(Fig. 1) while $\gamma < 1/2$ [1] without such a node (Fig. 2). Consequently, the addition of a relay node decreases the maximal throughput by 33 percent in this case.

In Fig. 6, the minimal average *waiting time* of packets in the network is plotted versus the throughput for equal arrival rates at nodes 2 and 3 for the network with and without the relay node. This minimal average waiting time is obtained from the minimal average delay time minus one unit for the network of Fig. 2, and minus two units for the network of Fig. 1. As is seen from Fig. 6, the addition of a relay node significantly deteriorates the performance of the network.

References

[1] M. Sidi and A. Segall, "Two interfering queues in packet-radio networks," *IEEE Trans. Commun.*, vol. COM-31, pp. 123-127, Jan. 1983.

[2] ——, "A three-node packet-radio network," in *Proc. INFOCOM'83*, San Diego, CA, Apr. 1983, pp. 222-228.

[3] G. Fayolle and R. Iasnogorodski, "Two-coupled processors—The reduction to a Riemann-Hilbert problem," *Wahrscheinhilchkeitstheorie*, pp. 1-27, 1979.

[4] M. Eisenberg, "Two queues with alternating service," *SIAM J. Appl. Math.*, vol. 36, pp. 287-303, Apr. 1979.

[5] J. D. C. Little, "A proof for the queueing formula $L = \lambda W$," *Oper. Res.*, vol. 9, pp. 383-387, 1961.

[6] E. T. Copson, *Theory of Functions of a Complex Variable.* London, England: Oxford Univ. Press, 1948.

Adaptive Load Balancing for Parallel Queues with Traffic Constraints

TAKSHING P. YUM AND HUA-CHUN LIN

Abstract—**A new adaptive rule for balancing the load on many parallel queues is designed. The queueing system can accommodate different types of customers where each type is persistent in joining a particular set of queues. The rule makes use of a set of bias levels to compare the queue lengths and makes use of the majority-vote rule for propagating the routing decisions to the different types of customers. Delay and blocking probability comparisons between this rule and three other adaptive load balancing rules, the JSQ (join-the-shortest-queue) rule, the GBQ (generalized biased queue) rule, and the MRT (minimum response time) rule, show that it is always superior under widely different conditions on a three-parallel-queue system.**

I. Introduction

The purpose of load balancing in a queueing system is to distribute arriving customers to a number of parallel queues for services. The use of load balancing rules is common in multiprocessor computer systems. Load balancing rules with traffic constraints can also be used for realizing local adaptive multidestination routing strategies in computer communication networks.

We can generally classify load balancing rules as either stochastic, deterministic, or adaptive. Stochastic rules distribute arriving customers to different queues by fixed probability assignments. The problem of finding optimum bifurcation probabilities for minimum average delay has been solved in [1] in a more general context of a network of queues. We call the rule using the optimum probabilities the best stochastic (BS) load balancing rule. Load balancing in a network environment is similar to the routing function. In [2] and [3] the authors are actually making use of the BS rule for designing decentralized routing algorithms for computer networks. In [4], it was shown that smaller overall average delay can be achieved by distributing the customers to the set of queues according to a predetermined routing sequence. The particular deterministic rule that gives the same bifurcation ratios as the optimum bifurcation probabilities of the BS rule is referred to in [4] as the best deterministic (BD) rule.

The extensively studied join-the-shortest-queue (JSQ) rule [5]-[7] is an example of an adaptive load balancing rule. It can also perform the function of adaptive routing [8]-[10]. Another example is the join-biased queue (JBQ) rule [11]. It introduces a bias term in comparing the queue lengths.

In this paper, we propose a new adaptive load balancing rule for parallel queues with traffic constraints. These constraints are due to the presence of different types of customers (differentiated by the set of queues which they can join) in the queueing system. The rule is a combination of the JBQ rule and the majority-vote rule, and is referred to as the BQ/MV rule. It degenerates to the JSQ rule when the system has similar queues and balanced (or symmetrical) input traffic rates and degenerates to the JBQ rule when the system has only two queues.

II. The Queueing Model

Consider a system of M parallel queues serving different types of customers. We assume the arrival of each type of customer forms a Poisson process with an arbitrary rate. The amount of work brought by each customer, regardless of type, is exponentially distributed with mean equal to 1. Among the various types, the "persistent" customers are those requiring special services, and so they persist in joining a particular queue of their choice. This situation may occur in a multiprocessor system where some jobs must be processed by a special purpose processor. It may also occur in a computer network node where some messages must be routed to a particular output line. We call the customers that are constrained to join queue $i(Q_i)$ the P_i customers. The "nonpersistent" customers, on the other hand, can join two or more queues. If they can join m out of a total of M queues, we refer to them as the N_m customers. It is clear that there are a total of M types of P customers, and $\binom{M}{m}$ types of N_m customers. The total number of customer types, therefore, is $2^M - 1$. We further let an M-bit indicator, $x = x_1 x_2 \cdots x_M$, denote the $2^M - M - 1$ types of N customers with

$$x_i = \begin{cases} 1, & \text{if that type of customers can join } Q_i, \\ 0, & \text{otherwise,} \end{cases} \qquad i = 1, 2, \cdots, M.$$

Thus, for $M = 3$, $N_2(110)$ is the type of customer that can join only Q_1 and Q_2. We assume that when a customer arrives at the system, its type is first determined by a dispatcher. If it is a P_i customer, the dispatcher sends it directly to Q_i. If it is an $N_m(x)$ type customer, the dispatcher sends it to switch $S_m(x)$. The switch then sends the customer to one of the m queues indicated in x according to the load balancing

Paper approved by the Editor for Computer Communication of the IEEE Communications Society for publication after presentation at the International Conference on Communications, Amsterdam, The Netherlands, May 1984. Manuscript received January 15, 1983; revised September 12, 1983.

T. P. Yum is with the Department of Electronics, Chinese University of Hong Kong, NT, Hong Kong.

H.-C. Lin is with the Institute of Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan, R. O. C.

0090-6778/84/1200-1339$01.00 © 1984 IEEE

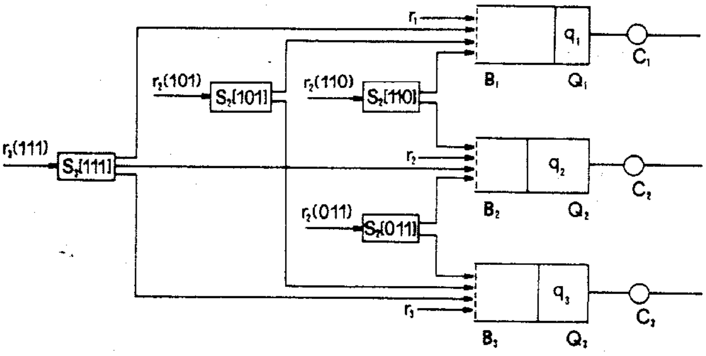

Fig. 1. Three-parallel-queue system with seven types of customers.

rule. Fig. 1 shows the seven types of customers for a three-parallel-queue system. There, switch $S_2(101)$, for example, will send the $N_2(101)$ customers to Q_1 or Q_3. We have also let the r's be the arrival rates of the different types of customers, the C_i's be the service rates, the B_i's be the buffer sizes, and the q_i's be the length of these queues.

We now refer the readers to [1] for the calculation of the optimum flow rates for the minimum delay criterion. The optimum flow rates for the minimum blocking criterion can be found similarly using the Lagrange multiplier technique. After that, the maximum entropy method [13] is used to find the set of bifurcation probabilities [4]. These flow rates and bifurcation probabilities found are to be used in the next section for the design of the BQ/MV rule.

III. The BQ/MV Load Balancing Rule

The BQ/MV rule makes use of the JBQ rule to assign the N_2 customers and the majority-vote rule to assign the N_3, N_4, $\cdots$, N_M customers. For two similar queues with no persistent arrivals, the JSQ rule is intuitively good. But with the presence of unbalanced persistent arrivals, comparing the length of $Q_1(q_1)$ and the length of Q_2 plus a suitably chosen bias value Δ_{12} was shown to give better delay performance [11]. Besides, by adjusting the bias term, we can also regulate the proportion of traffic to be sent to each queue [11].

We shall show later how the bias values are set. For the moment, let us assume that we have a three-queue system and that we know how to assign the N_2 customers. Let an $N_3(111)$ customer enter the system and be sent to $S_3(111)$. $S_3(111)$ then makes a decision as to which queue the customer should join based on the available information, such as $\{q_i\}$, $\{B_i\}$, $\{C_i\}$, and $\{r_i(x)\}$ so as to optimize certain performance criteria. This turns out to be a very difficult problem in general, for we need to partition the three-dimensional state space "optimally" into three regions for the three decisions: to Q_1, to Q_2, or to Q_3. Thus, instead of searching for the optimum decision rule, we choose to make the decision based on the recommendations provided by the three S_2 switches and send the $N_3(111)$ customer to the queue with the highest recommendation (or vote). In case of a tie, select one randomly. For an M-parallel-queue system, however, the recommendations for a particular S_3 switch must be obtained from the "relevant" S_2 swtiches. A switch is relevant if it makes recommendations from a subset of queues which the particular N_3 customer can join. Thus, for $M = 4$, the relevant S_2 switches for $S_3(0111)$ are $S_2(0011)$, $S_2(0101)$, and $S_2(0110)$. All other S_2 switches are "irrelevant." In a similar fashion, the S_4 switches can make the queue assignment decisions by the majority-vote rule based on the votes of the relevant S_3 switches. Thus, we see that the instantaneous queue length information for load balancing is propagated from S_2 to S_3 to S_4, etc., and this is how the N_3, N_4, $\cdots$ customers are routed to the various queues.

We now turn to the setting of the bias values (Δ_{ij}'s) for assigning the N_2 customers. Consider again a three-queue system and let us focus on the $N_2(110)$ customers. To solve for Δ_{12}, we let the arrivals to Q_1 (except the $N_2(110)$ arrivals) be represented by a composite arrival process and the arrival to Q_2 be represented by another composite process. We then use different bias values on this two-queue system, solve for their corresponding average delays and average blocking probabilities, and choose the Δ_{12} that minimizes the average delay or average blocking, depending on the performance criterion used. Other values of Δ_{ij} can be found similarly. Details can be found in [11].

IV. Performance Comparisons

In this section, we attempt to compare five load balancing rules: the BS rule, the JSQ rule, the GBQ rule, the MRT rule, and the BQ/MV rule. For the BS rule, we use the minimum blocking/delay criterion to find the $\{\lambda_i\}$ and the maximum entropy bifurcation probabilities to split the N type traffic. This is therefore the optimum stochastic rule. The GBQ (generalized biased queue) rule is a generalization of the JBQ rule for three or more queue systems. It uses random bifurcation to split the N_3, N_4, $\cdots$ types of traffic. But instead of routing them to the queues directly, it groups these streams of bifurcated traffic that belong to the same $N_i(x)$ type two by two in descending order of probability values and sends them to the corresponding S_2 switches for adaptive bifurcation using the JBQ rule. All streams of bifurcated traffic that are left over from the two-by-two grouping are to be sent directly to the corresponding queues. A more detailed discussion of the GBQ rule can be found in [11].

The MRT (minimum response time) rule is first proposed in [12]. It is specified on a queueing system where there is only one arrival stream and, therefore, cannot be used directly to balance the load of the queueing system in Fig. 1. We generalized it as follows: among the queues a particular customer is allowed to join, select the one that will give the minimum expected response time (waiting time plus service time) *for that customer*. Note that this is an individual optimization policy. It does not imply in any sense system optimality.

We now focus on a three-queue system. For the first set of comparisons, we choose the minimum blocking criterion. Fig. 2 shows the average blocking probabilities of the five load balancing rules. As k increases, that is, as the amount of N_3 traffic increases relative to the N_2 traffic, the blocking probability decreases for the JSQ (MRT) and BQ/MV rules. This is to be expected, as when $r_3(111)$ increases, more traffic has more freedom in choosing queues for load balancing. The GBQ rule shows, however, a slight increase in blocking probability. This is because as $r_3(111)$ increases, more traffic has to be randomly bifurcated before the final adaptive bifurcation.

Fig. 3 shows the case where the GBQ rule is better than the JSQ (MRT) rule. Here, as we increase k, the arrival rates become more unbalanced (or asymmetric). This is the case where we need large values of Δ_{ij}'s in order to invert the unbalanced condition. The JSQ rule always has $\Delta_{ij} = 0$ and does not make such a provision. The BQ/MV rule still has the best performance, however. For the above two cases, the difference in average delays between the four adaptive rules is about 1 percent.

We now focus on the comparisons when the buffer sizes are different and the input rates are highly uneven. Fig. 4 shows that the blocking performance between the BQ/MV rule and the other three rules accentuates. Hence, the BQ/MV rule is more robust.

Fig. 5 shows the blocking probabilities of the five rules for different service rates. In all cases, the BQ/MV rule is superior. Note that, for $d \neq 0$, the MRT rule is no longer identical to the JSQ rule. The MRT rule, in an attempt to minimize

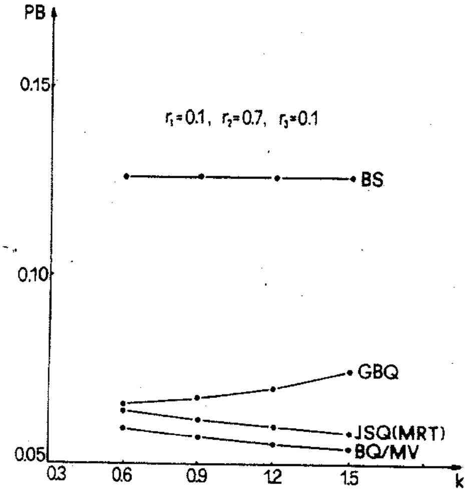

Fig. 2. Blocking comparisons: $r_3(111) = k$, all $r_2(\cdot) = 0.6 - 1/3\,k$, $B_1 = B_2 = B_3 = 5$, $C_1 = C_2 = C_3 = 1.0$.

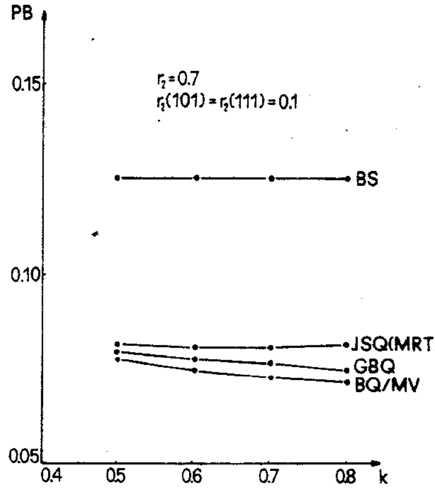

Fig. 3. Blocking comparisons: $r_2(110) = r_2(011) = k$, $r_1 = r_3 = 0.9 - k$, $B_1 = B_2 = B_2 = 5$, $C_1 = C_2 = C_3 = 1.0$.

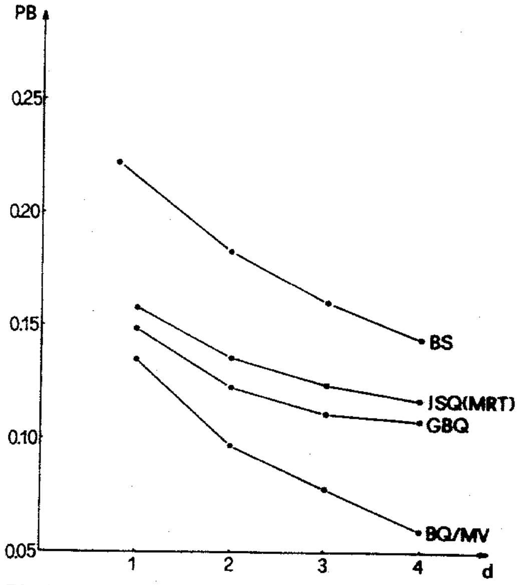

Fig. 4. Blocking comparisons: $B_1 = 1$, $B_2 = 2 + d$, $B_3 = 2 + 3\,d$, $r = 2.7$, $r_2 = 0.7$, $r_3 = 0.1$, $r_2(110) = r_2(101) = r_2(011) = 0.5 - 0.1d$, $r_3(111) = 0.3 + 0.3\,d$, $C_1 = C_2 = C_3 = 1.0$.

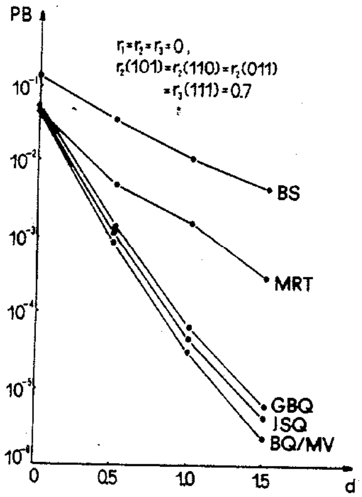

Fig. 5. Blocking comparisons: unequal service rates, $C_1 = 1.0$, $C_2 = 1.0 + d$, $C_3 = 1.0 + 2\,d$, $B_1 = B_2 = B_3 = 5$.

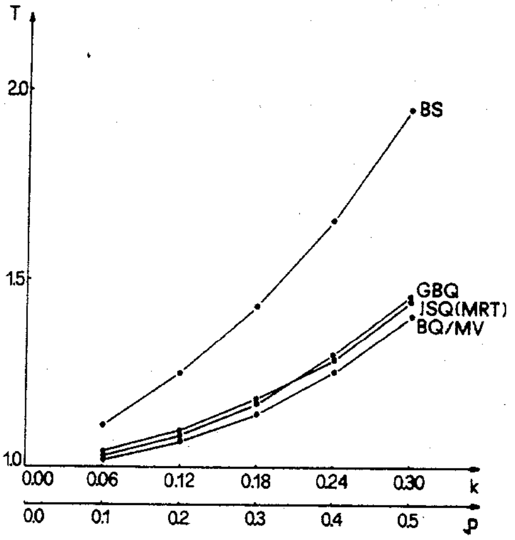

Fig. 6. Delay comparisons: $r_1 = r_2(110) = r_2(101) = r_2(011) = r_3(111) = k$, $r_2 = r_3 = 0$, $B_1 = B_2 = B_3 = 7$, $C_1 = C_2 = C_3 = 1.0$

the response time, would like to route the customers to the queue with a faster service rate, even though the waiting room for that queue may be full. Thus, the blocking performance is severely affected.

We now consider the performance of the rules under the minimum delay criterion. Fig. 6 shows the delay with the average system utilization factor ρ varying from 0.1 to 0.5. We see that the BQ/MV rule gives the smallest delay as well as the smallest blocking probability. For $\rho = 0.5$, the blocking probabilities for the BS, GBQ, JSQ (MRT), and BQ/MV rules are, respectively, 3.9×10^{-3}, 1.8×10^{-4}, 1.2×10^{-4}, and 7.9×10^{-5}. We also observed that as ρ gets larger, the delay difference between the rules also gets larger. Fig. 7 shows the delay performance with unequal service rate under the minimum delay criterion. The blocking probabilities of the four adaptive rules are all about 10^{-4}. Again we see that the BQ/MV rule is superior.

V. Conclusions

Load balancing is an important function in any resource sharing system. As the system gets larger and more complex, there are bound to be some constraints in allocating resources.

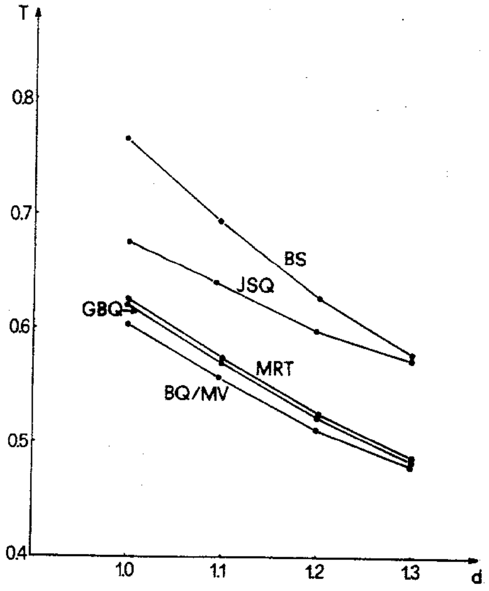

Fig. 7. Delay comparisons: $r_1 = 0.0, r_2 = 0.3, r_3 = 0.7$, all $r_2(\cdot) = 0.3, r_3(111) = 0.7, B_1 = 5, B_2 = 6, B_3 = 7, C_1 = 1.0, C_2 = 1.0 + d, C_3 = 1.0 + 2\,d$.

These constraints make the design of load balancing rules more difficult, especially when the system has to operate under widely different conditions.

This paper attempts to evaluate the comparative performance of four adaptive load balancing rules on a three-queue system with traffic constraints. Four-or-more-queue systems are not investigated due to the difficulties of solving four-or-more-dimensional Markov chains. The rules designed and studied, however, are perfectly general for any number of queues. All the rules can also be easily implemented. The most complicated one, the BQ/MV rule, requires only a few "compare" operations for determining the majority vote. The set of Δ_{ij}'s can either be calculated when needed or be tabulated for later references.

From the performance evaluation of the various rules on a three-queue system with traffic constraints and under widely different conditions, we found that the BQ/MV rule is the best under both the minimum delay and the minimum blocking criteria. In particular, the BQ/MV rule is significantly better than the JSQ rule under any kind of asymmetric condition, e.g., highly uneven input rates, unequal buffer sizes, and unequal service rates. It is also significantly better than the GBQ rule when the N_3 type traffic is increased relative to the N_2 type traffic.

References

[1] L. Fratta, M. Gerla, and L. Kleinrock, "The flow deviation method: An approach to store-and-forward communication network design," *Networks*, vol. 3, pp. 97–133, 1973.

[2] R. G. Gallager, "A minimum delay routing algorithm using distributed computation," *IEEE Trans. Commun.*, vol. COM-25, Jan. 1977.

[3] T. Stern, "A class of decentralized routing algorithm using relaxation," *IEEE Trans. Commun.*, vol. COM-25, Oct. 1977.

[4] T. Yum, "The design and analysis of a semidynamic deterministic routing rule," *IEEE Trans. Commun.*, vol. COM-29, Apr. 1981.

[5] J. F. C. Kingman, "Two similar queues in parallel," *Ann. Math. Statist.*, vol. 32, pp. 1314–1323, 1961.

[6] L. Flatto and H. McKean, "Two parallel queues with equal servicing rates," IBM Res. Rep. RC5816, Math., Mar. 24, 1976.

[7] G. Foschini and J. Salz, "A basic dynamic routing problem and diffusion approximation," *IEEE Trans. Commun.*, vol. COM-26, Mar. 1978.

[8] H. Rudin, "On routing and 'delta routing': A taxonomy and performance comparison of techniques for packet-switched networks," *IEEE Trans. Commun.*, vol. COM-24, Jan. 1976.

[9] G. L. Fultz, "Adaptive routing techniques for message switching computer communication networks," Univ. California, Los Angeles, Eng. Rep. UCLA-ENG-7252, July 1972.

[10] A. Livne, "Dynamic routing in computer communication networks," Ph.D. dissertation, Polytech. Inst. New York, Brooklyn, NY, July 1976.

[11] T. Yum and M. Schwartz, "The join-biased-queue (JBQ) rule and its application to routing in computer communication networks," *IEEE Trans. Commun.*, vol. COM-29, Apr. 1981.

[12] Y. Chow and W. H. Kohler, "Models for dynamic load balancing in a heterogeneous multiple processors system," *IEEE Trans. Comput.*, vol. C-28, May 1979.

[13] R. D. Levine and M. Tribus, Eds., *The Maximum Entropy Formalism*. Cambridge, MA: M.I.T. Press, 1979.

Series Representations for Rice's *Ie* Function

B. T. TAN, T. T. TJHUNG, C. H. TEO, AND P. Y. LEONG

Abstract—**Two new series representations for the Rice function *Ie* (*k*, *x*) are presented. One of the series involves the modified Struve functions and the other involves the modified Bessel functions. These two series complement each other in their convergence speeds as functions of the values of *k* and *x*. The truncation error bounds are derived for both series. Therefore, they can be used alternatively with high efficiency and known precision.**

Rice's *Ie* function [1, p. 267],

$$Ie(k, x) = \int_0^x e^{-r} I_0(kr)\, dr, \qquad 0 \leqslant k \leqslant 1, \tag{1}$$

appears frequently in the analysis of angle modulation systems [1], and recently, it was shown to play a central role in the study of error rates in differentially encoded systems [2]. A table of $Ie(k, x)$ computed to seven significant figures, with values of k ranging from 0.1 to 1.0 and x from 0.1 to 39.0, has been presented in [1, pp. 268–270]. However, due to the present widespread work in angle modulation and differentially encoded systems, efficient methods for computing $Ie(k, x)$ not included in the published tables are desirable. In this note, we introduce two computationally efficient series representations for $Ie(k, x)$. It can be shown from [2, eq. 16] and [3, eq. 13] that

$$Ie\left(\frac{V}{U}, U\right) = \frac{U}{W} - \frac{U}{\pi}\int_0^{\pi} d\theta\, \frac{e^{-(U - V\cos\theta)}}{U - V\cos\theta}$$

$$= U\sqrt{\frac{\pi}{2W}}\, e^{-U} \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{V}{2}\right)^{2n}\left(\frac{2}{W}\right)^{n} \cdot \left[\frac{U}{W} L_{n+(1/2)}(W) + L_{n-(1/2)}(W)\right] \tag{2}$$

Paper approved by the Editor for Data Communication Systems of the IEEE Communications Society for publication without oral presentation. Manuscript received October 17, 1983; revised April 4, 1984.

The authors are with the Department of Electrical Engineering, National University of Singapore, Kent Ridge, Singapore 0511.

0090-6778/84/1200-1342$01.00 © 1984 IEEE