Heuristic Approximations for the Mean Delay in the GI/G/s Queue

Toshikazu KIMURA
Associate Professor
Faculty of Economics
Hokkaido University

Abstract: We provide some heuristic two-moment approximation formulas for the mean waiting time in a GI/G/s queue. These formulas are certain combinations of the exact mean waiting times for D/M/s, M/D/s and M/M/s queues. To see the quality of the approximations, they are numerically compared with exact solutions and other approximations for some particular cases.

Key Words: GI/G/s queue; mean waiting time; two-moment approximation.

1. Introduction

In this paper we provide some heuristic approximation formulas for the mean waiting time in a multi-server queue. We consider the standard GI/G/s queueing system with s homogeneous servers in parallel, unlimited waiting room, the first-come first-served discipline and i.i.d (independent and identically distributed) service times which are independent of a renewal arrival process. We approximate the mean waiting time in this GI/G/s queue by combining those for analyzable systems such as D/M/s, M/D/s and M/M/s queues.

Let EW(GI/G/s) denote the mean waiting time (until beginning service) in the GI/G/s queue, assuming that the system is stable. Let u and v be generic interarrival times and service times, respectively. Then, among approximation formulas we provide in this paper, the best one in an average sense is

$$EW(GI/G/s) \cong \frac{c_u^2 + c_v^2}{2} \, g \, EW(M/M/1) \times$$

$$\times \prod_{n=2}^{s} \left\{ \frac{1-c_u^2}{c_u^2 + c_v^2} \frac{EW(D/M/n)}{EW(D/M/n-1)} + \frac{1-c_v^2}{c_u^2 + c_v^2} \frac{EW(M/D/n)}{EW(M/D/n-1)} + \frac{2(c_v^2 + c_v^2 - 1)}{c_u^2 + c_v^2} \frac{EW(M/M/n)}{EW(M/M/n-1)} \right\}, \quad (1)$$

where c_u^2 (c_v^2) is the squared coefficient of variation (variance devided by the square of mean) of u (v); $g \equiv g(\rho, c_u^2, c_v^2)$ is defined as

$$g(\rho, c_u^2, c_v^2) = \begin{cases} \exp\{-\frac{2(1-\rho)}{3\rho} \frac{(1-c_u^2)^2}{c_u^2 + c_v^2}\}, & c_u^2 \leq 1 \\ 1, & c_u^2 > 1 \end{cases} \quad (2)$$

with the traffic intensity $\rho \equiv Ev/Eu \in (0,1)$. Of course, suppose that both of the mean service time and the traffic intensity are common among the mean waiting times appeared in (1). The exact mean waiting times for the building-block systems, i.e., the D/M/n, M/D/n and M/M/n (n = 1,..., s) queues, can be obtained either by computing their analytical solutions or by using some queueing tables; see Hillier and Yu (1981) and Page (1982).

We see that the approximation (1) with (2) is exact for the M/D/s, M/M/s and M/G/1 queues, but not for the D/M/s queue. Hence, the approximation (1) is an incomplete *interpolation approximation* among these systems; cf. Page (1972) and Kimura (1986). We also see that the approximation (1) is asymptotically exact as $\delta \to 1$ from below, i.e., it is consistent with the heavy traffic limit theorem of Köllerström (1974). Several other approximations we provide in this paper also satisfy these properties.

The approximation (1) is a *two-moment approximation* for EW(GI/G/s), i.e., it depends only on the first two moments of u and v. Closely related two-moment approximations have been developed by Page (1972) and

Kimura (1986), in which three exact mean waiting times for the D/M/s, M/D/s and M/M/s queues are used as their building blocks. This paper shows that our approximations including (1) perform as well as Kimura's approximations and also that they are sometimes better than these two approximations especially for $E_m/E_k/s$ (m,k ≠ 1) queues.

Two-moment approximations for EW(GI/G/s) are of course useful for analyzing an individual GI/G/s queue. Moreover, they also are useful for designing and/or evaluating an open non-Markovian network of queues: We analyze each of nodes in a network as a separate GI/G/s queue characterized by the first two moments of the interarrival-time and service-time distributions. This approach is adopted in a software package called *QNA* (Queueing Network Analyzer) which has been developed to calculate approximate congestion measures for networks of queues; see Whitt (1983). Our approximation formulas can be used in QNA-like softwares to obtain several congestion measures for the whole network as well as each node.

This paper is organized as follows: In Section 2, we focus on the ratio EW(GI/G/m)/EW(GI/G/n) $(m > n)$ instead of EW(GI/G/s). We approximate this ratio by combining the corresponding ratios for the D/M/s, M/D/s and M/M/s queues. Some approximation formulas for EW(GI/G/s) can be derived from this approximation, using appropriate weights in the combination. In Section 3, we discuss the quality of the approximations by numerical comparisons for some particular cases.

2. Approximations for EW(GI/G/s)

For the M/G/s and GI/M/s queues, Cosmetatos (1974, 1976) derived the following approximate relations, taking into account the heavy traffic behavior of the mean waiting times for these systems: For $m > n \geqq 1$,

$$\frac{EW(M/G/m)}{EW(M/G/n)} \cong \frac{1-c_v^2}{1+c_v^2}\frac{EW(M/D/m)}{EW(M/D/n)} + \frac{2c_v^2}{1+c_v^2}\frac{EW(M/M/m)}{EW(M/M/n)}, \tag{3}$$

$$\frac{EW(GI/M/m)}{EW(GI/M/n)} \cong \frac{1-c_u^2}{1+c_u^2}\frac{EW(D/M/m)}{EW(D/M/n)} + \frac{2c_u^2}{1+c_u^2}\frac{EW(M/M/m)}{EW(M/M/n)}. \tag{4}$$

Using these relations, Cosmetatos suggested some approximation formulas for EW(M/G/s) and EW(GI/M/s). In this section we generalize the approximate relations (3) and (4) to the case of the GI/G/s queue.

A basic idea in the approximate relation (3) is that the ratio EW(M/G/m)/EW(M/G/n) $(m > n \geqq 1)$ can be approximately represented as a weighted sum of those for the M/D/s and M/M/s queues. Also, the approximate relation (4) is based on the same heuristic idea as (3). Applying this idea extensively to the GI/G/s queue, we propose the generalized approximate relation

$$\frac{EW(GI/G/m)}{EW(GI/G/n)} \cong w_{01}\frac{EW(D/M/m)}{EW(D/M/n)} + w_{10}\frac{EW(M/D/m)}{EW(M/D/n)} + w_{11}\frac{EW(M/M/m)}{EW(M/M/n)},$$

$$m > n \geqq 1, \tag{5}$$

where $w_{ij} = w_{ij}(c_u^2, c_v^2)$ $(i,j = 0,1)$ is a weighting factor with $w_{01} + w_{10} + w_{11} = 1$. Since the approximate relation (5) includes (3) and (4) as its particular cases, the weights $\{w_{ij}\}$ need to satisfy the conditions

$$w_{01}(1,c_v^2) = 0,\ w_{10}(1,c_v^2) = \frac{1-c_v^2}{1+c_v^2},\ w_{11}(1,c_v^2) = \frac{2c_v^2}{1+c_v^2}, \tag{6}$$

$$w_{01}(c_u^2,1) = \frac{1-c_u^2}{1+c_u^2},\ w_{10}(c_u^2,1) = 0,\ w_{11}(c_u^2,1) = \frac{2c_u^2}{1+c_u^2}. \tag{7}$$

To obtain appropriate weights satisfying these conditions, we will make use of Kimura's (1986) approximation for EW(GI/G/s): Kimura has recently developed a heuristic two-moment approximation for EW(GI/G/s) which is just a *weighted harmonic mean* of the exact mean waiting times EW(D/M/s), EW(M/D/s) and EW(M/M/s), i.e.,

$$\frac{1}{EW(GI/G/s)} \cong \frac{1-c_u^2}{c_u^2+c_v^2}\,\frac{1}{EW(D/M/s)} + \frac{1-c_v^2}{c_u^2+c_v^2}\,\frac{1}{EW(M/D/s)}$$

$$+ \frac{2(c_u^2+c_v^2-1)}{c_u^2+c_v^2}\,\frac{1}{EW(M/M/s)}, \tag{8}$$

It is easy to check that the weights used in Kimura's approximation certainly satisfy the conditions (6) and (7) which are desired in (5). Hence, as a candidate of the weights, we choose

$$\textit{Case A}: \quad w_{01} = \frac{1-c_u^2}{c_u^2+c_v^2}, \quad w_{10} = \frac{1-c_v^2}{c_u^2+c_v^2}, \quad w_{11} = \frac{2(c_u^2+c_v^2-1)}{c_u^2+c_v^2}. \tag{9}$$

However, as noted in Remark 3.1 of Kimura (1986), the way of generalization is *not* unique. In fact, there is a simple alternative of (9):

$$\textit{Case B}: \quad w_{01} = \frac{(1-c_u^2)c_v^2}{c_u^2+c_v^2}, \quad w_{10} = \frac{c_u^2(1-c_v^2)}{c_u^2+c_v^2}, \quad w_{11} = \frac{2c_u^2c_v^2}{c_u^2+c_v^2}. \tag{10}$$

In this paper we restrict our attention to these two cases as the weights in (5) because of their simplicies.

From (5) with the weights (9) or (10), we now derive two different-type approximations for EW(GI/G/s): (i) if we let $n = 1$ and $m = s$ in (5), then we obtain

$$EW(GI/G/s) \cong EW(GI/G/1)\left(w_{01}\frac{EW(D/M/s)}{EW(D/M/1)} + w_{10}\frac{EW(M/D/s)}{EW(M/D/1)}\right.$$

$$\left. + w_{11}\frac{EW(M/M/s)}{EW(M/M/1)}\right); \tag{11}$$

(ii) if we let $n = m-1$ and $m = s$ in (5), then (5) can be regarded as a recursion formula for EW(GI/G/s), so that we obtain

$$EW(GI/G/s) \cong EW(GI/G/1)\prod_{n=2}^{s}\left(w_{01}\frac{EW(D/M/n)}{EW(D/M/n-1)} + w_{10}\frac{EW(M/D/n)}{EW(M/D/n-1)}\right.$$

$$+ w_{11} \frac{EW(M/M/n)}{EW(M/M/n-1)} \Bigg) . \tag{12}$$

To derive tractable approximations for EW(GI/G/s), from (11) and (12), we need a certain simple approximation for EW(GI/G/1). Following Whitt (1983), we will approximate EW(GI/G/1) using

$$EW(GI/G/1) \cong \frac{c_u^2 + c_v^2}{2} g\ EW(M/M/1) \tag{13}$$

with the correction factor g defined in (2). The approximation (13) is the Kraemer and Langenbach-Belz (1976) approximation for $c_u^2 \leqq 1$.

Thus we obtain four different two-moment approximations by combining (9) and (19) with (11) and (12). For convenience, we call (11) ((12)) with (13) the approximation formula of Type I (II). In addition, if the approximation formula is, e.g., Type I with the weights of Case A, we call it the approximation IA, and so forth. For s = 2, the approximation IA (IB) coincides with IIA (IIB). For the M/G/s and GI/M/s queues, the approximation IA (IIA) coincides with IB (IIB). Hence, it is apparent that all of these four approximations are equal for the M/G/2 and GI/M/2 queues. Since the Kraemer and Langenbach-Belz approximation is exact for the M/G/1 queue, the approximations IA and IB for the M/G/s queue are equivalent to Page's approximation for EW(M/G/s):

$$EW(M/G/s) \cong (1-c_v^2)EW(M/D/s) + c_v^2 EW(M/M/s). \tag{14}$$

Remark. For $n > m \geqq 1$, is it appropriate to use (5) as another approximate relation? We solve this problem by the use of the *normed cooperation coefficient* introduced in Boxma et al. (1979): Assume that the arrival process is a Poisson process, i.e., $c_u^2 = 1$. Then, if we let m = 1 and n = s in (5), we have

$$\frac{EW(M/G/1)}{EW(M/G/s)} \cong \frac{1-c_v^2}{1+c_v^2}\,\frac{EM(M/D/1)}{EW(M/D/s)} + \frac{2c_v^2}{1+c_v^2}\,\frac{EW(M/M/1)}{EW(M/M/s)} \,. \tag{15}$$

As in Boxma et al. (1979), we define the normalized quantity

$$N_{Gs} \equiv \frac{EW(M/M/s)}{EW(M/M/1)}\,\frac{EW(M/G/1)}{EW(M/G/s)} \,. \tag{16}$$

From (15) and (16), we obtain

$$N_{Gs} \cong \frac{1}{2}\,\frac{1-c_v^2}{1+c_v^2}\,\frac{EW(M/M/s)}{EW(M/D/s)} + \frac{2c_v^2}{1+c_v^2} \equiv \tilde{N}_{Gs} \,. \tag{17}$$

Letting $s \to \infty$ in (17), we have

$$\lim_{s\uparrow\infty} \tilde{N}_{Gs} = \frac{3c_v^2+1}{2(1+c_v^2)} \,, \tag{18}$$

which is not consistent with the exact result $\lim_{s\uparrow\infty} N_{Gs} = (1+c_v^2)/2$ except that $c_v^2 = 1$; see Remark 1 of Boxma et al. (1979). It is easy to check that the consistency holds for the case $m > n \geqq 1$. Hence, the answer of the problem is in the negative.

3. Numerical Comparisons

To see the quality of the approximations provided in this paper, we compare them with exact solutions and other approximations for some particular cases.

Table 1 compares the approximations for the mean queue length in some M/PH/s queues. Approximations for the mean queue length can be derived from those of the mean waiting time by using Little's formula. Since there is no difference between the weights of Cases A and B for systems with Poisson arrivals, we simply denote our approximations by I(= (14)) and II in the table. The exact values are quoted for comparisons from Table

4.6 of Tijms et al. (1981). In Table 1, $E_{1,2}$ denotes a mixture of M and E_2, and $H_2(H_2^b)$ denotes a mixture of two exponentials (H_2 with balanced means); see Tijms et al. (1981) for their detailed definitions.

Table 1 shows that all of the approximations performs well, providing accuracy adequate for most practical applications. The table also shows that the application (8) is much better than the other approximations for highly variable service-time distributions; see Table 4 of Kimura (1986) for further comparisons with various distribution-dependent approximations for EW(M/PH/s) such as the Boxma et al. (1979) approximation.

Table 1. A Comparison of Approximations for the Mean Queue Length in M/PH/s Queues.

System	c_v^2	ρ	Exact	I	II	(8)
$M/E_2/2$	0.5	0.7	1.020	1.018	1.018	1.023
		0.9	5.773	5.769	5.769	5.776
$M/E_2/10$	0.5	0.7	0.4076	0.4052	0.4046	0.4122
		0.9	4.576	4.560	4.559	4.582
$M/H_2/3$	2.0	0.7	1.660	1.696	1.697	1.650
		0.9	10.92	10.99	10.99	10.90
$M/E_{1,2}/3$	3.0	0.7	1.933	2.244	2.245	2.111
		0.9	14.05	14.62	14.62	14.37
$M/H_2^b/5$	2.0	0.8	3.170	3.277	3.278	3.192
$M/H_2^b/10$	2.0	0.8	2.267	2.395	2.397	2.298
$M/H_2^b/5$	5.0	0.8	5.923	6.457	6.462	5.703
$M/H_2^b/10$	5.0	0.8	4.036	4.669	4.681	3.853
		0.9	16.53	17.69	17.70	16.14

Tables 2-4 give the relative percentage errors of the approximations for some $E_m/E_k/s$ ($m,k \neq 1$) queues. Table 2 deals with the $E_4/E_2/5$ queue; Table 3 deals with 7 $E_m/E_k/2$ queues; Table 4 deals with 6 $E_m/E_2/s$ ($s = 4,8$) queues. In

Table 2. Relative Percentage Errors of Approximations for $EW(E_4/E_2/5)$.

ρ	IA	IB	IIA	IIB	(8)	(19)
0.3	112.04	337.27	-37.19	144.92	-64.02	1249.78
0.5	2.67	51.59	-10.61	36.99	-16.47	118.52
0.7	-1.25	13.19	-2.81	11.68	-0.73	26.02
0.8	-0.81	6.49	-1.25	6.08	0.90	12.32
0.9	-0.33	2.52	-0.40	2.46	0.84	4.67
0.95	-0.14	1.13	-0.16	1.12	0.48	2.07

the approximation IA (IB) coincides IIA (IIB) for s = 2. The exact values of the mean waiting times for these systems can be derived from the Hillier and Yu (1981) tables. In Tables 2-4, our approximations are compared with the two-moment approximations of Kimura (1986) and Page (1972), where Page's approximation for EW(GI/G/s) is

$$EW(GI/G/s) \cong (1-c_u^2)c_V^2 EW(D/M/s) + c_u^2(1-c_V^2)EW(M/D/s) + c_u^2 c_V^2 EW(M/M/s). \tag{19}$$

Tables 2-4 show that the approximation IIA; see (1), is much better than the others not only in heavy traffic but also in moderate traffic. In heavy traffic, the approximations IA and (8) perform as well as IIA. The approximations IB and IIB are less accurate than IA and IIA in the average sense, but the formers are better than the latters for small c_u^2. Page's approximation performs very poorly for $E_m/E_k/s$ queues in moderate traffic.

It is sometimes observed that IA and IIA become negative in light traffic, especially when both c_u^2 and c_V^2 are extremely smaller than one. In practical applications, however, the mean waiting time in light traffic could be almost ignored. Hence, it does not matter so much in light traffic situations whether approximate values are positive or negative.

Table 3. Relative Percentage Errors of Approximations for $EW(E_m/E_k/2)$ (ρ = 0.5,0.9).

ρ	m	k	A	B	(8)	(19)
0.5	2	2	8.99	14.25	-19.57	21.76
	3	2	4.19	12.84	-12.97	34.76
	4	2	-0.99	9.84	-5.44	43.10
	9	2	-15.13	0.25	20.41	59.88
	2	3	7.42	15.76	-17.10	30.56
	3	3	-1.34	12.70	-6.97	52.47
	4	3	-10.19	7.50	4.41	68.66
	Average		-1.01 (6.89)	10.45	-5.32 (12.41)	44.46
0.9	2	2	1.35	1.77	-0.66	1.85
	3	2	0.95	1.62	0.06	2.74
	4	2	0.40	1.23	0.75	3.25
	9	2	-1.40	-0.20	2.65	4.17
	2	3	1.25	1.91	-0.26	2.75
	3	3	0.45	1.56	0.85	4.20
	4	3	-0.54	0.88	1.85	5.08
	Average		0.35 (0.91)	1.25 (1.31)	0.75 (1.01)	3.43

Note: The average absolute relative percentage error is in parentheses below the average.

Table 4. Relative Percentage Errors of Approximations for $EW(E_m/E_2/s)(\rho = 0.5, 0.9)$.

ρ	m	s	IA	IB	IIA	IIB	(8)	(19)
0.5	2	4	14.80	28.88	7.65	24.47	-33.78	42.65
	3	4	8.24	34.23	0.05	27.88	-23.80	70.13
	4	4	-0.74	34.25	-7.51	27.22	-12.74	88.04
	9	4	-31.47	26.85	-27.54	20.62	26.68	120.64
	2	8	46.69	75.00	3.79	46.44	-58.86	101.43
	3	8	57.46	127.36	-2.13	74.96	-40.19	208.63
	Average		15.83 (26.57)	54.42	-4.28 (8.11)	36.93	-23.78 (32.68)	105.25
0.9	2	4	1.28	2.39	1.24	2.36	-0.92	2.53
	3	4	0.66	2.44	0.61	2.41	-0.03	3.65
	4	4	-0.12	2.12	-0.16	2.08	0.82	4.23
	9	4	-2.62	0.64	-2.61	0.61	3.11	5.10
	2	8	1.22	3.40	1.01	3.27	-1.45	3.62
	3	8	1.95	5.57	1.72	5.41	1.41	6.92
	Average		0.40 (1.31)	2.76	0.30 (1.23)	2.69	0.49 (1.29)	4.34

Note: The average absolute relative percentage error is in parentheses below the average.

Acknowledgment:

This research was supported in part by the Grant in Aid for Scientific Research of the Ministry of Education, Science and Culture of Japan under Contract No. 61780016 (1986).

References

Boxma, O.J., Cohen, J.W. and Huffels, N. (1979), "Approximations of the mean waiting time in an M/G/s queueing system", *Operations Research* 27, 1115-1127.

Cosmetatos, G.P. (1974), "Approximate equilibrium results for the multi-server queue (GI/M/r)", *Operational Research Quarterly* 25, 625-634.

Cosmetatos, G.P. (1976), "Some approximate equilibrium results for the multi-server queue (M/G/r)", *Operational Research Quarterly* 27, 615-620.

Hillier, F.S. and Yu, O.S. (1981) *Queueing Tables and Graphs*, North-Holland, New York.

Kimura, T. (1986), "A Two-moment approximation for the mean waiting time in the GI/G/s queue", *Management Science*, 32, 751-763.

Köllerström, J. (1974), "Heavy traffic limit theory for queues with several servers, I", *Journal of Applied Probability* 11, 544-552.

Kraemer, W. and Langenbach-Belz, M. (1976), "Approximate formulae for the delay in the queueing system GI/G/1", Proceedings of the 8th International Teletraffic Congress, Merbourne, 235-1/8.

Page, E. (1972), *Queueing Theory in OR*, Butterworth, London.

Page, E. (1982), "Tables of waiting times for M/M/n, M/D/n and D/M/n and their use to give approximate waiting times in more general queues", *Journal of the Operational Research Society* 33, 453-473.

Tijms, H.C., van Hoorn, M.H. and Federgruen. A. (1981), "Approximations for the steady-state probabilities in the M/G/c queue", *Advances in Applied Probability* 13, 186-206.

Whitt, W. (1983), "The queueing network analyzer", *Bell System Technical Journal* 62, 2779-2815.