times specified, the maximum relative error in EW (the upper bound minus the lower bound divided by the lower bound) is

$$\mathrm{MRE}(EW) = \frac{(c_a^2 - 1)}{\rho(c_a^2 + 1)}. \tag{5}$$

By Little's formula $L = \lambda W$, the associated maxi-

0167-6377/82/0000-0000/$02.75 © 1982 North-Holland

mum relative error for the expected number in system, EN, is

$$\mathrm{MRE}(EN)=\frac{c_a^2-1}{2+\rho(c_s^2-1)}. \qquad (6)$$

We can also identify the set of all possible values for EW in an IMRL/G/1 queue with the first two moments of the interarrival times and service times specified: It is the interval $(a, b]$ with a and b given by (3) and (2). (It is not difficult to show that all values in the interval can be realized.)

Since the interarrival-time distributions attaining the bounds are mixtures of two exponential distributions, no further improvement can be obtained by assuming the additional shape constraints of complete monotonicity or log-convexity for the interarrival-time distribution; see Chapter 5 of Keilson [10]. It turns out that the extremal interarrival-time distributions do not depend on the service-time distribution at all. Moreover, the mean waiting time given the extremal interarrival-time distribution depends on the service-time distribution only through its mean and variance.

For the special case of exponential service-time distributions and completely monotone interarrival-time distributions (mixtures of exponential distributions), the extremal distributions and the associated MRE in (6) were derived in Whitt [19]. In that setting there is a systematic procedure for deriving the extremal distributions based on complete Techbycheff systems. Here we show that these same interarrival-time distributions are extremal for more general service-time and interarrival-time distributions.

The bounds (2) and (3) with the inequalities reversed were obtained for IFR and NBUE distributions, respectively, at the same time as (2) and (3), but the tightness properties are not the same. In the IFR case, we show that the analogues of (2) and (3) are not tight.

It is natural to combine the bounds (2) and (3) to obtain approximations for general GI/G/1 queues based on the DFR and IFR structure. In particular, it is reasonable to require that approximations for EW fall between (2) and (3). Moreover, since (2) is asymptotically correct in heavy traffic (as $\rho \to 1$) while (3) is not unless $c_a^2 = 1$, it is also reasonable to require that approximations should approach (2) as ρ increases. We call approximations *MFR approximations* (monotone failure rate) if they are of the form

$$\mathrm{Approx}(EW)=f(\rho)\mathrm{Bd}(2)+[1-f(\rho)]\,\mathrm{Bd}(3) \qquad (7)$$

where

$$0 \leqslant f(\rho) \leqslant 1$$

for all ρ and $f(\rho)$ is an increasing function of ρ. The special case of $f(\rho)=\rho$ yields

$$\mathrm{Approx}(EW)=\frac{\rho(c_a^2+c_s^2)}{2\mu(1-\rho)}, \qquad (8)$$

which is an approximation that is often proposed; see p. 221 of Arnold [1], Sakasegawa [14], Shanthikumar and Buzacott [15], and Yu [20]. However, as we indicate in Section 5, it is perhaps better to have $f(\rho) > \rho$.

The rest of this note is organized as follows. In Section 2 we present the interarrival-time distribution that attains the upper bound (2). In Section 3 we present the interarrival-time distributions that asymptotically attain the lower bound (3). In Section 4 we discuss the case of DMRL interarrival-time distributions. Finally, in Section 5 we discuss related work by Daley, Karpelevich, and Kreinen [8,9,12].

2. The upper IMRL bound

The interarrival-time distribution yielding the upper bound (2) is a mixture, being distributed according to an exponential distribution having mean $(1+c_a^2)/2\lambda$ with probability $2/(c_a^2+1)$ and taking the value 0 with probability $(c_a^2-1)/(c_a^2+1)$; its cdf is thus

$$F(x)=1-\left[2/(c_a^2+1)\right]\mathrm{e}^{-2\lambda x/(c_a^2+1)}, \quad x \geqslant 0. \qquad (9)$$

The mass at 0 can be interpreted as the limit of exponential distributions with asymptotically negligible means, so the upper bound can be approximated arbitrarily closely with mixtures of two proper exponential distributions.

Note that the arrival process with the interarrival-time cdf F in (9) is equivalent to a compound Poisson process with geometrically distributed batches. Thus the GI/G/1 system we claim attains the upper bound in (2) is none other than an M/G/1 system with batch arrivals. Hence, it is a simple matter to prove that the cdf F in (9) attains the bound: We apply known formulas for the mean waiting time in an M/G/1 queue with batch arrivals. From §5.10 of Cooper [5], it follows that equality holds in (2) with F. (We use Poisson arrival rate $2/m_1(1+c_a^2)$, mean of the batch size $(1+c_a^2)/2$, and variance of the batch size $(c_a^2+1)(c_a^2-1)/4$; then $EW_1=\rho^2(c_a^2+c_s^2)/2\lambda(1-\rho)$ and $EW_2=\rho(c_a^2-1)/2\lambda$ in (10.1) and (10.2) of [5].)

3. The lower IMRL bound

The interarrival-time distribution yielding the lower bound (3) is the exponential distribution with mean λ^{-1}. Of course, this distribution does not have the correct squared coefficient of variation, c_a^2, but it is the limit of distributions that do. For each b, let the approximating interarrival-time distribution be the mixture of two exponential distributions: one having mean b/λ with probability $(c_a^2-1)/(c_a^2-1+2(b-1)^2)$ and the other having mean $(1-(c_a^2-1)/2(b-1))/\lambda$ with probability $2(b-1)^2/(c_a^2-1+2(b-1)^2)$. For each b, the approximating interarrival-time distribution has mean λ^{-1} and squared coefficient of variation c_a^2.

As $b\to\infty$, the approximating interarrival-time distribution approaches the exponential distribution with mean λ^{-1}. Moreover, the mean waiting time as a function of b, $EW(b)$, converges to the mean waiting time in the associated M/G/1 queue with arrival rate λ. By the corollary on p. 58 of Borovkov [4], the steady-state waiting-time distributions converge. By uniform integrability, the means converge too; see p. 32 of Billingsley [3]. To get uniform integrability, note that the interarrival-time distribution indexed by b is stochastically larger than the exponential distribution with mean $(1-(c_a^2-1)/2(b-1))/\lambda$. Hence, the associated steady-state waiting time $W(b)$ is stochastically smaller than the waiting time of the M/G/1 system, which implies uniform integrability.

4. DMRL interarrival-time distributions

When the inequalities are reversed, the bounds (2) and (3) hold for IFR distributions, in fact, for DMRL and NBUE distributions, respectively. The NBUE upper bound, i.e., the analogue of (3) obviously cannot be tight because for $c_a^2<\rho^2$ it is greater than Kingman's upper bound (1).

For the DMRL lower bound, there obviously can be no analogue of Section 2 because for some parameter values the bound (2) is negative. In particular, for $c_a^2+\rho(c_s^2+1)<1$, the bound is negative. Of course, this can only occur if $c_a^2\leqslant 1$, as is the case when the interarrival-time distribution is DMRL.

We now provide additional information about the DMRL lower bound in the case of deterministic service times, i.e., when $c_s^2=0$. We show that there is a DMRL interarrival-time distribution with $EW=0$ if and only if $c_a^2\leqslant(1-\rho)^2$. We *conjecture* that the infimum of EW over all DMRL distributions is positive when $c_a^2>(1-\rho)^2$. If the conjecture is true, then for $(1-\rho)^2<c_a^2<1-\rho$ the bound (2) is negative while the infimum of EW is positive. For the following proof, let u and v be generic interarrival-time and service-time random variables.

The actual steady-state mean waiting time obviously if 0 if and only if $P(u>Ev)=1$. Hence, to achieve $EW=0$, we must have $u=Ev+X$ where X is a nonnegative random variable with mean $Eu-Ev$ and variance $(Eu)^2c_a^2$. Hence, X has squared coefficient of variation $c_a^2/(1-\rho)^2$. Since u has a DMRL distribution, so must X. Hence, the squared coefficient of variation of X must be less than or equal to 1. In other words, it is possible to represent u as $Ev+X$ if and only if $c_a^2\leqslant(1-\rho)^2$. If $c_a^2\leqslant(1-\rho)^2$, then the minimum value of EW is 0 and it is attained by any nonnegative random variable X having a DMRL distribution with squared coefficient of variation $c_a^2/(1-\rho)^2$. In each case, X can be given a shifted exponential distribution, i.e., we can let $X=a+Y$ where $a\geqslant 0$ and Y is an exponential distribution. More generally, the shifted-exponential distribution appears to be a good candidate for achieving the minimum GI/G/1 mean waiting time over all DMRL interarrival-time distributions for other service-time distributions and other values of c_a^2 and ρ. However, we conjecture that the shifted-exponential distribution does not always yield the minimum.

In closing, we note that a better lower bound in the DMRL case cannot be contained from Marshall's argument [13], which is based on a formula relating EW to the steady-state idle-time, say I. There is a sequence of service-time distributions for which the key inequality (20) in [13] is asymptotically an equality. Consider the service-time distribution with mass $c_s^2/(1+c_s^2)$ on 0 and mass $1/(c_s^2+1)$ on $Ev(1+c_s^2)$. As $c_s^2 \to \infty$, almost all mass is on 0 and, for fixed DMRL interarrival-time distributions, the idle-time distribution and its first two moments converge as $c_s^2 \to \infty$ to the interarrival-time distribution and its first two moments. Hence, $EI^2/2EI \to Eu^2/2Eu$ as $c_s^2 \to \infty$ and Marshall's bound on $EI^2/2EI$ is tight.

5. The Karpelevich–Kreinen curve

Daryl Daley (personal communication) has shown that the tightness of the IMRL/G/1 bounds can also be derived from the Karpelevich–Kreinen [8] curve for the interarrival time u. This curve Γ_u is specified parametrically by

$$\Gamma_u = \{(E(u-x)_+, E(u-x)_+^2): 0 \leqslant x \leqslant \infty\}, \quad (10)$$

where $(x)_+ = \max\{x, 0\}$. The curve moves from (Eu, Eu^2) to $(0, 0)$ as x increases. The slope of the curve at x is just the mean residual life $E(u-x|u > x)$. Since

$$EW = \frac{E(v-u)^2 - E(u-v-W)_+^2}{2(Eu-Ev)}, \quad (11)$$

see §3 of Daley [6], bounds on EW given the first two moments of u and v are equivalent to bounds on $E(u-v-W)_+^2$. Daley's analysis of the Karpelevich–Kreinen curve shows that in the IMRL case the extreme cases of the Karpelevich–Kreinen curve given two moments of u are linear functions. Moreover, they are obtained by the extremal distributions in Sections 2 and 3. The linearity implies that the extreme values of $E(u-X)_+^2$ are attained by the same distributions of u for any random variable X, e.g., $X = v + W$.

In the DMRL case, the extremal curves given two moments of u are not linear, so it is not possible to consider the extremal Karpelevich–Kreinen curves independently of v and W in order to obtain bounds for EW.

Kreinen [12] has also applied the curve (10) to show that the approximation (8) is an upper (lower) bound for EW in $E_k/G/1$ queues (gamma interarrival times) having $c^2 \leqslant 1$ ($c^2 \geqslant 1$). This suggests that the approximation (8) might be smaller (larger) when $c^2 \leqslant 1$ ($c^2 \geqslant 1$). This can be achieved, for example, by making $f(\rho) \geqslant \rho$ in (7). A natural choice is $f(\rho) = \rho^d$ where $0 < d < 1$. Even more promising for the case $c^2 \leqslant 1$, when the bounds (2) and (3) are not tight, is to replace the upper bound (3) in (7) with (8).

Appendix. Some notions of aging

Here we define classes of probability distributions on the nonnegative real line. For more discussion, see Barlow and Proschan [2] or Stoyan [16].

Definition 1. A real-valued random variable X is *stochastically less than or equal to* another real-valued random variable Y, which we denote by $X \leqslant_{st} Y$, if

$$P(X > t) \leqslant P(Y > t) \quad \text{for all } t.$$

Definition 2. Given a random variable X on the nonnegative real line, the *residual lifetime* of X after t is a random variable X_t with

$$P(X_t > x) = \begin{cases} P(X > x + t | X > t) \\ \quad \text{if } P(X > t) > 0, \\ 0, \quad \text{if } P(X > t) = 0 \end{cases}$$

for $x \geqslant 0$.

Definition 3. The distribution of the nonnegative random variable X

(a) has a decreasing failure rate (DFR) if

$$X_s \leqslant_{st} X_t, \quad 0 \leqslant s < t < \infty,$$

(b) has increasing mean residual life (IMRL) if

$$EX_s \leqslant EX_t, \quad 0 \leqslant s < t < \infty,$$

(c) is new worse than used (NWU) if

$$X \leqslant_{st} X_t, \quad t \geqslant 0,$$

(d) is new worse than used in expectation (NWUE) if

$$EX \leqslant EX_t, \quad t \geqslant 0.$$

If the inequalities are reversed in Definition 3, the notions are called, respectively, increasing failure rate (IFR), decreasing mean residual life (DMRL), new better than used (NBU) and new better than used in expectation (NBUE).

Acknowledgment

I am grateful to Daryl Daley for his comments and his contribution in Section 5 and to the referee for several helpful comments.

References

[1] A.O. Allen, *Probability, Statistics and Queueing Theory, with Computer Applications*, Academic Press, New York (1978).

[2] R.E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing: Probability Models*, Holt, Rinehart and Winston, New York (1975).

[3] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York (1968).

[4] A.A. Borovkov, *Stochastic Processes in Queueing Theory*, Springer, New York (1976).

[5] R.B. Cooper, *Introduction to Queueing Theory*, 2nd ed., North-Holland, New York (1981).

[6] D.J. Daley, "Inequalities for moments of tails of random variables, with a queueing application", *Z. Wahrsch. Verw. Gebiete* **41**, 139-143 (1977).

[7] D.J. Daley and C.D. Trengove, "Bounds for mean waiting times in single-server queues: A survey", Unpublished paper, Department of Statistics, The Australian National University (1977).

[8] F.I. Karpelevich and A.J. Kreinen, "Estimates of the mean waiting time in a single-queue system (GI/G/1)", *Engrg. Cybernet.* **14**, 81–83 (1976).

[9] F.I. Karpelevich and A.J. Kreinen, "Some bounds for the mean waiting time in E_k/GI/1 queues", *Math. Operationsforsch. Statist.* **11**, 85–88 (1980).

[10] J. Keilson, *Markov Chain Models — Rarity and Exponentiality*, Springer, New York (1979).

[11] J.F.C. Kingman, "Some inequalities for the queue GI/G/1", *Biometrika* **49**, 315–324 (1962).

[12] A.J. Kreinen, "Estimating mean characteristics of systems with Erlang input and general service times", *Izv. Akad. Nauk SSSR Tehn. Kibernet.* **3**, 187–192 (1981) (in Russian).

[13] K.T. Marshall, "Some inequalities in queueing", *Operations Res.* **16**, 651–665 (1968).

[14] H. Sakasegawa, "An approximation formula $L_q \cong \alpha\rho^\beta/(1-\rho)$", *Ann. Inst. Statist. Math.* **29**, 67–75 (1977).

[15] J.G. Shanthikumar and J.A. Buzacott, On the approximations to the single server queue, *Internat. J. Production Res.* **18**, 761–773 (1980).

[16] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, New York (1982). (English translation and revision, D.J. Daley, Ed., of *Qualitative Properties and Bounds for Stochastic Models*, in German, 1977).

[17] D. Stoyan and H. Stoyan, "Monotonicity properties of the waiting times in the GI/G/1 queue", *Z. Angew. Math. Mech.* **49**, 729–734 (1969) (in German).

[18] W. Whitt, "The effect of variability in the GI/G/s queue", *J. Appl. Probability* **17**, 1062–1071 (1980).

[19] W. Whitt, "On the quality of two-moment approximations for queues, III: Mixtures of exponential distributions", Submitted for publication.

[20] P.S. Yu, "On accuracy improvement and applicability conditions of diffusion approximation with applications to modeling of computer systems", TR-129, Digital Systems Laboratory, Stanford University (1977).