Journal of the Operations Research
Society of Japan
Vol. 34, No. 2, June 1991

APPROXIMATIONS FOR THE WAITING TIME IN THE $GI/G/s$ QUEUE

Toshikazu Kimura
Hokkaido University

(Received March 12, 1990; Revised July 18, 1990)

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

1. Introduction and Summary

In this paper we provide some two-moment approximation formulas for the mean waiting time and the delay probability 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 using those for analyzable systems such as $M/M/s$, $M/D/s$ and $D/M/s$ queues. In addition, combining these approximations with the two-moment approximations for the conditional mean waiting time provided in [21], we approximate the delay probability in the $GI/G/s$ queue.

For the $M/G/s$ queue, there are several elaborate approximations which depend on the service-time distribution [11, 17, 28]. These distribution-dependent approximations often have the advantage of producing the entire waiting time distribution and its higher moments, since one can utilize some known analytical results to refine approximations for the $M/G/s$ queue; cf. [2, 32]. However, for the $GI/G/s$ queue, such results have not been available as yet in the absence of an exact analysis. Thus, possible approaches to the $GI/G/s$ case are quite limited and are essentially heuristic by nature.

Great progress is currently being made on computational methods for obtaining exact solutions of $GI/G/s$ queues; see, e.g., [23, 24, 25, 26] and references therein. For some applications, these methods will eliminate the need for approximations. However, simple closed-form two-moment formulas will still be desired for other applications, e.g., when $GI/G/s$ models appear as submodels in large-scale queueing systems. It is helpful to have simple approximations as concise summaries.

Let W denote the waiting time before beginning service and let EW be its expected value, assuming that the system is stable and in steady state. We write $EW(M/M/s)$ to indicate EW for the $M/M/s$ queue and so forth. Let u and v be generic interarrival time and service time, respectively; let $\rho = Ev/sEu \in [0, 1)$ 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). Then, among approximation formulas we provide in this paper, a pair of approximations we

recommend to use is as follows: For $c_a^2 \leq 1$,

$$EW(GI/G/s) \simeq k(c_a^2 + c_s^2)\left(\frac{2(c_a^2 + c_s^2 - 1)}{EW(M/M/s)} + \frac{1 - c_s^2}{EW(M/D/s)} + \frac{k_0(1 - c_a^2)}{EW(D/M/s)}\right)^{-1}, \quad (1)$$

and for $c_a^2 > 1$,

$$EW(GI/G/s) \simeq$$
$$(c_a^2 + c_s^2 - 1)EW(M/M/s) + (1 - c_s^2)EW(M/D/s) + \frac{1}{k_0}(1 - c_a^2)EW(D/M/s), \quad (2)$$

where

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

$$k_0 = k(\rho, 0, 1) = \exp\left\{-\frac{2(1-\rho)}{3\rho}\right\}. \quad (4)$$

Our studies indicate that (1) and (2) will usually yield satisfactory approximations (in the order of 10% relative error), *provided* that (i) the valiability parameters c_a^2 and c_s^2 (especially c_a^2) are not too large, e.g., $c_a^2 \leq 2$ and $c_s^2 \leq 4$, and (ii) the traffic intensity ρ is not too small, e.g., $\rho \geq 0.3$ for $s = 2$ and $\rho \geq 0.8$ for $s = 20$. In particular, the approximation (1) has excellent performance when $c_a^2 \leq 1$ and $c_s^2 \leq 2.5$; the relative percentage errors are in the order of 5% when $\rho = 0.5$ and in the order of 1% when $\rho = 0.9$ for almost all cases satisfying the condition (ii) in our numerical experiments. In other words, we can roughly say that the relative percentage error is in the order of 1% if the approximate value of EW is greater than $10Ev$. The studies also indicate that the accuracy of our approximations does not so strongly depend on the number of servers if c_a^2 is not too large. This property is practically important because the computational methods for exact solutions become difficult to carry out for cases with large s. Theorems and numerical examples in this paper will help clarify these points.

In (1) and (2), the mean waiting times for the building-block systems, i.e., the $M/M/s$, $M/D/s$ and $D/M/s$ queues, have the same mean service times and traffic intensities as those of the queue in question. The exact values of these mean waiting times can be obtained either by computing their analytical solutions or by using some queueing tables [10, 19, 22]. We should note that data of the building-block systems required for computing our approximations can be considerably reduced by using some interpolation techniques in [22, pp. 12–14]. In Section 3, we further propose simple closed-form formulas in which only $EW(M/M/s)$ is used as their building blocks.

We see that the approximation (1) (and also (2)) is exact for the $M/M/s$, $M/D/s$, $D/M/s$ and $M/G/1$ queues. Hence the approximation (1) is an *interpolation approximation* among these systems when $c_a^2 \leq 1$ and $c_s^2 \leq 1$. It will be shown that (1) and (2) perform very well as extrapolation approximations when $c_a^2 > 1$ or $c_s^2 > 1$.

The approximations (1) and (2) are *two-moment approximations* for $EW(GI/G/s)$, i.e., they depend only on the first two moments of u and v. Closely related two-moment approximations have been developed by Page [18] and Kimura [13], in which three exact mean waiting times for the $M/M/s$, $M/D/s$ and $D/M/s$ queues are also used as their building blocks; see (27) and (28). We will see in Section 2 that the approximations of Page and Kimura can be produced by using our approach as its special cases. Other simple two-moment approximations for $EW(GI/G/s)$ can be found in [20, 29]. This paper shows that (1) and (2) are much better than these approximations in both moderate and heavy traffic.

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 software packages such as QNA (Queueing Network Analyzer) which has been developed to calculate approximate congestion measures for networks of queues [12, 29]. Typically the arrival process at each node is not actually renewal, but the two-moment characterization can be viewed as an approximation by a renewal process. The idea in QNA is *not* to ignore the dependence among successive interarrival times, but to try to capture the essential properties of this dependence in the variability parameters c_a^2. Our approximation formulas can be used in QNA-like softwares to obtain several congestion measures for the whole network as well as each node if the departure process from a node can be well approximated by a renewal process; cf. [30].

This paper is organized as follows: In Section 2, we focus on a ratio of the mean waiting times for systems with different number of servers. We approximate this ratio by a linearly weighted sum of the corresponding ratios for the $M/M/s$, $M/D/s$ and $D/M/s$ queues. Using several sets of weights consistent with exact properties for particular cases, we derive four two-moment approximation formulas for $EW(GI/G/s)$. Combining these approximations with the approximations by Cosmetatos [5] and Seelen and Tijms [21], we provide simpler approximations for $EW(GI/G/s)$ in Section 3 and approximations for the delay probability in Section 4. In each section, we discuss the quality of the approximations by numerical comparisons for some particular cases.

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

A frequently used approach to obtain approximations for EW is to approximate a normalized mean waiting time instead of EW itself; see [1, 6, 13] for the $M/G/s$ case and [4] for the $GI/M/s$ case.

In this paper, we focus on the quantity $EW(GI/G/m)/EW(GI/G/n)$ $(m \neq n)$ for the $GI/G/s$ queue. This quantity denotes the ratio of the mean waiting times for two systems with different number of servers which have the same mean service times and traffic intensities as those of the approximating $GI/G/s$ queue. We approximate this ratio by a linearly weighted sum of the corresponding ratios for the $M/M/s$ $M/D/s$ and $D/M/s$ queues, i.e.,

$$\frac{EW(GI/G/m)}{EW(GI/G/n)} \simeq w_{11}\frac{EW(M/M/m)}{EW(M/M/n)} + w_{10}\frac{EW(M/D/m)}{EW(M/D/n)} + w_{01}\frac{EW(D/M/m)}{EW(D/M/n)}, \tag{5}$$

where $\{w_{c_a^2 c_s^2}\} = \{w_{11}, w_{10}, w_{01}\}$ denotes a set of weighting coefficients. In (5), the exact ratios for the building-block systems can be calculated in a numerically stable way for given s and ρ, or they can be found in some queueing tables. Thus we need to determine the weights to identify our approximation completely.

For the weights $\{w_{ij}\}$, we restrict their class to the function $w_{ij} \equiv w_{ij}(c_a^2, c_s^2)$ which depends only on the squared coefficients of variation of u and v, and not on m, n and ρ. From the consistency with the building-block systems, we immediately see that $\{w_{ij}\}$ satisfies the condition

$$\text{C1}: \quad w_{11}(1,1) = w_{10}(1,0) = w_{01}(0,1) = 1. \tag{6}$$

As a natural condition for the interpolation approximation, we assume

$$\text{C2}: \quad w_{10}(c_a^2, 1) = w_{01}(1, c_s^2) = 0. \tag{7}$$

This condition is essentially based on an idea that we approximate a general interarrival-time or service-time distribution by combining the exponential and deterministic distributions [1, 13, 18, 27].

Since the approximate relation (5) can be identified by c_a^2 and c_s^2, we simply denote the relation (5) by $R_{mn} \equiv R_{mn}(c_a^2, c_s^2)$ for convenience. Then we have

Theorem 2.1 *Assume that* $E[v^3] < \infty$. *Then, the approximate relation* $R_{mn}(c_a^2, c_s^2)$ *is asymptotically correct as* $\rho \to 1$ *if the condition*

$$\text{C3}: \quad w_{11} + w_{10} + w_{01} = 1 \tag{8}$$

holds.

Proof: By the heavy traffic limit theorem in [15], we have

$$\lim_{\rho \to 1}(1-\rho)EW(GI/G/s) = \frac{c_a^2 + c_s^2}{2s}Ev, \tag{9}$$

if $E[v^3] < \infty$. Multiplying both the denominators and numerators in the relation $R_{mn}(c_a^2, c_s^2)$ by the term $(1-\rho)$ and letting $\rho \to 1$ from below, we obtain the desired result. ∎

From Theorem 2.1, we assume that the condition C3 holds to ensure the accuracy of (5) in heavy traffic. It is difficult to obtain further useful properties of $\{w_{ij}\}$ from (5) for arbitrary m and n. Hence, we hereafter restrict the values of m and n to two cases with (i) $m = s$, $n = 1$ and (ii) $m = 1$, $n = s$, and we call the relations R_{s1} and R_{1s} as Type I and Type II relations, respectively.

Theorem 2.2 *For the M/G/s queue,*
(i) *the approximate relation* $R_{s1}(1, c_s^2)$ *is asymptotically correct as* $s \to \infty$ *if*

$$\text{C4}: \quad w_{11}(1, c_s^2) = \frac{2c_s^2}{1+c_s^2}, \quad w_{10}(1, c_s^2) = \frac{1-c_s^2}{1+c_s^2}, \tag{10}$$

(ii) *the approximate relation* $R_{1s}(1, c_s^2)$ *is asymptotically correct as* $s \to \infty$ *if*

$$\text{C5}: \quad w_{11}(1, c_s^2) = c_s^2, \quad w_{10}(1, c_s^2) = 1 - c_s^2. \tag{11}$$

Proof: For notational convenience, let $w_{11}(1, c_s^2) = \tilde{w}_{11}$ and $w_{10}(1, c_s^2) = \tilde{w}_{10}$ for a moment. It is clear from the conditions C2 and C3 that

$$\tilde{w}_{11} + \tilde{w}_{10} = 1. \tag{12}$$

Following Boxma et al. [1], we introduce the quantity

$$N_{Gs} = \frac{1+c_s^2}{2}\frac{EW(M/M/s)}{EW(M/G/s)}, \tag{13}$$

to investigate the asymptotic consistency of $R_{s1}(1, c_s^2)$ and $R_{1s}(1, c_s^2)$ as $s \to \infty$. From (5), (13) and $N_{G1} = 1$, we have the approximation for N_{Gs} as

$$N_{Gs} \simeq \begin{cases} \left(\tilde{w}_{11} + \tilde{w}_{10}N_{Ds}^{-1}\right)^{-1}, & \text{for } R_{s1}(1, c_s^2) \\ \tilde{w}_{11} + \tilde{w}_{10}N_{Ds}, & \text{for } R_{1s}(1, c_s^2), \end{cases} \tag{14}$$

where N_{Ds} is the quantity (13) for the $M/D/s$ queue. As shown in Remark 1 of [1], the quantity N_{Gs} satisfies

$$\lim_{s\to\infty} N_{Gs} = \frac{1+c_s^2}{2}, \tag{15}$$

which states the fact that the $M/G/\infty$ queue is *insensitive* to the service-time distribution. Letting $s \to \infty$ in (14) and using (12) and (15), we obtain (10) and (11). ∎

We now determine the weights $\{w_{ij}\}$ satisfying the conditions C1–C5. Unfortunately, the weights satisfying all of these conditions are not uniquely determined; cf. [13]. Hence, we restrict the weighting coefficients to those which are simple and symmetric with respect to c_a^2 and c_s^2, taking account of the symmetry in the heavy traffic result (9). As such weights, we propose the following: For Type I relation $R_{s1}(c_a^2, c_s^2)$,

$$\text{Case IA}: \quad w_{11} = \frac{2(c_a^2+c_s^2-1)}{c_a^2+c_s^2}, \quad w_{10} = \frac{1-c_s^2}{c_a^2+c_s^2}, \quad w_{01} = \frac{1-c_a^2}{c_a^2+c_s^2}, \tag{16}$$

$$\text{Case IB}: \quad w_{11} = \frac{2c_a^2c_s^2}{c_a^2+c_s^2}, \quad w_{10} = \frac{c_a^2(1-c_s^2)}{c_a^2+c_s^2}, \quad w_{01} = \frac{(1-c_a^2)c_s^2}{c_a^2+c_s^2}; \tag{17}$$

for Type II relation $R_{1s}(c_a^2, c_s^2)$,

$$\text{Case IIA}: \quad w_{11} = c_a^2+c_s^2-1, \quad w_{10} = 1-c_s^2, \quad w_{01} = 1-c_a^2, \tag{18}$$

$$\text{Case IIB}: \quad w_{11} = \frac{c_a^2c_s^2}{c_a^2+c_s^2-c_a^2c_s^2}, \quad w_{10} = \frac{c_a^2(1-c_s^2)}{c_a^2+c_s^2-c_a^2c_s^2}, \quad w_{01} = \frac{(1-c_a^2)c_s^2}{c_a^2+c_s^2-c_a^2c_s^2}. \tag{19}$$

Hereafter we simply call the approximate relation $R_{s1}(c_a^2, c_s^2)$ with the weights of Case IA the approximation IA and so forth.

Remark 2.1 For the $M/G/s$ and $GI/M/s$ queues, the weights of Case IA (IIA) coincide with the weights of IB (IIB). They also coincide with the weights appeared in a similar approximate relation of [4, 6].

Remark 2.2 As a heuristic extension of approximations for $EW(M/G/s)$ and $EW(GI/M/s)$ in [4, 6], Cosmetatos [7] derived a similar approximate relation for $EW(E_m/E_k/s)$ with the weighting coefficients of Case IB.

From the two different types of the approximate relation with the weights (16)–(19), we will derive some approximations for $EW(GI/G/s)$: From the relation of Type I, we have

$$EW(GI/G/s) \simeq EW(GI/G/1)\left(w_{11}\frac{EW(M/M/s)}{EW(M/M/1)} + w_{10}\frac{EW(M/D/s)}{EW(M/D/1)} + w_{01}\frac{EW(D/M/s)}{EW(D/M/1)}\right), \tag{20}$$

and from the relation of Type II, we have

$$EW(GI/G/s) \simeq EW(GI/G/1)\left(w_{11}\frac{EW(M/M/1)}{EW(M/M/s)} + w_{10}\frac{EW(M/D/1)}{EW(M/D/s)} + w_{01}\frac{EW(D/M/1)}{EW(D/M/s)}\right)^{-1}. \tag{21}$$

Therefore we have four different approximations for $EW(GI/G/s)$, i.e., (20) with the weights (16) or (17) and (21) with the weights (18) or (19). Clearly, these approximations contain

$EW(GI/G/1)$ for a single server queue with the same mean service time and traffic intensity as in the approximating $GI/G/s$ queue. It is, however, difficult to obtain the exact value of $EW(GI/G/1)$ except for some special cases, e.g., the $M/G/1$ case. Hence, to simplify (20) and (21), we replace three mean waiting times for the single server queues in (20) and (21) by a common two-moment approximation; see Remark 2.3. For such an approximation, we use the approximation provided in [29], which is

$$EW(GI/G/1) \simeq \frac{c_a^2 + c_s^2}{2} g EW(M/M/1), \tag{22}$$

where the coefficient $g \equiv g(\rho, c_a^2, c_s^2)$ is defined as

$$g(\rho, c_a^2, c_s^2) = \begin{cases} k(\rho, c_a^2, c_s^2), & \text{if } c_a^2 \leq 1 \\ 1, & \text{if } c_a^2 > 1, \end{cases} \tag{23}$$

for $k(\rho, c_a^2, c_s^2)$ in (3). We see that the approximation (22) together with (23) is the Krämer and Langenbach-Belz [16] approximation for $c_a^2 \leq 1$. Taking into account that (22) gives the exact results for $EW(M/M/1)$ and $EW(M/D/1)$, we obtain, from (20),

$$EW(GI/G/s) \simeq (c_a^2 + c_s^2) g \left(\frac{w_{11}}{2} EW(M/M/s) + w_{10} EW(M/D/s) + \frac{w_{01}}{k_0} EW(D/M/s) \right), \tag{24}$$

and from (21)

$$EW(GI/G/s) \simeq (c_a^2 + c_s^2) g \left(\frac{2w_{11}}{EW(M/M/s)} + \frac{w_{10}}{EW(M/D/s)} + \frac{k_0 w_{01}}{EW(D/M/s)} \right)^{-1}. \tag{25}$$

Remark 2.3 Instead of (22), it is possible to use the exact value for $EW(D/M/1)$ in (20) and (21). However, we can easily see that the resultant formulas are *not* exact for the $D/M/s$ queue. This is why we use the approximation (22) for $EW(D/M/1)$.

Remark 2.4 If we replace the mean waiting times for the single server queues in (20) and (21) by

$$EW(GI/G/1) \simeq \frac{c_a^2 + c_s^2}{2} EW(M/M/1), \tag{26}$$

then the approximation (20) with the weights of Case IB coincides with Page's [18] approximation

$$EW(GI/G/s) \simeq c_a^2 c_s^2 EW(M/M/s) + c_a^2(1 - c_s^2) EW(M/D/s) + (1 - c_a^2) c_s^2 EW(D/M/s), \tag{27}$$

and the approximation (21) with the weights of Case IIA coincides with Kimura's [13] approximation

$$EW(GI/G/s) \simeq (c_a^2 + c_s^2) \left(\frac{2(c_a^2 + c_s^2 - 1)}{EW(M/M/s)} + \frac{1 - c_s^2}{EW(M/D/s)} + \frac{1 - c_a^2}{EW(D/M/s)} \right)^{-1}. \tag{28}$$

Hence, we see that our approach unifies the above two-moment approximations for $EW(GI/G/s)$.

NUMERICAL COMPARISONS

Table 1 gives a list of 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, and $E_{1,2}$ ($E_{1,3}$) denotes a mixture of M and E_2 (E_3) which is defined in Groenevelt et al. [9]. A number of combinations of the parameters s, ρ, c_a^2 and c_s^2 are specified in the table. The exact mean waiting times for the systems in the first three rows in Table 1 can be found in [9], while those for all the other systems are given in Seelen et al. [22]. It should be noted that all of the exact results are not necessarily available for these systems. For example, the exact results 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.

Table 1: A List of Numerical Experiments.

System	c_a^2	c_s^2	s	ρ
$M/H_2^b/s$	1	1.5625, 2.25, 9		
$M/E_{1,2}/s$	1	0.5, 0.64, 0.75, 0.81	2(1)5,8,10(5)25	0.3,0.5,0.7,0.8,0.9,0.95
$M/E_{1,3}/s$	1	0.333, 0.4, 0.45, 0.5		
$M/E_2/s$	1	0.5		
$E_2/M/s$	0.5	1		
$E_{10}/M/s$	0.1	1		
$E_2/E_2/s$	0.5	0.5		
$H_2^b/E_2/s$	2,3,4	0.5	2(1)10(5)25	0.3,0.5,0.7,0.8,0.9,0.95
$H_2^b/M/s$	2,3,4	1		
$H_2^b/H_2^b/s$	2,3,4	1.5, 2.5, 4		
$M/H_2^b/s$	1	1.5, 2.5, 4		
$E_2/H_2^b/s$	0.5	1.5, 2.5, 4		

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

ρ	Method	$s = 2$	$s = 5$	$s = 10$	$s = 20$
0.5	Exact	0.74	0.24	0.05	—
	New	0.71	0.23	0.05	0.00
	Simplified	0.72	0.23	0.06	0.01
	Page	0.80	0.29	0.07	0.01
0.7	Exact	3.17	1.87	0.99	0.36
	New	3.11	1.82	0.96	0.35
	Simplified	3.15	1.83	0.96	0.36
	Page	3.31	2.10	1.19	0.47
0.9	Exact	18.87	16.40	13.94	10.99
	New	18.78	16.26	13.82	10.90
	Simplified	18.84	16.26	13.71	10.74
	Page	19.10	16.96	14.77	12.04

Table 3: A Comparison of Approximations of the Mean Queue Length for $PH/PH/2$ Queues.

c_a^2	ρ	Method	$c_s^2 = 0.5$	$c_s^2 = 1.0$	$c_s^2 = 1.5$	$c_s^2 = 2.5$	$c_s^2 = 4.0$
0.5	0.3	Exact	0.01	0.02	0.03	0.05	0.08
		New	0.01	0.02	0.03	0.04	0.06
		Page	0.02	0.03	0.04	0.05	0.07
	0.5	Exact	0.12	0.19	0.25	0.38	0.57
		New	0.12	0.19	0.26	0.38	0.54
		Page	0.14	0.20	0.25	0.37	0.53
	0.7	Exact	0.58	0.89	1.19	1.79	2.67
		New	0.60	0.91	1.22	1.80	2.63
		Page	0.62	0.90	1.18	1.73	2.57
	0.9	Exact	3.69	5.56	7.43	11.15	16.71
		New	3.74	5.62	7.48	11.19	16.64
		Page	3.75	5.58	7.40	11.04	16.51
2.0	0.3	Exact	0.08	0.10	0.11	0.14	0.17
		New	0.09	0.11	0.12	0.14	0.18
		Page	0.09	0.12	0.14	0.19	0.26
	0.5	Exact	0.43	0.53	0.61	0.76	0.98
		New	0.46	0.54	0.62	0.78	1.01
		Page	0.48	0.60	0.73	0.97	1.35
	0.7	Exact	1.72	2.09	2.42	3.07	4.00
		New	1.76	2.08	2.41	3.06	4.05
		Page	1.81	2.23	2.66	3.51	4.79
	0.9	Exact	9.69	11.65	13.56	17.35	22.98
		New	9.70	11.60	13.51	17.31	23.03
		Page	9.80	11.87	13.94	18.08	24.29
4.0	0.3	Exact	0.12	0.15	0.17	0.20	0.25
		New	0.18	0.20	0.21	0.24	0.28
		Page	0.18	0.23	0.28	0.37	0.52
	0.5	Exact	0.72	0.86	0.96	1.15	1.40
		New	0.88	0.95	1.03	1.19	1.42
		Page	0.92	1.14	1.35	1.79	2.43
	0.7	Exact	3.10	3.53	3.89	4.59	5.57
		New	3.23	3.56	3.89	4.54	5.52
		Page	3.39	4.01	4.64	5.88	7.76
	0.9	Exact	17.62	19.63	21.57	25.41	31.10
		New	17.55	19.45	21.36	25.17	30.88
		Page	17.86	20.26	22.66	27.45	34.65

Table 4: A Comparison of Approximations of the Mean Queue Length for $PH/PH/20$ Queues.

c_a^2	ρ	Method	$c_s^2 = 0.5$	$c_s^2 = 1.0$	$c_s^2 = 1.5$	$c_s^2 = 2.5$	$c_s^2 = 4.0$
0.5	0.7	Exact	0.06	0.09	0.12	0.16	0.23
		New	0.04	0.06	0.08	0.12	0.16
		Page	0.09	0.12	0.15	0.20	0.28
	0.9	Exact	2.20	3.32	4.33	6.29	9.19
		New	2.19	3.24	4.25	6.15	8.74
		Page	2.34	3.39	4.43	6.52	9.65
2.0	0.7	Exact	0.47	0.51	0.54	0.58	0.65
		New	0.37	0.41	0.45	0.54	0.66
		Page	0.34	0.42	0.49	0.64	0.86
	0.9	Exact	7.07	8.25	9.33	11.44	14.49
		New	6.78	7.96	9.14	11.50	15.04
		Page	6.65	8.10	9.55	12.46	16.81
4.0	0.7	Exact	1.28	1.28	1.28	1.29	1.34
		New	0.75	0.79	0.83	0.92	1.04
		Page	0.67	0.81	0.95	1.22	1.64
	0.9	Exact	14.13	15.24	16.32	18.45	21.60
		New	12.78	13.96	15.14	17.50	21.04
		Page	12.39	14.38	16.38	20.37	26.36

These experiments have clarified some qualitative properties of our approximations: The approximation IA is stably accurate even for highly variable interarrival-time or service-time distribution. The approximation IIA is much better than the others when $c_a^2 \leq 1$, but when $c_a^2 > 1$ it produces very bad approximations (e.g., negative). We observe that the approximations of Type II including (28) do not fit for cases with $c_a^2 > 1$. The approximation IB (IIB) is less accurate than IA (IIA) in moderate traffic, but performs about the same in heavy traffic. From these observations, we will use IIA (i.e., (1)) or IA (i.e., (2)) as a new approximation according as $c_a^2 \leq 1$ or $c_a^2 > 1$. We denote this approximation as "New" in Tables 2–5.

Table 2 compares three approximations with the exact values of the mean queue length (excluding customers in service) for $M/H_2^b/s$ queues with $c_s^2 = 4$. Approximations of the mean queue length can be derived from those of EW by using Little's formula. In Table 2, "Simplified" denotes a simplified version of New which will be discussed in Section 3. We add the closely-related approximation of Page (27) in the table. However, we omit Kimura's approximation (28) from comparisons, since it coincides with New for $M/G/s$ queues. Table 2 shows that New is sufficiently accurate for most practical applications. The relative percentage errors of New are less than 5% for $\rho = 0.5$ and less than 1% for $\rho = 0.9$ for $M/G/s$ queues with $c_s^2 \leq 4$; see also Tables 1–4 in [13].

Tables 3 and 4 compare the approximations with the exact values of the mean queue length for $PH/PH/2$ and $PH/PH/20$ queues, respectively. The interarrival-time (service-time) distribution is E_2 when c_a^2 (c_s^2) $= 0.5$ and H_2^b when c_a^2 (c_s^2) > 1. We again omit (28) from comparisons because it is less accurate than the others when $c_a^2 > 1$. Based on comparisons in Tables 3 and 4, we can conclude that the new approximation is stably more accurate

than Page's approximation especially for small s. For highly variable cases with $c_a^2 \leq 2$ and $c_s^2 \leq 4$, the new approximation has the reasonable accuracy for most applications; see the rough practical guideline in Section 1 for the use of the new approximation. However, when $c_a^2 > 2$ or $c_s^2 > 4$, the new approximation becomes relatively unreliable due to the fact that the set of possible exact values of EW, which is consistent with the first two moments of u and v, grows as c_a^2 or c_s^2 grows.

3. Simplified Formulas

In this section we simplify our approximations for EW to let them be more tractable. If we have extensive queueing tables containing the exact mean waiting times for the $M/M/s$, $M/D/s$ and $D/M/s$ queues with given s and ρ, it is easy to obtain our approximations. However, we usually need to calculate each of these means for the given parameters. Among these mean waiting times, $EW(M/M/s)$ can be easily calculated by a programmable desk calculator, while the others involve certain difficulties in their calculations: For $EW(M/D/s)$, the calculation tends to be unstable when s is large and ρ is close to one; for $EW(D/M/s)$, it has a little bit complicated form including a root of a transcendental equation. These numerical difficulties imply that it takes much time to calculate these means accurately.

To avoid these difficulties, we will express the approximations (24) and (25) in terms of $EW(M/M/s)$ and the first two moments of u and v. For this purpose, it is necessary to approximate $EW(M/D/s)$ and $EW(D/M/s)$ by using $EW(M/M/s)$. Cosmetatos [5] provided the following approximations:

$$EW(M/D/s) \simeq \frac{1}{2}\phi_{10}(s,\rho)EW(M/M/s) \tag{29}$$

$$EW(D/M/s) \simeq \frac{EW(D/M/1)}{EW(M/M/1)}\phi_{01}(s,\rho)EW(M/M/s), \tag{30}$$

where $\phi_{10}(s,\rho)$ and $\phi_{01}(s,\rho)$ are defined by

$$\phi_{10}(s,\rho) = 1 + \gamma(s,\rho) \tag{31}$$

$$\phi_{01}(s,\rho) = 1 - 4\gamma(s,\rho) \tag{32}$$

$$\gamma(s,\rho) = \min\left\{(1-\rho)(s-1)\frac{\sqrt{4+5s}-2}{16s\rho}, 0.25(1-10^{-6})\right\}. \tag{33}$$

Following Whitt [31], we have modified the approximations of Cosmetatos [5] by inserting the minimum with $0.25(1-10^{-6})$ in (33). Without it, the approximation (30) becomes negative and hence meaningless for $\gamma(s,\rho) > 0.25$; cf. Kimura [14].

For $EW(D/M/s)$, we can obtain a simpler approximation by inserting a certain approximation for $EW(D/M/1)$ into (30). In particular, if we use the Krämer and Langenbach-Belz approximation for $EW(D/M/1)$; see (22) and (23), then we have

$$EW(D/M/s) \simeq \frac{1}{2}k_0\phi_{01}(s,\rho)EW(M/M/s), \tag{34}$$

for k_0 in (4). From some numerical tests, we saw that the approximations (29) and (34) perform well unless ρ is close to zero.

Applying these approximations for $EW(M/D/s)$ and $EW(D/M/s)$ in (24) and (25), we obtain simplified formulas: From (24),

$$EW(GI/G/s) \simeq \frac{c_a^2+c_s^2}{2}g\left(w_{11} + w_{10}\phi_{10}(s,\rho) + w_{01}\phi_{01}(s,\rho)\right)EW(M/M/s) \tag{35}$$

with the weights given in (16) or (17); and from (25),

$$EW(GI/G/s) \simeq \frac{c_a^2 + c_s^2}{2} g \left(w_{11} + \frac{w_{10}}{\phi_{10}(s,\rho)} + \frac{w_{01}}{\phi_{01}(s,\rho)} \right)^{-1} EW(M/M/s) \tag{36}$$

with the weights given in (18) or (19).

NUMERICAL COMPARISONS

In Table 2, we have given the simplified approximations of the mean queue length for some $M/H_2^b/s$ queues. We use (35) or (36) according as $c_a^2 > 1$ or $c_a^2 \leq 1$ as "Simplified" in the table. Table 2 shows that Simplified performs as well as New. This indicates the excellence of the quality of the approximations for the $M/D/s$ and $D/M/s$ queues (i.e., (29) and (34)). We see from the other experiments that the simplified approximation has almost the same accuracy as New and is good enough for practical applications.

4. Delay Probability

We now focus on delay probability, $P(W > 0)$, i.e., the probability that an arriving customer has to wait before beginning service. There are considerable works on $M/G/s$ queues. For the $M/G/s$ queues, it is well known that the delay probability for the $M/M/s$ queue, i.e., the Erlang-C formula [3, p. 91], is usually an excellent approximation for other service-time distributions [17]. However, there have been relatively few works on approximations of $P(W > 0)$ for $GI/G/s$ queues with non-Poisson arrivals. In this section we approximate $P(W > 0)$ for the $GI/G/s$ queue by combining our approximations for EW with an approximation for the conditional mean waiting time $E(W \mid W > 0)$.

Let D be the conditional waiting time given that the server is busy, i.e., $D = (W \mid W > 0)$ and let ED be its expected value. Seelen and Tijms [21] proposed the following two-moment approximation for ED: For $c_a^2 \leq 1$ and $c_s^2 \leq 1$,

$$ED \simeq \left\{ \left(1 - \frac{\rho}{2} - \frac{\rho^2}{2} \right) \left(\frac{c_s^2}{s} + \frac{1 - c_s^2}{s+1} \right) + \frac{(1+\rho)(c_a^2 - 1) + (3\rho - \rho^3)(1 + c_s^2)}{4s(1-\rho)} \right\} Ev; \tag{37}$$

for $c_a^2 > 1$ or $c_s^2 > 1$,

$$ED \simeq \left\{ (1+\rho) \left(\frac{c_s^2}{s} + \frac{1 - c_s^2}{s+1} \right) + \frac{\rho^2(c_a^2 + c_s^2)}{2s(1-\rho)} \right\} Ev. \tag{38}$$

These approximations for ED have essentially been obtained by taking weighted combinations of the heavy-traffic and light-traffic results for ED and by making sure that the approximations are exact for the $M/G/1$ case. Extensive numerical experiments have shown that (37) and (38) are excellent approximations for ED.

Inserting these approximations for ED into the obvious relation

$$P(W > 0) = \frac{EW}{ED}, \tag{39}$$

we can obtain four different approximations for $P(W > 0)$ corresponding to our approximations for EW in Section 2.

NUMERICAL COMPARISONS

Table 5 compares our approximation and the $M/M/s$ approximation with the exact values of the delay probability for some $H_2^b/H_2^b/s$ queues with $c_a^2 = 2$ and $c_s^2 = 4$. "New" in

Table 5: A Comparison of Approximations of the Delay Probability for $H_2^b/H_2^b/s$ Queues ($c_a^2 = 2$, $c_s^2 = 4$).

ρ	Method	$s = 2$	$s = 5$	$s = 10$	$s = 20$
0.5	Exact	0.4156	0.1950	0.0682	—
	New	0.4489	0.2124	0.0630	0.0064
	$M/M/s$	0.3333	0.1304	0.0361	0.0037
0.7	Exact	0.6536	0.4770	0.3072	0.1533
	New	0.6963	0.5130	0.3195	0.1383
	$M/M/s$	0.4644	0.3778	0.2217	0.0936
0.9	Exact	0.8871	0.8146	0.7355	0.6312
	New	0.9106	0.8449	0.7556	0.6312
	$M/M/s$	0.8526	0.7625	0.6687	0.5508

Table 5 denotes the approximation of $P(W > 0)$ obtained by combining (37), (38) and New for EW. Table 5 indicates that the new approximation is satisfactory even for such highly variable cases as $H_2^b/H_2^b/s$ queues. Table 5 also indicates that the $M/M/s$ approximation for $P(W > 0)$ is not good enough for these cases. From the other numerical experiments, we see that New is much better than "$M/M/s$" except for $c_a^2 = 1$.

Acknowledgments

I am grateful to Dr. Ward Whitt of AT&T Bell Laboratories for his interest in this work, and 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).

References

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

[2] Burman, D.Y. and D.R. Smith, "A light-traffic theorem for multi-server queues," *Mathematics of Operations Research*, **8** (1983), 15–25.

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

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

[5] 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.

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

[7] Cosmetatos, G.P., "Some approximate equilibrium results for the multi-server queue $(E_m/E_k/r)$," *Opsearch*, **14** (1977), 108–117.

[8] 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.

[9] GROENEVELT, H., M.H. VAN HOORN AND H.C. TIJMS, "Tables for $M/G/c$ queuing systems with phase-type service," *European Journal of Operational Research*, **16** (1984), 257–269.

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

[11] HOKSTAD, P., "Approximations for the $M/G/m$ queue," *Operations Research*, **26** (1978), 510–523.

[12] KIMURA, T., "The queueing network analyzer: a survey (1)–(3)," [in Japanese] *Communications of the Operations Research Society of Japan*, **29** (1984), 366–371, 431–439, 494–500.

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

[14] 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.

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

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

[17] MIYAZAWA, M., "Approximations of the queue length distribution of an $M/GI/s$ queue by the basic equations," *Journal of Applied Probability*, **23** (1986), 443–458.

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

[19] 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.

[20] SAKASEGAWA, H., "An approximation formula $L_q \simeq \alpha\rho^{\beta}/(1-\rho)$," *Annals of the Institute of Statistical Mathematics*, **29** (1977), Part A, 67–75.

[21] SEELEN, L.P. AND H.C. TIJMS, "Approximations for the conditional waiting times in the $GI/G/c$ queue," *Operations Research Letters*, **3** (1984), 183–190.

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

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

[24] SUMITA, U. AND M. RIEDERS, "A new algorithm for computing the ergodic probability vector for large Markov chains: replacement process approach," *Probability in the Engineering and Information Sciences*, **4** (1990), 89–116.

[25] SUMITA, U. AND M. RIEDERS, "Application of the replacement process approach for computing the ergodic probability vector of large scale row-continuous Markov chains," *Journal of the Operations Research Society of Japan*, **33** (1990), 279–307.

[26] TAKAHASHI, Y. AND Y. TAKAMI, "A numerical method for the steady-state probabilities of a $GI/G/c$ queueing system in a general class," *Journal of the Operations Research Society of Japan*, **19** (1976), 147–157.

[27] TAKAHASHI, Y., "An approximation formula for the mean waiting time of an $M/G/c$ queue," *Journal of the Operations Research Society of Japan*, **20** (1977), 150–163.

[28] TIJMS, H.C., M.H. VAN HOORN AND A. FEDERGRUEN, "Approximations for the steady-state probabilities in the $M/G/c$ queue," *Advances in Applied Probability*, **13** (1981), 186–206.

[29] WHITT, W., "The queueing network analyzer," *Bell System Technical Journal*, **62** (1983), 2779–2815.

[30] WHITT, W., "Approximations for departure processes and queues in series," *Naval Research Logistics Quarterly*, **31** (1984), 499–521.

[31] WHITT, W., personal communication, 1985.

[32] WOLFF, R.W., "Poisson arrivals see time averages," *Operations Research*, **30** (1982), 223–231.

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