Taking Advantage of Partial Customer Flexibility
An Inexpensive Means of Improving Performance

Rhonda Righter
Department of Industrial Engineering, University of California, Berkeley, CA, U.S.A.

Keywords: Customer Flexibility, Routing, Scheduling, Service Systems, Call Centers.

Abstract: In many service systems with multiple types of customers, providing server flexibility, e.g., by cross-training servers, is very expensive. On the other hand, there is often inherent flexibility in some of the customers that is not used by the system. I argue that taking advantage of such flexibility can create a win-win-win situation, in which overall performance can be greatly improved, and in which both flexible and non-flexible customers benefit. Moreover, only a small subset of customers needs to be flexible to obtain nearly the benefit of full flexibility.

1 INTRODUCTION

In many service, production, and traffic systems there are multiple types, or classes, of customers requiring different types of "servers," i.e., different services, products, or routes. Often, the underlying infrastructure is expensive, and hence so are the opportunity costs incurred when servers of one type are idle while others are congested. This cost can be reduced by introducing flexible servers that can serve multiple types of customers, but the cost of providing this flexibility may be very high. On the other hand, in many situations a proportion of the customers may be flexible, i.e., may be willing to change their type in order to reduce their time waiting for service, and the infrastructure to take advantage of this customer flexibility is often relatively inexpensive.

Consider a call center (in, say, California), which provides service in both English and Spanish. Callers currently have the option of pressing "1" for English and "2" for Spanish, but there are times when many Spanish speakers, for example, are on hold while all the Spanish speaking agents are busy, and yet there are idle English speaking agents. Because of the training expense, the cost of errors, and the high turnover of agents, agents are typically trained to only handle calls in one language. In such situations, I argue that the call center should add a "Press 0" option for bilingual customers willing to have their question answered in either language in exchange for reducing their waiting time. Note that this option has a small incremental infrastructure cost only, because it is taking advantage of flexibility that is already present in the customers.

There are many other examples of systems with partial customer flexibility. An example is the Mobile Millennium project for reducing traffic congestion at UC Berkeley, in which participating drivers collect data on current highway speed through GPS-enabled cell phones. The data is sent to a central system that provides information back to the participating drivers for personal use in choosing alternate routes (http://www.traffic.berkeley.edu/). A similar application is to communications and Internet routing, in which some but not all users have the ability to query alternate routes and use the shortest. In a make-to-order manufacturing context, some customers may not care, for example, about the color of the product they are ordering. Another application is to national border crossings with different queues for different nationalities, and where some customers may have dual citizenship.

Note that the flexibility I am considering is customer flexibility, not server flexibility. The latter has received a lot of attention in the operations research community, and in particular for call centers (Aksin et al., 2007); (Graves and Tomlin, 2003); (Hopp et al., 2004); (Hopp and van Oyen, 2004) and (Jordan and Graves, 1995). However, such flexibility is still generally expensive, particularly in terms of training costs. Customer flexibility, on the other hand, is often already present, but may not be exploited, and generally, it is

Righter R..
Taking Advantage of Partial Customer Flexibility - An Inexpensive Means of Improving Performance.
DOI: 10.5220/0004923403010304
In *Proceedings of the 3rd International Conference on Operations Research and Enterprise Systems* (ICORES-2014), pages 301-304
ISBN: 978-989-758-017-8

inexpensive to take advantage of customer flexibility.

In earlier research (Akgun et al., 2011; 2012; 2013) we have shown the benefit of partial flexibility in homogeneous systems, in which different groups of servers are stochastically identical. More work needs to be done, in terms of investigating the performance of, and developing protocols for, systems with heterogeneous server stations and user populations.

2 RESULTS FOR HOMOGENEOUS SERVERS

If flexible customers are given the option to choose which queue to join based on queue length, they would clearly join the shortest queue (JSQ) assuming server stations are homogeneous, service times are exponential, and the service discipline is FCFS. My co-authors and I showed the optimality of JSQ (join-the-shortest-queue) routing in a very strong, sample-path, sense (Akgune et al., 2011). The system is modelled as a queuing system with c parallel multiple-server stations that have exponentially distributed service times with the same service rate μ. All of these servers follow a nonidling but otherwise arbitrary service discipline (FCFS, LCFS, etc.). Arrivals to the system form an arbitrary process that is independent of the state of the system. Some (*dedicated*) arrivals are obliged to use a particular station, while others (*flexible*) have the ability to use any of the c stations (or, more generally, they can use an arbitrary subset of size at least two of the stations. Dedicated arrivals are equally likely to require any particular station, so the arrival process is homogeneous across stations. Let A be the set of arrival points, and let $F \subseteq A$ denote the time points where a flexible arrival occurs. Note that F is an arbitrary subset of A.

We used weak majorization and developed a new approach for coupling potential service completions to prove the optimality of JSQ (join-the-shortest-queue) in the sample path sense. We also showed that when flexible customers follow JSQ, the total number of customers in the system is stochastically decreasing in the proportion of flexible customers, so there is a strong advantage to having customer flexibility. Note that minimizing the total number in the system is equivalent to minimizing the mean waiting time from Little's law. We also showed that the waiting time for dedicated customers is decreasing in the proportion of flexible customers. That is, the monolingual customers, on average, benefit from having bilingual customers.

We also considered several practically important extensions. For example, suppose customers may abandon, but they only abandon from the queue (a reasonable assumption for, e.g., a call center), and suppose the abandonment rate is greater than the service rate. We showed, under these abandonment assumptions, that JSQ no longer minimizes the number of customers in the system, but it still maximizes the service completion process. Other extensions that we considered included finite buffers, resequencing, random yields, and randomly varying service rates.

While in Akgun et al., (2011) we showed that stationary waiting time is stochastically *decreasing* in the proportion of flexible customers, p, in Akgun, Righter, and Wolff (2012) we studied the *marginal* impact of customer flexibility, that is, the *convexity* of waiting time in p. Convexity means that the marginal advantage of flexibility is largest at small proportions. That is, roughly, "a little bit of flexibility goes a long way." Unfortunately, it is not possible to obtain convexity in the strong sense for which monotonicity holds. We considered a modified model, which we called the inventory model, where we obtained a sample-path convexity result using majorization of the queue lengths. In this model, there are two servers and they never idle but instead build up inventory when no customers are waiting. This may be reasonable in production environments where demand is high and where it is expensive to idle machines, e.g., due to high backorder costs or server shutdown costs. Although convexity in p is intuitive for our original model, in which servers idle when they have no customers, it does not hold in the same strong sense that monotonicity holds, and it is surprisingly difficult to prove. We developed a new approach that combines marginal analysis with coupling to show that the stationary mean waiting time is convex in p. We considered a tagged customer in steady-state that has lowest preemptive priority relative to the other customers so that the other customers are unaffected by the tagged customer. We showed that the derivative of the stationary waiting time with respect to p (the marginal value of customer flexibility) can be expressed in terms of the difference in expected waiting time between going to the long and the short queue for the tagged customer. We then showed, using another coupling argument, that this difference is decreasing in p.

Now consider a slightly different policy, where flexible customers "virtually" join all of the queues

for which they are eligible. Once a flexible customer enters service at one of the stations, its "virtual copies" are removed from the other stations. Again, this should not be hard to implement in many service systems, such as call centers. Such a policy is equivalent to JSW (join-the-smallest-work) for flexible customers. We showed in Akgun et al., (2013) that such a policy outperforms JSQ.

We were able to improve overall performance, as well as performance for flexible customers as a group and dedicated customers as a group, by moving from JSQ to JSW routing for the flexible customers. This leads us to ask whether we can do even better. Of course, the most efficient (overall) alternative for handling both flexible and dedicated customers is to maintain a separate queue for flexible customers, and to follow an optimal scheduling policy for each server. That is, to decide whether a given server should serve a dedicated or a flexible customer next. We showed in Akgun et al., (2013) that the optimal policy, under a range of fairly general conditions, is DCF (serve-dedicated-customers-first). This policy indeed outperforms both JSQ and JSW in terms of minimizing overall mean waiting time, and is especially good for dedicated customers. However, it is unfair for flexible customers, and, unlike JSQ and JSW, is not incentive compatible for them (it is not the policy that they would choose for themselves).

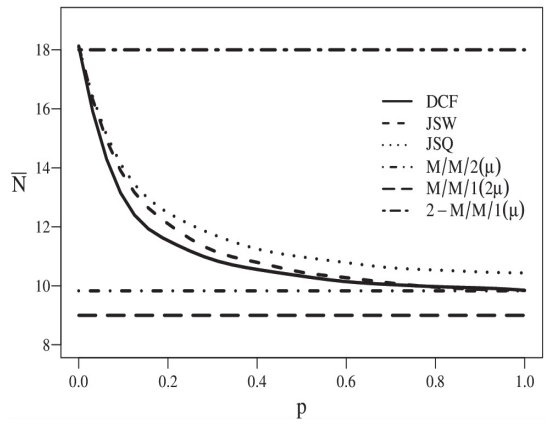

Figure 1: Comparison of policies. Total number in system vs. proportion of flexible customers (p).

In Figure 1 we show simulation results for two M/M/1 queues with overall traffic intensity ρ = .9, and where $\overline{N}$ denotes the long run average number of customers in the system. This is directly proportional to mean waiting time through Little's law. When none of the arrivals are flexible (p = 0), this system becomes two separate M/M/1 queues with service rate μ (the upper bound line). On the other hand when all customers are flexible (i.e. p = 1), we see that under JSW and DCF, the mean number of customers in the system converges to that for an M/M/2 queue with each server having a service rate of μ. In other words, when customers are fully flexible, the system performance is the same as the performance for the generally more expensive system in which servers are fully flexible. The lower bound is a single-server system with twice the service rate (M/M/1(2μ)), which would represent an ideal (and generally unattainable) pooled scenario where servers can collaboratively serve each customer with no loss in efficiency.

We see from Figure 1, as expected, that DCF outperforms JSW, which in turn outperforms JSQ. Note, however, that the DCF performance is not much better than that of JSW, and, as mentioned before, it is unfair to flexible customers. Therefore, the best overall policy is JSW. The figure also clearly shows the convexity of performance. In particular, we have an "80-20 rule" where at about p = 20% we have about 80% of the benefit relative to the total benefit that could be obtained by going from p = 0 to p = 1.

3 FUTURE WORK

Results for homogeneous stations clearly indicate the benefit of exploiting customer flexibility. Much work remains to be done to study systems with heterogeneous servers and multiple classes of customers.

REFERENCES

Akgun, O., Righter, R., and Wolff, R., 2011. Multiple Server System with Flexible Arrivals. *Advances in Applied Probability* 43: 985-1004.

Akgun, O., Righter, R., and Wolff, R., 2012. Understanding the Marginal Impact of Customer Flexibility. *Queueing Systems* 71: 5-23.

Akgun, O., Righter, R., and Wolff, R., 2013. Partial Flexibility in Routing and Scheduling. *Advances in Applied Probability* 45: 673-691.

Aksin, O. Z., Karaesmen, F. and Ormeci, E. L., 2007. A review of workforce cross-training in call centers from an operations management perspective. In *Workforce Cross Training Handbook*, ed. D. Nembhard. CRC Press.

Graves, S. C. and Tomlin, B. T., 2003. Process flexibility in supply chains. *Management Science* 49: 907–919.

Hopp, W. J., Tekin, E. and van Oyen, M. P., 2004.

Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science* 50: 83–98.

Hopp, W. J. and van Oyen, M. P., 2004. Agile workforce evaluation: A framework for cross-training and coordination. *IIE Transactions* 36: 919–940.

Jordan, W. C. and Graves, S. C., 1995. Principles on the benefits of manufacturing process flexibility. *Management Science* 41: 577–594.