Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

APPLIED SCIENCE DIVISION

Submitted to Operations Research

ON THE BENEFITS OF COMBINING QUEUES

M.H. Rothkopf and P. Rech

February 1987

APPLIED SCIENCE DIVISION

Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

DISCLAIMER

ON THE BENEFITS OF COMBINING QUEUES

Michael H. Rothkopf amd Paul Rech

Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

February 1987

This work was supported by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

LBL-22917

ON THE BENEFITS OF COMBINING QUEUES

by

Michael H. Rothkopf and Paul Rech

A recent seminar in the Operations Research Department of a major university given by a panel of faculty had as a theme the value of operations research. During the seminar, the advantage of combining separate queues was used as a key and uncontested example of the benefits that accrue from operations research. While there was a claim that many banks and other counter systems have gone to combined queues, there was also an observation that many managers were slow to adopt such an obvious improvement and that more selling effort was required.

A member of the audience, a graduate of the Department now employed in industry, raised the only caution. He suggested that managers might be reluctant to combine queues because of concern about customer acceptance; the combined queue looks long. This was taken as shifting the need for "education" from the managers to the customers. No one present questioned whether combining queues is, in fact, a good idea. But, is it? Are there factors, other than customer ignorance, that should lead one to question the benefits of combining queues--especially queues of people? It is the purpose of this writing to suggest that there are such factors and that they deserve careful investigation. First, however, for completeness we set forth the case for combining queues.

For Markovian queues, it is a simple matter to compare the steady state average wait in one of s standard (i.e., unlimited waiting room, no balking or reneging, etc.) M/M/1 systems having service rate μ and arrival rate λ with the steady state average wait in a standard M/M/s system with the same service rate μ and an arrival rate $s\lambda$. This comparison will always show that the second

expression is less than the first for any positive λ and any integer $s > 1$. For a proof, see Smith and Whitt. Indeed, the improvement can be dramatic. Wolff's forthcoming book shows that the average wait in the combined queue is less by at least a factor of s. A similar comparison for the variance of the wait in the system leads to a similar result. Furthermore, numerical results for similar comparisons with various non-Markovian queues will yield similar results. Wolff as well as Smith and Whitt have some relevant proofs. Indeed, comparisons such as these are sometimes discussed in introductory texts. For example, see Wagner, pp. 882-883 or Baker and Kropp, pp. 474. Hillier and Lieberman have an extensive example involving combining geographically dispersed servers. Other texts assign simple comparisons to students. For example, see Truman, problem 12.12; Anderson and Lievano, problem 8-4; or Davis *et al.*, problem 20-13. Bierman, Bonini and Hausman, problems 20-5 and 20-13 assign such comparisons with appropriate calls for examination of assumptions. Do such comparisons or proofs settle the matter? No, and for several reasons.

First, *all* of the reduction in the workload (measured in server minutes) waiting for service occurs because multiple queue systems sometimes have an idle server while a customer is waiting for service in another queue. If there is jockeying between queues, at least when a server is idle ("No waiting on aisle 7."), such a situation will not arise. If the jockeying is nondiscriminatory with respect to waiting time, then the number of customers in the queue and, hence, the *average* wait in the queue will be the same in both types of systems. Moreover, if the jockeying favors shorter jobs, then the average wait will be *less* in the separate queue system. Shorter jobs might be favored because customers with them can jockey more effectively (e.g., those with a basket vs. those with a full shopping cart), or because cultural factors such as politeness favor letting customers with short jobs have preference going to an open server. Even if the jockeying is-

nondiscriminatory with respect to *service time*, it may reduce average *delay costs* if those in a hurry due to high delay costs are more prone to jockey.

Even with jockeying that is nondiscriminatory as to service time, the average wait in the single queue system will be *longer* if for any reason combining the queues increases the service times even slightly. There are several possible reasons why such increases might happen. It may take time to get from the central queue to the next available server. It may be that the physical proximity of the customer to his or her ultimate server can allow some overlapping of service such as unloading a shopping cart or getting a form to fill out while the current customer signs one. It may also be that a server is more accountable and feels more responsible for his or her own queue and may therefore be inclined to work faster, especially when the queue is long. With individual queues, customers "belong" to individual servers. A server cannot easily rely on other servers to take care of his or her customers. Finally, it may also be faster because of a subtle degree of specialization. If I usually use the same teller, that teller may become more efficient in dealing with my typical transactions.

While jockeying will not eliminate all of the difference in the variance of time in the system, it will reduce some of the variance in the multi-queue case. On the other hand, any increase in service time in the single queue case will increase the variance as well as the mean time in the system. When server utilization is high, this can be a dominating effect.

Even if there is no jockeying, if customers tend to join shorter queues, the separate systems are no longer independent, and much of the extra delay of independent systems would be eliminated. In particular, if arriving customers can observe accurately the workload in each separate queue and then join the queue with the shortest workload, the distribution of delays will be the same as it would be with a combined queue. If arriving customers join the queue with the fewest

customers, the result is almost as good. The less variability there is in service times, and the longer the queues, the more accurately the number of customers in a queue predicts that queue's workload. This suggests that practices that tend to reduce the variance in service times--especially the unobservable part of the variance-- are not only effective tools in the management of queuing systems, but also reduce the incentive to combine queues.

Even if service times do not change and there is no jockeying, if the separate queues tend to segregate jobs by service time (as with an express line) combining the queues will increase the variance of the service times in the combined queue and may increase average delay. Smith and Whitt give examples to illustrate this effect.

Going to a single queue system may involve a variety of disadvantages. First, it makes it more difficult for customers to choose servers and, thus, may contribute to depersonalizing the server-customer relationship and decreasing the customer's sense of satisfaction and perhaps the server's as well. Second, it makes it difficult for the system to use opportunities for specialized servers ("Money orders at windows 5, 6, or 7 only."), forcing it to incur the additional cost of training and equipping each server for each allowed type of service. An important kind of specialization is geographic dispersion. Third, it may deprive management of an opportunity to use delay avoidance to motivate customer behavior it desires ("Cash only in the express line." "Diamond lanes for car pools only."). Finally, it may focus scarce managerial attention on a relatively unimportant question. The critical question regarding the queue's management may well be the rules for adding or subtracting servers (e.g., by sending them to stock shelves or sort letters), the rules for assigning priorities, the selection of equipment for increasing the speed of servers, or improving the quality of the experience and the satisfaction of customers while waiting.

Finally, we have another reason for questioning the efficiency in practice of combining queues. We have seen a number of papers on the successful application of queuing theory. Papers by Edie, Sze, Gilliam, McKeown, Koopman, Vogel, Bouland, and Bluel, and the work of the New York City RAND Institute come to mind. But we do not recall seeing a convincing practice paper describing a successful unification of a multi-queue system into a single queue system. Bleuel's practice prize competition winning paper comes closest. It analyzes service team size for a geographically dispersed service operation and recommends medium size teams. A paper by Jones *et al.*, discusses a survey of shopper attitudes toward combining grocery store queues and ends up with a slightly negative attitude toward such combining based in part, upon the presence of jockeying.

In a recent draft paper, "Social Justice and Other Attributes of Queueing," Larson argues that the psychological experiences of people in queues may be more important than the actual amount of delay. In particular, he argues that the "social injustice" of the violation of first-come-first-served order may contribute to dissatisfaction. Of course, combined queues avoid violations of first-come-first-served order. We heartily agree with Larson's views on the importance of the psychology of queueing. We also agree that perceptions of fairness can play a role. However, queue managers may be clever enough to deal with this aspect of the problem without seriously compromising other aspects of the operational effectiveness of the system. Furthermore, the power to choose a server and to jockey may be an important psychological benefit that helps to offset the sense of powerlessness that waiting can cause. It may also reduce feelings of frustration and injustice associated with violations of first-come-first-served order.

In summary, there are significant reasons for believing combining queues may at times not be a good thing to do. These reasons include customer reaction, jockeying in separate queues, increased service times and costs for combined

queues, and the absence of published before-and-after studies. This is not to say that combining queues is always a bad idea. We are confident that in some engineering contexts in which arrivals cannot jockey, it is a manifestly good idea. It may also be advantageous in some counter service systems. However, it is not *automatically* a winner.

We hope that when operations researchers think about and analyze service systems, they will pay attention to the concerns we have raised. In addition, we hope that some of these concerns will inspire research on some new queueing questions and reports of experience in combining queues in counter service systems.

Acknowledgement

We are indebted for perceptive comments by Fred Hillier, Ernest Koenigsberg, Ronald Wolff, David Wood and especially Ward Whitt.

References

Anderson, Michael Q. and Lievano, R.J., *Quantitative Management: An Introduction*, Second Edition, Kent Publishing Co., 1976.

Baker, Kenneth R. and Kropp, Dean H., *Management Science: An Introduction to the Use of Decision Models*, John Wiley and Sons, 1985.

Bierman, Harold Jr., Bonini, Charles P. and Hausman, Warren H., *Quantitative Analysis for Business Decisions*, Seventh Edition, Irwin, 1986.

Bleuel, William H., "Management Science's Impact on Service Strategy," *Interfaces* 6, No. 1, Part 2, pp. 4-12, 1975.

Bouland, Heber D., "Truck Queues at County Grain Elevator," *Oper. Res.* 15, pp. 649-659, 1967.

Davis, K. Roscoe, McKeown, Patrick G. and Raher, Terry R., *Management Science: An Introduction*, Kent Publishing Co., 1986.

Edie, Leslie C., "Traffic Delays at Toll Booths," *Oper. Res.* 2, pp. 107-138, 1954.

Jones, Michael T., Oberski, Arlene M., and Tour, Gail, "Quickening the Queue in Grocery Stores," *Interfaces* 10, No. 3, pp. 90-92, 1980.

Larson, Richard C., "Social Justice and Other Attributes of Queueing," Draft paper, January 1987.

Koopman, Bernard, O., "Air-Terminal Delays Under Time-Dependent Conditions," *Oper. Res.* 20, pp. 1089-1114, 1972.

McKeown, Patrick G., "An Application of Queuing Analysis to the New York State Child Abuse and Maltreatment Register Telephone," *Interfaces* 9, No. 3, pp. 20-25, 1979.

Smith, D. R., and Whitt, W., "Resource Sharing Efficiency in Traffic Systems," *Bell System Technical Journal* 60, pp 39-55, 1981.

Sze, David Y., "A Queuing Model for Telephone Operator Staffing," *Oper. Res* 32, pp. 229-249, 1984.

Trueman, Richard C., "An Introduction to Quantitative Methods for Decision Making," Second Edition, Holt, Rinehart and Winston, 1977.

Vogel, Myles A., "Queuing Theory Applied to Machine Manning," *Interfaces* 9, No. 4, pp. 1-8, 1979.

Wagner, Harvey M., *Principles of Operations Research,* 2nd Edition, Prentice Hall, 1975.

Wolff, Ronald W., "An Upper Bound for Multi-Channel Queues," *J. Appl. Prob.* 14, No. 4, pp. 884-8, 1977.

Wolff, Ronald W., "Upper Bounds on Work in Systems for Multi-channel Queues," To appear in *J. Appl. Prob.*

Wolff, Ronald W., *Stochastic Modeling and the Theory of Queues,* Chapter 5, Prentice Hall, forthcoming.

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720