Assigning Customers to Two Parallel Servers with Resequencing

Nitin R. Gogate, *Member, IEEE,* and Shivendra S. Panwar, *Senior Member, IEEE*

***Abstract*— In this letter, we consider the resequencing delay characteristics of a system with two heterogeneous servers serving a single queue with Poisson packet arrivals. An ordered delivery of packets is guaranteed at the receiver, hence packets arriving out of order suffer additional resequencing delay. We introduce the concept of threshold policies with variable routing positions and obtain analytical expressions for the end-to-end delay. With numerical examples we show that some improvement is obtained in the end-to-end delay values with variable routing position policies over fixed position routing policies. Further, initial simulation studies seem to indicate that this observation is true even with bursty arrivals.**

***Index Terms*—Packet-switched networks, queueing models, resequencing, routing policies.**

I. Introduction

A SYSTEM with multiple heterogeneous servers serving a single queue has been considered extensively in the literature [1]–[10]. The variations include different arrival and service distributions, number of available servers, or strategies for activation of the servers. The optimality of a *threshold type* of scheduling policy for minimizing the mean sojourn time of the customers (not including the resequencing delay) in a two-server system was proved in [1]. The results were extended in [2] to a more general arrival and service distributions. In [3] the authors obtained closed-form expressions for the resequencing and the queueing delay for the two heterogeneous server system (with Poisson arrivals and exponential service rates) under a *threshold type* of policy, where the slow server was activated by serving a customer from a *fixed position* in the queue if the number of customers waiting in the system exceeded a certain threshold. It was proved in [4] that the threshold type of policies are also optimal under the resequencing constraint, but only for some routing positions. It was conjectured in [6] that an optimal policy within the class of all nonpreemptive policies is of a threshold type with *variable routing position*. The optimality of a variable routing position policy is extremely difficult to prove [4] and is an open problem. In this work, we have obtained analytical expressions for the total end-to-end delay (including the resequencing delay) to gain some insight to this problem.

We can model an end-to-end connection between a pair of nodes communicating on multiple channels, either by a set of parallel queues and servers (e.g., a bottleneck within the backbone WAN) or by a queue served by multiple parallel servers (e.g., a bottleneck at the network interface point). In the latter scenario, queuing occurs mainly at the edge of the network rather than within the backbone WAN. Another way of viewing the latter model is having a rate controlled or credit based access mechanism at the interface point, one for each path. In addition, a single queue served by multiple parallel servers has also been used to model a pair of communicating nodes connected by multiple links.

The paper is organized as follows. In Section II we present the motivation for our model. Section III deals with the evaluation of the total expected end-to-end delay. The numerical results from our analysis are reported in Section IV, followed by our conclusions in Section V.

II. Motivation and Model Description

Here we consider a system of two heterogeneous servers with service rates μ_1 and μ_2, and Poisson arrivals with rate λ. The service times are assumed to be exponentially distributed. We assume a stable system $(\lambda < \mu_1 + \mu_2)$ with an infinite buffer and $\mu_1 \geq \mu_2$. Also, without loss of generality, we further assume that $\lambda + \mu_1 + \mu_2 = 1$. A *fixed position* routing policy is that which always routes a customer to the slow server from a fixed position in the queue. A *fixed position threshold type* of policy is that which always routes a customer from a fixed position in the queue to the slow server if the number of customers in the queue exceeds a certain threshold. Two fixed position threshold type of policies were analyzed and compared in [3]. We consider a routing policy with variable routing position, with the objective of minimizing the sum of the queueing delay and resequencing delay. In general, in a *variable routing position* policy we can route a customer to the slow server from any position in the queue. The system schematic is shown in Fig. 1.

The routing policy we shall consider is as follows. The fast server $[S_1]$ is always activated from the first position in the queue and the slow server $[S_2]$ is activated only if the total number of customers in the queue, at the time of the routing decision, exceeds a certain threshold value (N). The routing policy routes the customer to S_2 from a position nearest to the

Manuscript received February 22, 1998. The associate editor coordinating the review of this letter and approving it for publication was Prof. Y. Bar-Ness. This work was supported by the New York State Science and Technology Foundation's Center for Advanced Technology in Telecommunications, Polytechnic University, Brooklyn, NY, and by the National Science Foundation under Grant NCR-9115864. This letter was presented at the Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore, MD, March 1995.

N. R. Gogate was with the Department of Electrical Engineering, Polytechnic University, Brooklyn, NY 11201. He is now with Fujitsu Networ Communications, Pearl River, NY 10965 USA (e-mail: gogate@tddny.fujitsu.com).

S. S. Panwar is with the Department of Electrical Engineering, Polytechnic University, Brooklyn, NY 11201 USA.

Publisher Item Identifier S 1089-7798(99)02679-4.

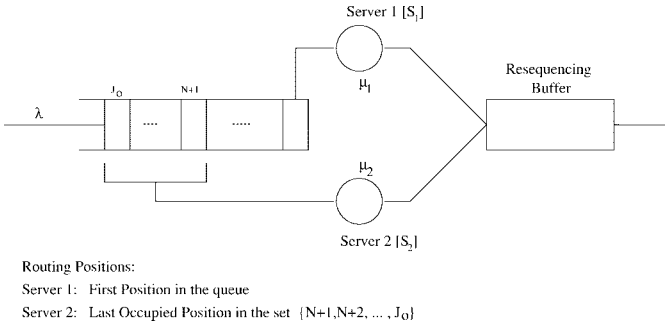

Fig. 1. A heterogeneous $M/M/2$ system with variable routing positions.

"optimal" position (J_0), which is generally greater than N. The optimality of the position J_0 for fixed position policies was shown in [3] and [4]. J_0 is given as follows:

$$J_0 = \left\lfloor \frac{\ln(1-\alpha)}{\ln \alpha} \right\rfloor, \qquad \text{where } \alpha = \frac{\mu_1}{\mu_1 + \mu_2}.$$

If we consider a position (x) which is the last occupied position in the queue and an inactive slow server, then at the time of a routing decision:

$x \le N$	do not activate the slow server;
$N < x \le J_0$	activate the slow server with a customer from position x;
$x > J_0$	activate the slow server with a customer from position J_0.

We will compare our results with the best fixed position threshold policy given in [3]. In this scenario, the fixed position threshold policy routes a customer from the position $N+1$ to the slow server. We considered the above variable routing position policy because of the following argument. It was shown in [3] that the queueing delay is dependent only on the choice of the threshold and is independent of the choice of the routing position. It was also shown in [4] there exists an optimal position J_0, such that it always pays to route from that position, if it is occupied. If both the policies have the same threshold then the queueing portion of the total end-to-end delay (TD) is the same for both of them. In the variable position routing policy, since we are routing from a position as close to J_0 as possible the resequencing delay should be less than a fixed position policy that routes from position $N+1$. So a variable position routing policy of the above nature will perform at least as well as a fixed position routing policy, if they have the same threshold $(<J_0)$. Hence we expect that overall end-to-end delay will be less for this variable routing position policy. Based on the above argument we conjecture that this policy is optimal, but have been unable to prove it.

III. Analysis

For the sake of brevity, the detailed analysis is not presented here. We use a technique similar to [3] wherein fixed position routing policies were considered and a closed-form expression was obtained for the resequencing delay. Let us take the system state to be (x_0, x_1, x_2), where x_0 is the number of customers waiting in queue, x_1 the state of the fast server, and x_2 the state of the slow server. Server state (0) indicates an idle server and state (1) indicates a busy server. We evaluate the total expected system time (including the resequencing delay) for a tagged arriving customer for the policy described in the previous section. Depending on the system state seen by an arriving customer, we considered five different cases. For the two cases corresponding to the tagged customer (TC) being routed to the fast and the slow server, the problem is formulated as a 2-D random walk with two absorbing barriers—barrier-1 and barrier-2—which correspond to crossing the N-boundary and service by S_2, respectively, as shown in Fig. 2. By evaluating the expected time spent by a tagged customer in these two random walks (i.e., the expected durations of the random walks) and combining that with the other cases we find the expression for the expected system time of the tagged customer. A detailed analysis is given in [13].

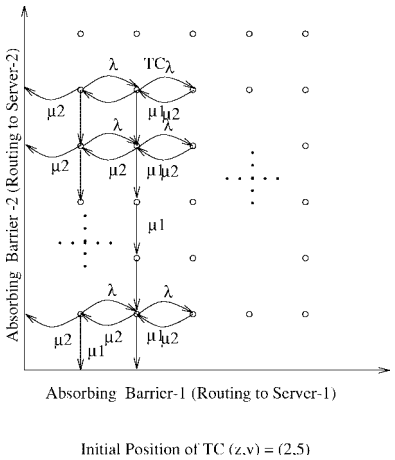

Fig. 2. Two-dimensional random walk model.

IV. Numerical Results and Discussion

In this section, with the help of some numerical examples we compare the performance of the fixed position and the variable position threshold policies.

We onserved that for the threshold type of policies, at the optimum threshold the average resequencing delay is only a small fraction of the total end-to-end delay. In all of the cases that have been considered, the optimal threshold for the variable position policy was always less than or equal to the optimal threshold (N_0) for the fixed position policy, with a difference of at most 2. Hence, the queueing portion of the end-to-end delay is somewhat less for the variable position policy as compared to the fixed position policy, when optimum thresholds were different. However, for these cases the resequencing delay, for the variable position policy is slightly higher than that for the fixed position policy. So the net reduction in the end-to-end delay was marginal. When the optimum thresholds were identical, the resequencing delay using the variable position policy was always lower than for the fixed position policy. In all of the cases considered the improvement ranged from negligible to 3.9%. Some typical end-to-end delay results for different values of arrival and service rates are presented in Table I.

TABLE I
END-TO-END DELAY VALUES FOR VARIABLE AND FIXED POSITION POLICIES

λ	μ_1	μ_2	J_0	Fixed Pos		Variable Pos.	
				N_o	TD	N_o	TD
49	56	14	7	3	0.07479	3	0.07429
56	56	14	7	3	0.10480	2	0.10294
63	56	14	7	2	0.18200	2	0.17766
46.9	56	11	10	4	0.07814	4	0.07751
53.6	56	11	10	4	0.11206	3	0.11028
60.3	56	11	10	3	0.19630	2	0.19008
43.12	56	5.6	25	12	0.07596	12	0.07588
49.28	56	5.6	25	10	0.12269	10	0.12153
55.44	56	5.6	25	9	0.23438	7	0.22530

Here we have presented a general analysis method for evaluating the total expected end-to-end delay for the system shown in Fig. 1. As the average resequencing delay is only a small fraction of the total end-to-end delay, the improvements in the end-to-end delay values (over fixed position threshold policies) with variable routing position threshold policies is only marginal, for the arrival and service distributions considered in this study. How well this policy performs with different arrival and service distributions is an open problem. Initial simulation results in [13] seem to indicate that for two-stage hyperexponential arrivals and exponential service, the above observations hold good.

In this study we see that the variable routing position policy results in only a modest reduction in end-to-end delay; at least for Poisson arrivals and exponential service, it might still be worthwhile to minimize resequencing delay to reduce either the resequencing buffer size or in the case of bursty service. It was observed in [9] that both the queueing delay as well as the resequencing delay increase almost linearly with the squared coefficient of variation of the service time distribution.

V. CONCLUSIONS

In this letter we have presented a problem of routing customers to a set of two heterogeneous servers under a variable position threshold type of policy. We have developed an analytical technique to evaluate the total end-to-end delay (including the resequencing delay). With numerical examples we also show that the variable position threshold policy we have considered, and which has been conjectured to be optimal, performs better than the fixed position threshold type of policies, though the improvement obtained in the end-to-end delay values is only marginal. Some initial simulation studies seem to indicate that this is also true in the case of bursty arrivals. Determining the optimal resequencing policy for this model is an interesting open problem.

REFERENCES

[1] W. Lin and P. R. Kumar, "Optimal control of a queuing system with two heterogeneous servers," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 696–703, Aug. 1984.
[2] I. Viniotis and A. Ephremides, "Extension of the optimality of the threshold policy in heterogeneous multiserver queuing systems," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 104–109, Jan. 1988.
[3] I. Iliadis and L. Y.-C. Lien, "Resequencing delay for a queuing system with two heterogeneous servers under a threshold type scheduling," *IEEE Trans. Commun.*, vol. 36, pp. 692–702, June 1988.
[4] S. Ayoun and Z. Rosberg, "Optimal routing to two parallel heterogeneous servers with resequencing," *IEEE Trans. Automat, Contr.*, vol. 36, pp. 1436–1449, Dec. 1991.
[5] I. Iliadis and L. Y.-C. Lien, "Resequencing control for a queuing system with two heterogeneous servers," *IEEE Trans. Commun.*, vol. 41, pp. 951–961, June 1993.
[6] S. X. Zu, "On a job resequencing issue in parallel processor stochastic scheduling," *Adv. Appl. Probability*, vol. 24, pp. 915–933, 1992.
[7] S. Varma, "Optimal allocation of customers in a two server queue with resequencing," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1288–1293, Nov. 1991.
[8] T. S. P. Yum and T. Y. Ngai, "Resequencing of messages in communication network," *IEEE Trans. Commun.*, vol. COM-34, pp. 143–149, Feb. 1986.
[9] S. Chowdhury, "An analysis of virtual circuit with parallel links," *IEEE Trans. Commun.*, vol. 39, pp. 1184–1188, Aug. 1991.
[10] ———, "Distribution of the total delay of packets in virtual circuits," in *Proc IEEE INFOCOM'91*, Apr. 1991, pp. 911–918.
[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York: Wiley,1986, vol. 1.
[12] N. Gogate and S. S. Panwar, "On a resequencing model for high speed networks," in *Proc. IEEE INFOCOM'94*, June 1994, pp. 40–47.
[13] ———, "Assigning customers to two parallel servers with resequencing," Dep. Electrical Eng., Polytechnic Univ., Brooklyn, NY, Tech. Rep. TR 94-78.