Journal of the Operations Research
Society of Japan
Vol. 35, No. 1, March 1992

INTERPOLATION APPROXIMATIONS FOR THE MEAN WAITING TIME IN A MULTI-SERVER QUEUE

Toshikazu Kimura
Hokkaido University

(Received January 10, 1991; Revised May 27, 1991)

Abstract This paper gives numerical validation of a couple of interpolation approximations for the mean waiting time in a $GI/G/s$ queue, which are provided by a unified approach similar to that in Kimura (1991). Both approximations are represented as certain combinations of the mean waiting times for the $GI/M/s$ and $GI/D/s$ queues in which the arrival processes and the mean service times are the same as in the approximating $GI/G/s$ queue. To let these approximations be more tractable, we further provide simple interpolation approximations for the mean waiting times in $GI/M/s$ and $GI/D/s$ queues with low variable interarrival times. The quality of the approximations is tested by comparing them with exact solutions and previous two-moment approximations for a variety of cases. Extensive numerical comparisons indicate that our approximations are more accurate than the two-moment approximations and that the relative percentage errors are in the order of 5% in moderate traffic and in the order of 1% in heavy traffic.

1. Introduction and Summary

In this paper we provide a couple of approximations for the mean waiting time in a multi-server queue, which is an improvement of the approximation proposed in Kimura [14] when interarrival times are highly variable and/or the traffic is not heavily loaded. As in [14], we consider the standard $GI/G/s$ queueing system with s (≥ 2) homogeneous servers in parallel, unlimited waiting room, the first-come first-served discipline and independent sequences of i.i.d. (independent and identically distributed) interarrival times and service times. Let u and v be generic interarrival- and service-time, respectively; let A (B) denote the interarrival- (service-) time cdf with mean λ^{-1} (μ^{-1}); let $\rho = \lambda/s\mu \in [0, 1)$ be the traffic intensity; and let c_a^2 (c_s^2) be the squared coefficient of variation (variance divided by the square of the mean) of u (v). Assume that the cdf A is not deterministic, i.e., $c_a^2 \neq 0$. In addition, let $EW(GI/G/s)$ denote the mean waiting time (until beginning service) in this $GI/G/s$ queue, assuming that the system is in equilibrium state. We approximate $EW(GI/G/s)$ by combining the exact mean waiting times for the $GI/M/s$ and $GI/D/s$ queues both having the interarrival-time cdf A.

There are some two-moment approximations for $EW(GI/G/s)$, which are weighted combinations of the exact mean waiting times for the $M/M/s$, $M/D/s$ and $D/M/s$ queues [11, 13, 14, 22]. Although these approximations are tractable and sufficiently accurate for most practical purposes, it has been known that they become less accurate as the variability parameters c_a^2 and c_s^2 (especially c_a^2) get large, i.e., for the cases where detailed information about A and/or B is important. Our experience seems to suggest that $EW(GI/G/s)$ is not sensitive to the shape of B. However, this is not the case with A. The shape of A affects $EW(GI/G/s)$ quite considerably when the traffic is not heavy and c_a^2 not too small. Two-moment approximations are no longer reliable in such circumstances. In this paper we give full information on A to our approximations by using the $GI/M/s$ and $GI/D/s$ queues as

their building blocks. Extensive numerical studies show that our approximations are more accurate than the previous two-moment approximations in the circumstances above.

The approximations we recommend in this paper are

$$EW(GI/G/s) \simeq \begin{cases} (c_a^2+c_s^2)\left\{\dfrac{(c_a^2+1)c_s^2}{EW(GI/M/s)}+\dfrac{c_a^2(1-c_s^2)}{EW(GI/D/s)}\right\}^{-1}, & \text{if } 1 \le c_a^2 < 2 \\ c_s^2 EW(GI/M/s)+(1-c_s^2)EW(GI/D/s), & \text{otherwise,} \end{cases} \tag{1.1}$$

which will be obtained in a *unified* way. We immediately see that the approximations in (1.1) have the following characteristics: (i) They are, of course, exact for the $GI/M/s$ and $GI/D/s$ queues. (ii) They are asymptotically exact as $\rho \to 1$. (iii) For the $M/G/s$ queue, the approximation for $c_a^2 = 1$ coincides with the excellent approximation in Kimura [11] and is exact for $s = 1$.

Our studies indicate that (1.1) will usually yield satisfactory approximations at least for the cases that (i) $c_a^2 \le 4$ and $c_s^2 \le 4$, and (ii) the traffic intensity is not too small, e.g., $\rho \ge 0.3$ for $s = 2$ and $\rho \ge 0.7$ for $s = 20$. Roughly speaking, the relative percentage errors of (1.1) are in the order of 5% (1%) if the approximate value of $EW(GI/G/s)$ is less (greater) than $10/\mu$. The studies also indicate that the accuracy of our approximations does not so strongly depend on the number s of servers at least for $c_a^2 \le 4$. This property is practically important because algorithmic methods for computing exact solutions of $GI/G/s$ queues, e.g., [25], become infeasible for systems with large s.

In (1.1), the mean waiting times for the building-block systems, i.e., the $GI/M/s$ and $GI/D/s$ queues, have the same mean service times and traffic intensities as those of the approximating $GI/G/s$ queue. The exact values of these mean waiting times can be obtained either by using the queueing tables of Seelen, Tijms and van Hoorn [24] or by computing their analytic and/or algorithmic solutions; see, e.g., [2, 21] for the $GI/M/s$ queue and [20, 29] for the $GI/D/s$ queue with phase-type (abbreviated as Ph) arrival distributions. The algorithms for the $Ph/M/s$ and $Ph/D/s$ queues are computationally feasible for very large values of s, e.g., up to $s = 250$ servers when the number of arrival phases is 10; see [29].

The approximations in (1.1) are useful not only for quick calculation of $EW(GI/G/s)$ but also for obtaining approximations for the distributions of the number of customers and of the waiting time: Wu and Chan [30] proposed simple approximations for these queueing characteristics in the $GI/G/s$ queue by the use of maximum entropy analysis. Shore [26, 27] derived more heuristic approximations for the characteristics by viewing the $GI/G/s$ queue as a system alternating between associated $GI/G/1$ and $GI/G/\infty$ queues. In both of the approximations, $EW(GI/G/s)$ plays a key role to derive explicit formulas. Our approximations can be directly applied to these approaches.

This paper is organized as follows: In Section 2, following Kimura [14], we focus on the ratio $EW(GI/G/s)/EW(GI/G/1)$ and its reciprocal. We approximate each of these quantities by a linearly weighted sum of the corresponding quantities for the $GI/M/s$ and $GI/D/s$ queues. From these approximate relations, we derive a couple of approximations which are consistent with exact properties for particular cases. To let our approximations be more tractable for systems with $c_a^2 \le 1$, we further provide simple approximations for $EW(GI/M/s)$ and $EW(GI/D/s)$ in Section 3. In each section, the quality of the approximations is tested by comparing them with exact solutions for a variety of cases. Finally, in Section 4, we give a concluding remark on a possible direction for future extensions.

2. Approximating $EW(GI/G/s)$

To obtain accurate approximations for queueing characteristics, we often need the extreme limiting behavior as $s \to \infty$ or $\rho \to 0$ or $\rho \to 1$. For $EW \equiv EW(GI/G/s)$, it is obvious that $EW \to 0$ as $s \to \infty$ or $\rho \to 0$ and $EW \to \infty$ as $\rho \to 1$. A good way to analyze asymptotic properties of EW in these extreme cases is to normalize EW so that nondegenerate limits occur.

In this paper, as in Kimura [14], we focus on the ratio $EW(GI/G/s)/EW(GI/G/1)$ and its reciprocal as such normalized quantities. In these quantities, we assume that the $GI/G/1$ queue has the same mean service time and traffic intensity as those of the approximating $GI/G/s$ queue. We approximate each of these ratios by a linearly weighted sum of the corresponding ratios for the $GI/M/s$ and $GI/D/s$ queues both having the interarrival-time cdf A, i.e.,

$$\frac{EW(GI/G/s)}{EW(GI/G/1)} \simeq \omega \frac{EW(GI/M/s)}{EW(GI/M/1)} + (1-\omega)\frac{EW(GI/D/s)}{EW(GI/D/1)}, \tag{2.1}$$

and

$$\frac{EW(GI/G/1)}{EW(GI/G/s)} \simeq \nu \frac{EW(GI/M/1)}{EW(GI/M/s)} + (1-\nu)\frac{EW(GI/D/1)}{EW(GI/D/s)}, \tag{2.2}$$

where ω and ν denote weighting coefficients. This approximation is motivated by a similar idea of Tijms [28, (4.223)] for plain performance measures and the idea of Cosmetatos [3, 5] for $EW(GI/M/s)$ and $EW(M/G/s)$. For convenience, assume that these coefficients depend only on the variability parameters c_a^2 and c_s^2 and *not* on s and ρ. To show the dependency of the weights on c_a^2 and c_s^2 definitely, we write $\omega \equiv \omega(c_a^2, c_s^2)$ and $\nu \equiv \nu(c_a^2, c_s^2)$ henceforth.

From the consistency with the building-block systems in (2.1) and (2.2), we see that these weighting functions satisfy the following conditions: From the consistency with the $GI/M/s$ queue,

$$\omega(c_a^2, 1) = \nu(c_a^2, 1) = 1, \tag{2.3}$$

and from the consistency with the $GI/D/s$ queue,

$$\omega(c_a^2, 0) = \nu(c_a^2, 0) = 0. \tag{2.4}$$

Moreover, we have the next two theorems.

Theorem 2.1 *Assume that* $E[v^3] < \infty$. *Then the approximate relations* (2.1) *and* (2.2) *are asymptotically correct as* $\rho \to 1$ *for any finite weights.*

Theorem 2.2 *For the* $M/G/s$ *queue with the service-time cdf* B, *the approximate relations* (2.1) *and* (2.2) *are asymptotically correct as* $s \to \infty$ *if*

$$\omega(1, c_s^2) = \frac{2c_s^2}{1+c_s^2}, \quad \nu(1, c_s^2) = c_s^2. \tag{2.5}$$

Theorem 2.1 ensures that our approximations are structurally accurate in heavy traffic. The proofs of these theorems are given by virtue of Theorems 2.1 and 2.2 in Kimura [14] and hence they are omitted.

To keep the consistency with the building-block systems and the $M/G/\infty$ queue, we will determine the weighting functions in such a way that they satisfy the conditions (2.3), (2.4) and (2.5) at the same time. However, the weights satisfying these conditions are *not*

uniquely determined; cf. [11]. Hence, taking account of the well-known symmetry of c_a^2 and c_s^2 in the heavy traffic limit theorem [17], we simply approximate ω and ν by

$$\omega(c_a^2, c_s^2) \simeq \frac{(c_a^2+1)c_s^2}{c_a^2+c_s^2}, \tag{2.6}$$

and

$$\nu(c_a^2, c_s^2) \simeq c_s^2. \tag{2.7}$$

From the approximate relation (2.1) ((2.2)) with the weight (2.6) ((2.7)), we have

$$EW(GI/G/s) \simeq \frac{EW(GI/G/1)}{c_a^2+c_s^2}\left\{(c_a^2+1)c_s^2\frac{EW(GI/M/s)}{EW(GI/M/1)} + c_a^2(1-c_s^2)\frac{EW(GI/D/s)}{EW(GI/D/1)}\right\}, \tag{2.8}$$

and

$$EW(GI/G/s) \simeq EW(GI/G/1)\left\{c_s^2\frac{EW(GI/M/1)}{EW(GI/M/s)} + (1-c_s^2)\frac{EW(GI/D/1)}{EW(GI/D/s)}\right\}^{-1}. \tag{2.9}$$

Obviously, these two approximations contain the mean waiting times for three single server queues with the same mean service times and traffic intensities as in the approximating $GI/G/s$ queue. Among these mean waiting times, $EW(GI/G/1)$ and $EW(GI/D/1)$ are difficult to compute except some special cases, e.g., Poisson arrival case. Hence, to simplify the approximations (2.8) and (2.9), we replace the mean waiting times for *all* single server queues by a simple two-moment approximation; see Remark 2.2. As such an approximation, we use the approximation

$$EW(GI/G/1) \simeq \frac{c_a^2+c_s^2}{2}EW(M/M/1); \tag{2.10}$$

see Remark 2.3. Substituting (2.10) into (2.8) and (2.9), we respectively obtain the approximations as

$$EW(GI/G/s) \simeq c_s^2 EW(GI/M/s) + (1-c_s^2)EW(GI/D/s), \tag{2.11}$$

and

$$EW(GI/G/s) \simeq (c_a^2+c_s^2)\left\{\frac{(c_a^2+1)c_s^2}{EW(GI/M/s)} + \frac{c_a^2(1-c_s^2)}{EW(GI/D/s)}\right\}^{-1}. \tag{2.12}$$

The approximation (2.11) is a simple linear interpolation between $EW(GI/M/s)$ and $EW(GI/D/s)$, and it coincides with Tijms' approximation (4.223) in [28]. The approximation (2.12) is a certain harmonic mean of $EW(GI/M/s)$ and $EW(GI/D/s)$. We will show the quality of these approximations through extensive numerical experiments.

Remark 2.1 For the $M/G/s$ queue, it should be noted that (2.11) coincides with Page's [22] approximation and (2.12) with Kimura's [11] approximation; see (2.17) and (2.18), respectively.

Remark 2.2 Instead of (2.10), it is possible to use the exact value of $EW(GI/M/1)$ in (2.8) and (2.9). However, we can easily see that the resultant formulas are *not* exact for the $GI/M/s$ queue. This is why we use the approximation (2.10) for $EW(GI/M/1)$.

Remark 2.3 It seems to be a good idea to replace the mean waiting times for the single server queues appeared in (2.8) and (2.9) by a more accurate approximation, e.g., the Krämer and Langenbach-Belz [18] approximation for $EW(GI/G/1)$, which also has a simple form similar to (2.10). However, from numerical comparisons with some other alternatives for (2.10), we saw that the simple approximation (2.10) fits for our approximations when we use the coefficients (2.6) and (2.7); cf. Kimura [14].

Remark 2.4 By using the light traffic limit theorem in Burman and Smith [1], we can prove for the $M/G/s$ queue that the approximate relations (2.1) and (2.2) are asymptotically correct as $\rho \to 0$ if

$$\omega(1, c_s^2) = \frac{2s}{s-1}\left\{1 - \frac{(s+1)\mu I(s)}{1+c_s^2}\right\} \tag{2.13}$$

and

$$\nu(1, c_s^2) = \frac{s+1}{s-1}\left\{\frac{1+c_s^2}{(s+1)\mu I(s)} - 1\right\}, \tag{2.14}$$

where

$$I(s) = \int_0^\infty \{1 - B_e(t)\}^s \, dt, \tag{2.15}$$

and B_e denote the stationary-excess cdf associated with the service-time cdf B, i.e.,

$$B_e(t) = \mu \int_0^t \{1 - B(x)\} \, dx, \qquad t \geq 0. \tag{2.16}$$

Note that the weighting coefficients depend on s. As in Remark 2.3, it is also a good idea to replace (2.5) by (2.13) and (2.14). However, it is relatively difficult to generalize the weighting coefficients in (2.13) and (2.14) to the $GI/G/s$ case. Taking the light traffic behavior into approximations would be an important subject of our future studies; see Kimura [16].

Table 1: A List of Numerical Experiments.

Arrival (c_a^2)		Service (c_s^2)		s	ρ
E_{10}	(0.1)	E_2	(0.5)	2(1)10(5)25	0.3,0.5,0.7,0.8,0.9,0.95
E_2	(0.5)	H_2^b	(1.5)		
M	(1.0)	H_2^b	(2.5)		
H_2^b	(1.5)	H_2^b	(4.0)		
H_2^b	(2.0)				
H_2^b	(3.0)				
H_2^b	(4.0)				

Numerical Comparisons

Table 1 gives a combination list of the parameters in queueing systems on which we have made numerical experiments to test the performance of our approximations. In Table 1, H_2^b denotes an H_2 distribution with balanced means. The exact values of the mean waiting times for these systems are given in Seelen et al. [24]. It should be noted that all of the exact values are not necessarily available; for example, the exact values for systems with $s = 10$ are available only when $\rho \geq 0.5$. Some typical results of these experiments are given

in Tables 2–5, in which we denote for convenience the approximations (2.11) and (2.12) as "New-I" and "New-II", respectively.

The experiments listed in Table 1 have clarified some qualitative properties of our approximations. First, we will summarize these properties: The approximation New-I is stably accurate for any combination of the variability parameters, while the approximation New-II becomes unstable (e.g., negative) when $c_a^2 < 1$ and $c_s^2 > 1$. The approximation New-II is less accurate than New-I when $c_a^2 < 1$ and $c_s^2 \leq 1$, but performs about the same as New-I when $c_a^2 \geq 1$. In particular, New-II becomes more accurate than New-I as $c_a^2 \to 1$ from above.

Table 2 compares five approximations with the exact values of the mean queue length (excluding customers in service) for $Ph/Ph/10$ queues with low variable interarrival times. Approximations of the mean queue length can be derived from those of EW by using Little's formula. Since the queue length is intuitively easy to capture the level of congestion, we use the mean queue length rather than the mean waiting time in Tables 2–4, 6 and 7. In Table 2, "Sim-I" denotes a simplified version of New-I which will be discussed in Section 3. We add three closely-related two-moment approximations of Page [22] and Kimura [11, 14] in the table. In terms of our notations, Page's approximation is given by

$$EW(GI/G/s) \simeq c_a^2 c_s^2 EW(M/M/s) + c_a^2(1-c_s^2)EW(M/D/s) + (1-c_a^2)c_s^2 EW(D/M/s), \tag{2.17}$$

while Kimura's [11, 14] approximations are respectively given by

$$EW(GI/G/s) \simeq \frac{c_a^2 + c_s^2}{\dfrac{2(c_a^2+c_s^2-1)}{EW(M/M/s)} + \dfrac{1-c_s^2}{EW(M/D/s)} + \dfrac{1-c_a^2}{EW(D/M/s)}} \tag{2.18}$$

and

$$EW(GI/G/s) \simeq \begin{cases} \dfrac{k(c_a^2+c_s^2)}{\dfrac{2(c_a^2+c_s^2-1)}{EW(M/M/s)} + \dfrac{1-c_s^2}{EW(M/D/s)} + \dfrac{k_{01}(1-c_a^2)}{EW(D/M/s)}}, & \text{if } c_a^2 \leq 1 \\ (c_a^2+c_s^2-1)EW(M/M/s) + (1-c_s^2)EW(M/D/s) & \\ \quad + \dfrac{1-c_a^2}{k_{01}} EW(D/M/s), & \text{if } c_a^2 > 1 \end{cases} \tag{2.19}$$

where

$$k \equiv k(\rho, c_a^2, c_s^2) = \exp\left\{-\frac{2(1-\rho)}{3\rho}\frac{(1-c_a^2)^2}{c_a^2+c_s^2}\right\} \tag{2.20}$$

and

$$k_{01} \equiv k(\rho, 0, 1). \tag{2.21}$$

Kimura's approximations (2.18) and (2.19) are denoted in the table by "Kimura86" and "Kimura91", respectively. We omit New-II from comparisons, since New-II is evidently less accurate than the others when $c_a^2 < 1$. Table 2 shows that New-I is more accurate than the two-moment approximations of Page and Kimura86 when $c_a^2 < 1$ and $c_s^2 < 1$. Table 2 also shows that Kimura91 performs as well as New-I for $c_s^2 < 1$. However, when $c_a^2 < 1$ and $c_s^2 > 1$, Page's approximation has the best performance. This tendency is remarkable especially when $0.5 \leq c_a^2 < 1$ and $c_s^2 > 2.5$.

Table 2: A Comparison of Approximations of the Mean Queue Length for $Ph/Ph/10$ Queues with $c_a^2 < 1$.

c_a^2	ρ	Method	$c_s^2 = 0.5$	$c_s^2 = 1.5$	$c_s^2 = 2.5$	$c_s^2 = 4.0$
0.1	0.7	Exact	0.05	0.19	0.33	0.52
		New-I	0.06	0.19	0.31	0.50
		Sim-I	0.07	0.20	0.34	0.55
		Page	0.08	0.19	0.30	0.46
		Kimura86	0.06	0.16	0.25	0.37
		Kimura91	0.06	0.18	0.29	0.44
	0.9	Exact	1.44	4.04	6.59	10.40
		New-I	1.44	4.10	6.77	10.76
		Sim-I	1.45	4.12	6.78	10.79
		Page	1.55	4.03	6.51	10.24
		Kimura86	1.51	3.92	6.22	9.47
		Kimura91	1.46	4.04	6.49	9.94
0.5	0.5	Exact	0.01	0.01	0.02	0.03
		New-I	0.01	0.02	0.03	0.04
		Sim-I	0.01	0.03	0.04	0.07
		Page	0.02	0.02	0.03	0.04
		Kimura86	0.00	0.00	0.01	0.01
		Kimura91	0.00	0.01	0.01	0.01
	0.7	Exact	0.18	0.36	0.50	0.72
		New-I	0.17	0.38	0.59	0.91
		Sim-I	0.19	0.43	0.67	1.03
		Page	0.23	0.39	0.55	0.79
		Kimura86	0.14	0.27	0.38	0.53
		Kimura91	0.17	0.32	0.45	0.61
	0.9	Exact	2.77	5.50	8.10	11.97
		New-I	2.74	5.62	8.50	12.83
		Sim-I	2.76	5.68	8.59	12.96
		Page	2.89	5.56	8.24	12.25
		Kimura86	2.73	5.31	7.76	11.22
		Kimura91	2.79	5.48	8.02	11.59

Table 3: A Comparison of Approximations of the Mean Queue Length for $Ph/Ph/10$ Queues with $c_a^2 \geq 1$.

c_a^2	ρ	Method	$c_s^2 = 0.5$	$c_s^2 = 1.5$	$c_s^2 = 2.5$	$c_s^2 = 4.0$
1.0	0.5	Exact	0.03	0.04	0.05	0.06
		New-I	0.03	0.04	0.06	0.07
		New-II	0.03	0.04	0.05	0.05
	0.7	Exact	0.41	0.60	0.76	0.99
		New-I	0.40	0.63	0.86	1.19
		New-II	0.41	0.61	0.77	0.96
	0.9	Exact	4.58	7.37	10.02	13.94
		New-I	4.56	7.47	10.38	14.74
		New-II	4.59	7.40	10.05	13.72
2.0	0.5	Exact	0.09	0.10	0.10	0.11
		New-I	0.09	0.10	0.11	0.12
		New-II	0.09	0.10	0.10	0.10
		Page	0.06	0.08	0.11	0.14
		Kimura91	0.06	0.08	0.09	0.11
	0.7	Exact	0.87	1.08	1.25	1.49
		New-I	0.88	1.10	1.33	1.66
		New-II	0.89	1.08	1.23	1.39
		Page	0.76	1.12	1.47	2.00
		Kimura91	0.80	1.02	1.24	1.58
	0.9	Exact	8.15	11.02	13.74	17.75
		New-I	8.16	11.09	14.01	18.39
		New-II	8.19	10.99	13.57	17.07
		Page	7.90	11.31	14.70	19.81
		Kimura91	7.96	10.88	13.79	18.17
4.0	0.5	Exact	0.23	0.24	0.24	0.24
		New-I	0.24	0.24	0.24	0.24
		New-II	0.24	0.24	0.24	0.24
		Page	0.12	0.17	0.21	0.28
		Kimura91	0.13	0.14	0.16	0.18
	0.7	Exact	1.96	2.15	2.31	2.56
		New-I	1.99	2.14	2.29	2.51
		New-II	2.00	2.13	2.23	2.34
		Page	1.48	2.09	2.70	3.61
		Kimura91	1.57	1.80	2.02	2.36
	0.9	Exact	15.65	18.49	21.26	25.35
		New-I	15.72	18.46	21.21	25.32
		New-II	15.75	18.37	20.75	23.96
		Page	14.59	18.96	23.33	29.89
		Kimura91	14.76	17.67	20.59	24.97

Table 3 compares four approximations with the exact values of the mean queue length for $Ph/Ph/10$ queues with highly variable interarrival times. In Table 3, we omit Kimura86 from comparisons because it becomes unstable when $c_a^2 > 1$. In addition, the approximation of Page (Kimura91) for $c_a^2 = 1$ is excluded from the table because it coincides with New-I (New-II) for the $M/G/s$ queue. Table 3 shows that both of New-I and New-II are more accurate than the two-moment approximations when $c_a^2 > 1$. When $c_a^2 = 1$, the quality of New-II (= Kimura86 = Kimura91) is quite excellent. Table 3 also shows that New-II performs as well as New-I except for $c_a^2 = c_s^2 = 4$, both providing satisfactory accuracy for practical applications.

Table 4: A Comparison of Approximations of the Mean Queue Length for $H_2^b/H_2^b/s$ Queues with $c_s^2 = 4$.

ρ	Method	$c_a^2 = 1.5$			$c_a^2 = 2.0$		
		$s = 2$	$s = 5$	$s = 10$	$s = 2$	$s = 5$	$s = 10$
0.3	Exact	0.15	0.02	—	0.17	0.03	—
	New-I	0.18	0.03	—	0.21	0.03	—
	New-II	0.20	0.02	—	0.30	0.03	—
	Page	0.20	0.03	0.00	0.26	0.03	0.00
	Kimura91	0.16	0.02	0.00	0.18	0.03	0.00
0.5	Exact	0.86	0.31	0.08	0.98	0.37	0.11
	New-I	0.98	0.38	0.10	1.15	0.46	0.12
	New-II	1.07	0.32	0.08	1.48	0.41	0.10
	Page	1.07	0.41	0.11	1.35	0.53	0.14
	Kimura91	0.91	0.35	0.09	1.01	0.40	0.11
0.7	Exact	3.59	2.22	1.24	4.00	2.56	1.49
	New-I	3.85	2.50	1.43	4.36	2.86	1.66
	New-II	4.06	2.26	1.17	4.83	2.65	1.39
	Page	4.05	2.69	1.59	4.79	3.28	2.00
	Kimura91	3.68	2.39	1.39	4.05	2.68	1.58
0.9	Exact	20.94	18.40	15.85	22.98	20.39	17.75
	New-I	21.36	19.03	16.63	23.50	20.99	18.39
	New-II	21.64	18.42	15.45	23.91	20.33	17.07
	Page	21.69	19.54	17.29	24.29	22.11	19.81
	Kimura91	21.06	18.80	16.47	23.03	20.64	18.17

To see the differences between New-I and New-II more clearly, we compare them with the two-moment approximations and the exact values of the mean queue length for some $H_2^b/H_2^b/s$ queues with $c_s^2 = 4$ in Table 4. Table 4 indicates that New-II becomes more accurate than New-I as $c_a^2 \to 1$. Table 4 also indicates that New-II tends to underestimate the exact value as s grows, and hence New-II tends to be less accurate than New-I, especially when $c_a^2 = 2$. Although it is relatively difficult to specify the region of c_a^2 where New-II surpasses New-I, a practical guideline for this region is that $1 \leq c_a^2 < 2$ if s is not too large, e.g., $s \leq 20$.

Table 5: Recommended Approximations for $EW(GI/G/s)$.

	$0 \le c_s^2 < 1$	$c_s^2 \ge 1$
$0 < c_a^2 < 0.5$	New-I or Kimura91	New-I
$0.5 \le c_a^2 < 1$		Page
$c_a^2 = 1$	New-II = Kimura86 = Kimura91	
$1 < c_a^2 < 2$	New-I or New-II	New-II
$c_a^2 \ge 2$	New-I	

Table 5 summarizes the results of Tables 2–4 and of the other numerical experiments, which gives a list of recommended approximations for $EW(GI/G/s)$. Of course, this recommendation is valid only for a subclass of $Ph/Ph/s$ queues, because we have not checked all of $GI/G/s$ queues due to the lack of authorized numerical results. Further numerical validation is necessary for more precise evaluation of our approximations. For almost all combinations of c_a^2 and c_s^2, the approximations provided in this paper are better than the two-moment approximations. The only exception is the case that $0.5 \le c_a^2 < 1$ and $c_s^2 > 1$, in which Page's approximation performs better than our approximations. It is one of the subjects of our future studies to improve the accuracy for this case, e.g., by refining the weighting coefficients ω and ν.

3. Simplified Formulas

To evaluate the approximations (2.11) and (2.12) actually, it is of course necessary to compute the mean waiting times in the building-block systems, i.e., the $GI/M/s$ and $GI/D/s$ queues. For $EW(GI/M/s)$ with given A, s and ρ, one can compute its value in a stable way through an explicit formula [2, pp. 267–273]. Neuts [21] provided an efficient algorithm for computing $EW(Ph/M/s)$. However, for $EW(GI/D/s)$ with a general interarrival-time cdf, it is *not* so easy to compute the exact value. Algorithms for computing $EW(GI/D/s)$ are available only for some special phase-type interarrival-time cdf's; see [20, 29].

In this section, to let our approximations be more tractable for systems with $c_a^2 \le 1$, we approximate $EW(GI/M/s)$ and $EW(GI/D/s)$ by the mean waiting times for more basic systems (e.g., the $M/M/s$ queue and so on) for which it is easy to compute the mean waiting times or extensive tables have been prepared. As shown in Section 2, the approximation (2.12) becomes unstable when $c_a^2 < 1$ and $c_s^2 > 1$. Hence, we are concerned only with the simplification of (2.11).

3.1 Approximating $EW(GI/M/s)$

Although approximating $EW(GI/M/s)$ is less important than $EW(GI/D/s)$, it is very useful if one can obtain approximate values for $EW(GI/M/s)$ only by using basic queueing tables. For this purpose, approximations of interpolating $EW(M/M/s)$ and $EW(D/M/s)$ are appropriate, because exact values for these mean waiting times for given s and ρ can be found in some queueing tables [9, 23, 24]. There are some approximations for $EW(GI/M/s)$ as special cases of interpolation approximations for $EW(GI/G/s)$: Page's approximation (2.17) for $EW(GI/M/s)$ can be written as a simple linear interpolation

$$EW(GI/M/s) \simeq c_a^2 EW(M/M/s) + (1 - c_a^2) EW(D/M/s), \tag{3.1}$$

and Kimura's approximation (2.18) for $EW(GI/M/s)$ can be considered as another interpolation which is a dual of (3.1) in some sense; cf. (2.19). Cosmetatos [3] has derived more accurate but complicated formulas which need the exact value of $EW(GI/M/1)$. From some numerical tests, we saw that Page's approximation (3.1) has satisfactory accuracy for practical applications, though it is not uniformly more accurate than the others. We are greatly concerned with the simplicity of the approximation, not with the relative percentage errors in light traffic if the absolute differences are small. Hence, we adopt (3.1) as our approximation for $EW(GI/M/s)$.

To let (3.1) be more tractable, we will further approximate $EW(D/M/s)$ by using $EW(M/M/s)$. Kimura [11, 14] suggested approximating $EW(D/M/s)$ by

$$EW(D/M/s) \simeq \frac{k_{01}}{2}\phi(s,\rho)EW(M/M/s), \tag{3.2}$$

which is obtained by combining the approximations of Cosmetatos [4] and Krämer and Langenbach-Belz [18], where k_{01} is defined by (2.21) and

$$\phi(s,\rho) = 1 - 4\min\left\{f(s)g(\rho),\ 0.25(1-10^{-6})\right\} \tag{3.3}$$

with

$$f(s) = \frac{(s-1)(\sqrt{4+5s}-2)}{16s}, \tag{3.4}$$

$$g(\rho) = \frac{1-\rho}{\rho}. \tag{3.5}$$

Note that we have slightly modified the original approximation in [4] by inserting the minimum with $0.25(1-10^{-6})$ in (3.3). Without it, the approximation (3.2) becomes negative and hence meaningless; cf. Tijms [28, (4.228)] and Kimura [14, Equations (32) and (33)]. From some numerical tests, we saw that (3.2) performs well unless ρ is close to zero. Combining (3.1) and (3.2), we obtain the final form of our simplified formula for $EW(GI/M/s)$ as

$$EW(GI/M/s) \simeq \left\{c_a^2 + \frac{k_{01}}{2}(1-c_a^2)\phi(s,\rho)\right\} EW(M/M/s). \tag{3.6}$$

NUMERICAL COMPARISONS

Table 6 compares the approximations (3.1) and (3.6) with the exact values of the mean queue length for $Ph/M/s$ queues. The exact values for $c_a^2 = 0.1, 0.25$ and 0.5 are computed via the analytic solution for the $E_m/M/s$ queue ($m = 10, 4$ and 2, respectively), while those for $c_a^2 = 0.8$ are quoted from the tables of Seelen et al. [24]. In Table 6, we refer to (3.6) as "Sim-Page". Table 6 shows that the simplified approximation (3.6) is stably accurate for various values of $c_a^2 \in (0,1)$ and s, and hence good enough for practical applications. It is interesting that (3.6) is more accurate than (3.1), due to underestimation of (3.2).

Remark 3.5 Of course, there is no problem to apply (3.6) to approximating the mean waiting times for queues with $c_a^2 > 1$. From some numerical comparisons, we saw that (3.6) is also accurate for $c_a^2 > 1$.

Table 6: A Comparison of Approximations of the Mean Queue Length for $Ph/M/s$ Queues.

s	ρ	Method	$c_a^2 = 0.1$	$c_a^2 = 0.25$	$c_a^2 = 0.5$	$c_a^2 = 0.8$
10	0.7	Exact	0.126	0.177	0.277	0.415
		Page	0.138	0.201	0.307	0.433
		Sim-Page	0.137	0.200	0.306	0.433
	0.8	Exact	0.583	0.740	1.021	1.383
		Page	0.600	0.773	1.061	1.406
		Sim-Page	0.600	0.772	1.061	1.406
	0.9	Exact	2.773	3.293	4.181	5.276
		Page	2.791	3.329	4.223	5.301
		Sim-Page	2.781	3.320	4.220	5.299
50	0.8	Exact	0.052	0.082	0.151	0.260
		Page	0.067	0.114	0.192	0.285
		Sim-Page	0.054	0.103	0.185	0.283
	0.9	Exact	1.148	1.460	2.024	2.758
		Page	1.185	1.533	2.114	2.810
		Sim-Page	1.177	1.526	2.110	2.809
100	0.9	Exact	0.525	0.715	1.080	1.583
		Page	0.566	0.797	1.187	1.644
		Sim-Page	0.552	0.786	1.175	1.641
	0.95	Exact	3.997	4.868	6.388	8.303
		Page	4.056	4.984	6.530	8.386
		Sim-Page	4.029	4.962	6.515	8.380

3.2 Approximating $EW(GI/D/s)$

It is difficult to obtain an explicit expression for the mean waiting time in the $GI/D/s$ queue with a general interarrival-time cdf A, except that A is the exponential distribution. In this subsection, we approximate $EW(GI/D/s)$ by combining the mean waiting times for two $M/M/c$ queues with different numbers of servers. There are three steps in our approximation: We first approximate $EW(GI/D/s)$ by $EW(E_m/D/s)$'s; we then utilize the equivalence between $EW(E_m/D/s)$ and $EW(M/D/ms)$; and finally we approximate $EW(M/D/s)$ by $EW(M/M/s)$.

For given c_a^2 (≤ 1) of the cdf A, there exists an integer m (≥ 1) such that

$$\frac{1}{m+1} < c_a^2 \leq \frac{1}{m}. \tag{3.7}$$

We approximate $EW(GI/D/s)$ by linearly interpolating $EW(E_m/D/s)$ and $EW(E_{m+1}/D/s)$ for m in (3.7), i.e.,

$$EW(GI/D/s) \simeq q_m EW(E_m/D/s) + (1 - q_m)EW(E_{m+1}/D/s), \tag{3.8}$$

where the interpolation coefficient $q_m \in (0, 1]$ is given by

$$q_m = m\{(m+1)c_a^2 - 1\}. \tag{3.9}$$

Using the fact that $EW(E_m/D/s) = EW(M/D/ms)$ which has been proved by Iversen [10], we can rewrite (3.8) as

$$EW(GI/D/s) \simeq q_m EW(M/D/ms) + (1 - q_m)EW(M/D/(m+1)s). \tag{3.10}$$

The mean waiting times for the building-block systems in (3.10) can be found in some queueing tables or can be computed by an explicit formula provided by Crommelin [7]. It is, however, known that the calculation by Crommelin's formula tends to be unstable when s grows or ρ tends to unity; see [6, 12]. To avoid this unstableness, Kimura [15] has recently suggested approximating $EW(M/D/s)$ by

$$EW(M/D/s) \simeq r_s EW(M/M/s), \tag{3.11}$$

as a refinement of Cosmetatos' approximation in [4], where $r_s \equiv r_s(\rho)$ is

$$r_s(\rho) = \frac{1}{2}\{1 + f(s)g(\rho)h(s,\rho)\} \tag{3.12}$$

for $f(s)$ in (3.4) and $g(\rho)$ in (3.5). The bivariate function $h(s,\rho)$ is introduced to correct some defects of the original approximation for large s and small ρ, which is given by

$$h(s,\rho) = \xi(s, a(\rho))\eta(b(s),\rho) \tag{3.13}$$

with

$$\xi(s,x) = \sqrt{1 - \exp\left(-\frac{2x}{s-1}\right)}, \quad x \geq 0, \tag{3.14}$$

$$\eta(y,\rho) = 1 - \exp\left(-\frac{\rho y}{1-\rho}\right), \quad y \geq 0. \tag{3.15}$$

In (3.13), the functions $a(\rho)$ and $b(s)$ are defined by

$$a(\rho) = \frac{25.6}{\{g(\rho)\eta(\beta,\rho)\}^2}, \tag{3.16}$$

and

$$b(s) = \frac{s-1}{(s+1)f(s)\xi(s,\alpha)}, \tag{3.17}$$

respectively, where α and β are arbitrary positive constants satisfying the relation

$$\alpha\beta^2 = 25.6. \tag{3.18}$$

It has been checked from many numerical experiments in [15] that $\alpha = 2.2$ (and hence $\beta = \sqrt{25.6/2.2}$) is an optimal value for the best performance of the approximation. It is shown in [15] that the approximation (3.11) is exact for $s = 1$ and asymptotically exact as $s \to \infty$ or $\rho \to 0$ or $\rho \to 1$. The relative percentage errors of (3.11) are less than 1% for most (s,ρ) combinations. By the use of (3.10), we obtain the approximation

$$EW(GI/D/s) \simeq q_m r_{ms} EW(M/M/ms) + (1 - q_m)r_{(m+1)s}EW(M/M/(m+1)s), \tag{3.19}$$

with m satisfying (3.7).

Remark 3.6 When the value of ms is small, it is sufficient to use $h(s,\rho) \equiv 1$ in the approximation (3.11) with (3.12). However, if ms is large, (3.19) fairly overestimates the true value unless the correcting function $h(s,\rho)$ is used. It should be noted that the value of ms increases quite rapidly as s grows for small c_a^2; e.g., consider the case with $c_a^2 = 0.1$.

NUMERICAL COMPARISONS

Table 7 compares the approximation (3.19) (referred as "New") with the exact values of the mean queue length for $Ph/D/s$ queues. All of the exact values are quoted from [24]. Table 7 shows that the quality of (3.19) is quite excellent.

Table 7: A Comparison of Approximations of the Mean Queue Length for $Ph/D/s$ Queues.

s	ρ	Method	$c_a^2 = 0.1$	$c_a^2 = 0.25$	$c_a^2 = 0.5$	$c_a^2 = 0.8$
10	0.7	Exact	0.000	0.009	0.067	0.198
		New	0.000	0.008	0.066	0.202
	0.8	Exact	0.005	0.073	0.287	0.639
		New	0.005	0.072	0.287	0.643
	0.9	Exact	0.111	0.498	1.299	2.380
		New	0.111	0.502	1.306	2.388
50	0.8	Exact	0.000	0.001	0.027	0.124
		New	0.000	0.001	0.026	0.137
	0.9	Exact	0.008	0.130	0.555	1.262
		New	0.007	0.128	0.557	1.298
100	0.9	Exact	—	0.036	0.260	0.732
		New	—	0.034	0.256	0.771
	0.95	Exact	—	0.594	1.883	3.753
		New	—	0.597	1.904	3.841

3.3 Accuracy of the Simplified Approximation

Substituting (3.6) and (3.19) into (2.11), we obtain the simplified formula for $EW(GI/G/s)$, which is a weighted combination of the mean waiting times for three $M/M/c$ queues ($c = s, ms$ and $(m+1)s$). In Table 2, we have given the simplified approximation (referred as "Sim-I") for some $Ph/Ph/10$ queues. Table 2 shows that Sim-I performs as well as New-I. This indicates that the approximations (3.6) and (3.19) have good quality when they are combined. For $c_a^2 < 1$, New-I in Table 5 can be replaced by Sim-I.

4. Concluding Remark

Almost no exact results have been known for the steady-state probabilities in the $GI/G/s$ queue. However, some invariance relations among characteristic quantities in general queues have been derived by the theory of point processes; see, e.g., Franken et al. [8] and Miyazawa [19]. Combining these invariance relations with our approximations for EW, we can derive simple approximations for the steady-state probabilities in the $GI/G/s$ queue. Extensions of our approximations to this direction are in progress and will be reported elsewhere.

Acknowledgments

I am grateful to the referees for their helpful suggestions. This research was supported in part by the Grants in Aid for Scientific Research of the Japanese Ministry of Education, Science and Culture under the Contracts No. 62302059 (1987–1989) and No. 63780017 (1988–1989), and by the Special Grant-in-Aid for Promotion of Education and Science in Hokkaido University provided by the Japanese Ministry of Education, Science and Culture (1987–1988).

References

[1] BURMAN, D.Y. AND SMITH, D.R., "A light-traffic theorem for multi-server queues," *Mathematics of Operations Research*, **8** (1983), 15–25.

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

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

[4] COSMETATOS, G.P., "Approximate explicit formulae for the average queueing time in the processes $(M/D/r)$ and $(D/M/r)$," *INFOR*, **13** (1975), 328–332.

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

[6] COSMETATOS, G.P., "On the implementation of Page's approximation for waiting times in general multi-server queues," *Journal of the Operational Research Society*, **33** (1982), 1158–1159.

[7] CROMMELIN, C.D., "Delay probability formulae," *P.O. Elec. Engrs.*, **26** (1934), 266–274.

[8] FRANKEN, P., KÖNIG, D., ARNDT, U., AND SCHMIDT, V., *Queues and Point Processes*, Akademie-Verlag, Berlin, 1981.

[9] HILLIER, F.S., AND YU, O.S., *Queueing Tables and Graphs*, North-Holland, New York, 1981.

[10] IVERSEN, V.B., "Decomposition of an $M/D/r \cdot k$ queue with FIFO into k $E_k/D/r$ queues with FIFO," *Operations Research Letters*, **2** (1983), 20–21.

[11] KIMURA, T., "A two-moment approximation for the mean waiting time in the $GI/G/s$ queue," *Management Science*, **32** (1986), 751–763.

[12] KIMURA, T., "Approximations for the mean delay in the $M/D/s$ queue," *Proceedings of the Seminar on Queueing Theory and Its Applications*, Kyoto, pp. 173–184, 1987.

[13] KIMURA, T., "Heuristic approximations for the mean delay in the $GI/G/s$ queue," *Economic Journal of Hokkaido University*, **16** (1987), 87–98.

[14] KIMURA, T., "Approximations for the waiting time in the $GI/G/s$ queue," *Journal of the Operations Research Society of Japan*, **34** (1991), 173–186.

[15] KIMURA, T., "Refining Cosmetatos' approximation for the mean waiting time in the $M/D/s$ queue," *Journal of the Operational Research Society*, **42** (1991), 595–603.

[16] KIMURA, T., "Approximating the mean waiting time in the $GI/G/s$ queue," *Journal of the Operational Research Society*, **42** (1991), 959–970.

[17] KÖLLERSTRÖM, J., "Heavy traffic theory for queues with several servers. I," *Journal of Applied Probability*, **11** (1974), 544–552.

[18] KRÄMER, W., AND LANGENBACH-BELZ, M., "Approximate formulae for the delay in the queueing system $GI/G/1$," *Proceedings of the 8th International Teletraffic Congress*, Melbourne, 1976, pp. 235–1/8.

[19] MIYAZAWA, M., "A formal approach to queueing processes in the steady state and their applications," *Journal of Applied Probability*, **16** (1979), 332–346.

[20] NEUTS, M.F., "The c-server queue with constant service times and a versatile Markovian arrival process," *Applied Probability – Computer Science, The Interface*, Vol. I, R.L. Disney and T.J. Ott (eds.), Birkhäuser, Boston, pp. 31–70, 1982.

[21] NEUTS, M.F., "Explicit steady-state solutions to some elementary queueing models," *Operations Research*, **30** (1982), 480–489.

[22] PAGE, E., *Queueing Theory in OR*, Butterworth, London, 1972.

[23] PAGE, E., "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** (1982), 453–473.

[24] SEELEN, L.P., TIJMS, H.C., AND VAN HOORN, M.H., *Tables for Multi-Server Queues*, North-Holland, Amsterdam, 1985.

[25] SEELEN, L.P., "An algorithm for $Ph/Ph/c$ queues," *European Journal of Operational Research*, **23** (1986), 118-127.

[26] SHORE, H., "Simple approximations for the $GI/G/c$ queue – I: the steady-state probabilities," *Journal of the Operational Research Society*, **39** (1988), 279–284.

[27] SHORE, H., "Simple approximations for the $GI/G/c$ queue – II: the moments, the inverse distribution function and the loss function of the number in the system and of the queue delay," *Journal of the Operational Research Society*, **39** (1988), 381–391.

[28] TIJMS, H.C., *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley & Sons, New York, 1986.

[29] VAN HOORN, M.H., "Numerical analysis of multi-server queues with deterministic service and special phase-type arrivals," *Zeitschrift für Operations Research*, **30** (1986), A15–A28.

[30] WU, J.-S. AND CHAN, W.C., "Maximum entropy analysis of multi-server queueing systems," *Journal of the Operational Research Society*, **40** (1989), 815–825.

Toshikazu KIMURA
Department of Business Administration
Faculty of Economics
Hokkaido University
Nishi 7, Kita 9, Kita-ku
Sapporo 060
Japan