On Accommodating Customer Flexibility in Service Systems

Yu-Tong He
Department of Computing and Software
McMaster University
1280 Main Street West, Hamilton, ON L8S 4L7, Canada
hey3@mcmaster.ca

Douglas G. Down
Department of Computing and Software
McMaster University
1280 Main Street West, Hamilton, ON L8S 4L7, Canada
downd@mcmaster.ca

Abstract

We consider simple parallel queueing models in which a proportion of arriving customers are *flexible*, i.e. they are willing to receive service at any one of some subset of the parallel servers. For the case of two parallel servers, we show that as the servers become fully utilized, the maximum improvement in mean waiting times is achieved for arbitrarily small levels of flexibility. The insights from this analytic model are supported by simulation results that show that large gains can be made with low levels of flexibility. The potential implications of these results for two motivating examples are discussed.

1 Introduction

There has been a large amount of work done in recent years on flexibility in queueing systems, and its potential benefits for performance. The vast majority of this work has been focused on how to exploit cross-trained servers or workers, either by choosing appropriate cross-training structures (see for example [10, 14, 18]), or through effectively utilizing an existing flexibility structure to improve performance measures such as throughput or holding costs (see for example [3, 5, 9, 15]). One area where this has been of much interest is in the call centre literature, which is by now quite extensive. The reader is referred to the comprehensive surveys in Aksin et al. [1, 2] and Gans et al. [9]. Of course, these two viewpoints are not mutually exclusive, typically one has to have some idea of how to effectively utilize a given set of flexible servers in order to actually design effective structures. The references above are far from exhaustive. We have chosen not to include a comprehensive list as this kind of flexibility is not the focus of this paper. Readers interested in the topic of flexible servers could use the reference lists in the works cited in this paragraph as a useful starting point.

In this paper, we would like to provide some initial thoughts on a different form of flexibility. Suppose that arriving customers are flexible, in that they might be willing to receive service in different manners. To make this a little more clear, it is probably instructive to discuss the two examples which motivated our interest.

1. In Canada, some call centres offer services in both French and English. On making a call to a call centre, the first choice one typically makes is a language choice. This struck us as somewhat unfair to bilingual customers, as they have to make a choice (usually with no information). We wondered what would happen if callers were provided with waiting time information, which they could use to make their language choice. As in the previous example, we are interested in the case when a (small) minority of customers would avail themselves of this option. (It may not be the case that all bilingual customers would choose to follow this option, even if it were available, so the flexible customers would be a subset of those that are bilingual. There may be other overriding considerations such as perceived quality of service in a particular language.)

2. In Ontario, Canada, there has been a recent initiative to publish waiting times for procedures (such as MRIs) at hospitals [21]. While this gives patients more information to make choices, there is doubt about how much benefit there will actually be, as there is a belief that not many patients will take advantage of this choice due to issues of travel costs for patients and reluctance of doctors to make appropriate referrals [6]. On reading [6], as the majority of patients may be dissuaded from acting on the waiting time information, we wondered what would happen if only a (small) fraction of patients availed themselves of such choice.

In order to get some insight into this issue, we examine a simple model, in which there are two queues in parallel. *Dedicated* arrivals occur to each of the two queues (by dedicated we mean that those arrivals will only accept service at that queue). The servers may have different service rates. In addition, there are *flexible* customers, who are willing to accept service at either queue.

Note that if all of the customers were flexible and willing to accept service at any of the queues, then under appropriate distributional assumptions, joining the shortest queue is an effective policy [25, 26]. In a two queue system where a fraction of the total arrivals joins the shortest queue (JSQ), Reiman [22] uses the methodology of diffusion approximations to study heavy traffic behaviour and gives conditions under which the joint queue length process collapses to a single dimension (so-called "state space collapse") and thus the system becomes statistically indistinguishable from a system in which *all* customers join the shortest queue. We adapt (the modifications are slight) his work to provide analytic insight into our question. Recently, in [12], we looked at larger systems (i.e. an arbitrary number of queues in parallel) and provided conditions under which state space collapse occurs. Foley and McDonald [7] also look at a two queue system, with all underlying distributions exponential. They are interested in studying large deviations (rare event) behaviour, providing conditions under which the rate at which the total queue length reaches a large level is the same as one in which all of the customers join the shortest queue.

It may be useful to mention another line of work that studies a different version of the problem in which we are interested. Suppose that all of the customers are flexible, but that each arrival is only allowed to choose a random subset of the servers to join, and joins the shortest queue within that random subset. The fact that the performance of a system where the subset is small can be close to one in which an arrival joins the shortest of all queues, is the subject of [4, 19], amongst others.

Our main observation is that if customer flexibility is accommodated, a significant decrease in mean waiting times may occur. An important point to note is that under certain conditions, such improvements are seen for *all* customer types in the system, not just the flexible ones. We also discuss implementation issues and suggest that if state information (or good estimates thereof) is not available in near real-time, performance may degrade considerably. We believe that this may be significant for our second motivating example.

The organization of the paper is as follows. Section 2 introduces the *bilingual customer model*. For this model, the application of results from [22] provide insight into how limited flexibility, if accommodated, can yield substantial performance improvement. This is supported by simulation results. Section 3 looks at the second motivating example and provides a cautionary note on the issue of the availability of real-time state information. Section 4 discusses the applicability of our insights for larger systems. Section 5 provides concluding remarks.

2 Bilingual Customer Model

Here, we study a model where there are two languages in use, which, using the terminology of [24], we will call the majority and minority languages. The following problem is considered in [24]. There are arrivals of monolingual customers, according to mutually independent Poisson processes of rate λ_1 for the majority language and λ_2 for the minority language. There are N servers who are capable of serving only majority language customers and M servers that are bilingual. They consider the problem of choosing the (minimal) number of bilingual servers and in addition, issues of how the bilingual servers should be assigned to customers.

Here, we turn the problem on its head. We suppose that we start with a system, that as above, has

arrivals following mutually independent Poisson processes with rates λ_1 and λ_2, corresponding respectively to the majority and minority language groups. These are held in separate queues, which we correspondingly label queue i, $i = 1, 2$. There is a single server for each language, where the service times at queue i are independent and identically distributed, with mean $1/\mu_i$ and variance s_i^2. We will comment later on why the restriction to a single server at each queue is not restrictive. Service at each queue is First Come First Served (FCFS). We will call this the *monolingual* system. Note that there may actually be bilingual customers, but in this system they have effectively been made monolingual, perhaps by the mechanism of having to make an initial language choice, with no information. (One could argue that bilingual customers make their choice based on past experience, so perhaps they in fact are using information. Modelling such a case would be of interest.)

We are interested in comparing the performance of the monolingual system with one where the service mechanism remains FCFS at each queue, but now a proportion of each of the original arrival streams are bilingual customers, in that they join the shortest of the two queues. (Note that this is not the best they can do if $\mu_1 \neq \mu_2$. In fact, it would be better to estimate their waiting times.) To be precise, with probability p_i, an arrival to queue i in the dedicated system is bilingual and joins the shorter of the two queues. We call this the *partially bilingual system*. We are interested in seeing what improvement can be seen if we move to this partially flexible setting, in particular whether significant improvement can be gained when $p_1 + p_2$ is small, i.e. if only a small proportion of arrivals are bilingual.

2.1 Analysis

In this section, we use the methodology of diffusion limits. In particular, the results here are not particularly novel. The results are in some sense a recast version of those in Reiman [22], bringing out the impact of making customers more flexible. To do this, we first construct the diffusion limit for the monolingual system and from that, construct the diffusion limit for the flexible system. We then compare performance measures of the appropriate diffusion limits to develop our insights. Let us first consider a sequence of monolingual systems, indexed by n. Assume that the following conditions hold:

$$\lim_{n\to\infty} \lambda_i(n) = \lambda_i, \tag{2.1}$$

$$\lim_{n\to\infty} \mu_i(n) = \mu_i, \quad \lim_{n\to\infty} s_i^2(n) = s_i^2, \tag{2.2}$$

where the limits are all finite. In addition, the heavy traffic condition

$$\lim_{n\to\infty} \sqrt{n}\Big(\lambda_i(n) - \mu_i(n)\Big) = c_d, \quad |c_d| < \infty, \tag{2.3}$$

is assumed to be true at each queue i, $i = 1, 2$.

From Theorem 3.3 of [22], we know that if the waiting time in the nth system for the kth arrival to the ith queue, $W_i^{d,(n)}(k)$ is scaled as

$$\hat{W}_i^{d,(n)}(t) = n^{-1/2} W_i^{d,(n)}(\lfloor \lambda_i n t \rfloor), \tag{2.4}$$

then $\hat{W}_i^{d,(n)}(t)$ converges (in the weak sense) to a reflected Brownian motion, which we will denote $\hat{W}_i^d$. Let $\bar{w}_i^d$ be the mean of the stationary distribution of $\hat{W}_i^d$. Using the expression for $\hat{W}_i^d$ in Theorem 3.3 in

[22], we have

$$\bar{w}_i^d = \frac{1}{2|c_d|\lambda_i}\left[\lambda_i^3 + \mu_i^3 s_i^2\right], \quad i = 1, 2. \tag{2.5}$$

For the partially bilingual system, the type i customers (type 1 are majority language, type 2 minority language) are partitioned into two sub-types. The customers that remain dedicated to queue i are referred to as type i_1 and the corresponding bilingual customers as type i_2. The arrival rate $\lambda_{i_j}^f$ for the corresponding Poisson process for type i_j customers, $i, j = 1, 2$, is given by

$$\lambda_{i_1}^f = (1 - p_i)\lambda_i, \quad \lambda_{i_2}^f = p_i\lambda_i.$$

Now let us consider a sequence of partially bilingual systems where the nth system corresponds to the nth monolingual system. We will assume that $p_i > 0$ for $i = 1, 2$. If p_i for $i = 1, 2$ is independent of n, then

$$\lim_{n\to\infty} \lambda_{i_j}^f(n) = \lambda_{i_j}^f, \quad i, j = 1, 2, \tag{2.6}$$

$$\lim_{n\to\infty} \mu_i(n) = \mu_i, \quad \lim_{n\to\infty} s_i^2(n) = s_i^2, \quad i = 1, 2, \tag{2.7}$$

where the limits are all finite. In addition, we also require the heavy traffic condition

$$\lim_{n\to\infty} \sqrt{n}\left(\sum_{i=1}^{2}\sum_{j=1}^{2} \lambda_{i_j}^f(n) - \sum_{i=1}^{2} \mu_i(n)\right) = c_f, \quad |c_f| < \infty. \tag{2.8}$$

It is easy to see that (2.8) follows from (2.3) with $c_f = 2c_d$.

Theorem 5.3 of [22] shows that under the scaling given in (2.4), the waiting time process for type i_j customers for the partially flexible system converges (weakly) to a reflected Brownian motion, which we denote by $\hat{W}_{i_j}^f$. From the form of $\hat{W}_{i_j}^f$ in Theorem 5.3 of [22], we have that its mean, $\bar{w}_{i_j}^f$, satisfies

$$\bar{w}_{i_j}^f = \frac{1}{4|c_d|(\lambda_1 + \lambda_2)}(\lambda_1 + \mu_1^3 s_1^2 + \lambda_2 + \mu_2^3 s_2^2), \quad i, j \in \{1, 2\}. \tag{2.9}$$

Note that (2.9) is independent of p_i, so that $\bar{w}_{i_j}^f$ for a partially bilingual system is the same as for the fully bilingual system. For the partially bilingual system to show improvement for all customer types, we require $\bar{w}_i^d/\bar{w}_{i_j}^f > 1$, for all i, which is equivalent to

$$\frac{1}{2 + \frac{\lambda_1}{\lambda_2}} < \frac{1 + \mu_1^2 s_1^2}{1 + \mu_2^2 s_2^2} < 2 + \frac{\lambda_1}{\lambda_2}. \tag{2.10}$$

There are two special cases where (2.10) is always satisfied that are worth mentioning, even though (2.10) is easy to check. The first is if $\mu_1 = \mu_2$ and $s_1 = s_2$. The second, perhaps more interesting case, is when the service time distributions are both exponential.

What does all of this mean practically? If one backs away from heavy traffic, while there would be a gap between the performance of the partially bilingual system and the fully bilingual system, the results here suggest that most of the gap is made up with a small amount of flexibility. (This can be strengthened by noting that a formal approximation of system performance may be constructed using the diffusion limit.) We will see this to be the case for the simulation studies in the next subsection.

We end this subsection with a comment on our assumption of a single server at each queue. We believe that this is not restrictive. Iglehart and Whitt [16, 17] under appropriate conditions show that the diffusion limits for a G/G/s queue with generic service time V are identical to the corresponding diffusion limits for a G/G/1 queue with generic service time V/s, so (at least in the heavy traffic limit), we can replace s homogeneous servers by a single server that works s times as fast.

2.2 Simulation Results

In our simulation work, we try to give some idea of the performance improvement that can be achieved as one backs off from heavy traffic. We see here that, as predicted by the analytic results in the previous subsection, substantial improvements are achieved if a small proportion of arrivals are bilingual, and the fact that they are bilingual is accommodated.

We adapt the simulation study in [24]. We begin by looking at a four server system, where three servers are majority language, the fourth is minority language. Here, the arrivals join the queue with the shortest expected waiting time. The service time distribution for either language is the same, taken to be exponential with mean 25.0. Note that this means that (2.10) is satisfied. We chose the arrival rates such that for the monolingual system, the loads on the servers are either 0.85 or 0.95. This corresponds to $\lambda_1 = 0.102$, $\lambda_2 = 0.034$ and $\lambda_1 = 0.114$, $\lambda_2 = 0.038$, respectively. For the corresponding partially bilingual systems, we let $p_2 = \min(3p_1, 1.0)$ and vary p_1. All of our simulations have an accuracy of no worse than 5 percent at a 95 percent confidence level. In the tables below, W_i is the mean waiting time at queue i. Tables 1 and 2 give the results.

p_1	0	0.05	0.10	0.20	1.00
W_1	68.6	64.5	63.2	60.6	54.9
W_2	175.7	102.6	80.9	64.2	59.8

Table 1: Four server system, load 0.85

p_1	0	0.05	0.10	0.20	1.00
W_1	175.2	162.4	154.0	145.7	132.9
W_2	595.7	209.2	166.9	145.1	137.7

Table 2: Four server system, load 0.95

Several observations can be made based on Tables 1 and 2. First note that the main effect here is to substantially improve the performance at the minority language server while not impacting the performance at the majority language servers (in fact, there is also a slight improvement in the performance at the majority language servers). This is consistent with (2.10) being satisfied. Note that in the higher load case, we see dramatic improvements under very low levels of bilingual customers. In fact, looking at the entry for $p_1 = 0.05$ ($p_2 = 0.15$) in Table 2 and comparing it to the entry with $p_1 = 1.00$ (the fully bilingual case), we see that we get just under 85 percent of the possible improvement at the minority language server. Under

more moderate loads, it is plain to see that the impact is not as significant, but still substantial. Finally, $p_1 = 1.00$ corresponds to a fully bilingual customer base, but the performance here is the same as if the servers were all bilingual. In other words, it is a lower bound not just for our model, but is a global lower bound.

To see how the insights scale, we increased the size of the system by a factor of two. In other words, we simulated a system with 8 servers, 6 majority language, the remainder minority language. The service time distribution was kept to be exponential with mean 25.0. The arrival rates were both increased by a factor of two. The results are presented in Tables 3 and 4.

p_1	0	0.05	0.10	0.20	1.00
W_1	42.8	41.5	40.6	40.1	37.8
W_2	89.0	57.5	48.9	44.4	40.2

Table 3: Eight server system, load 0.85

p_1	0	0.05	0.10	0.20	1.00
W_1	90.8	89.4	86.0	82.1	77.5
W_2	249.0	108.2	91.1	81.4	80.0

Table 4: Eight server system, load 0.95

There is little to say here other than the observations follow the same pattern as for the four server case.

2.3 Discussion

Our model assumes that all of the servers are monolingual. The problem of designing scheduling policies when a subset of servers is bilingual is studied in [23, 24]. It is expected that the benefit of accommodating customer flexibility would decrease as more servers are bilingual. At the extreme, it is easy to see that there is no need to accommodate customer flexibility if *all* servers are bilingual. Note that the results above show that at another extreme (heavy traffic), bilingual servers would not be required if (a small degree of) customer flexibility were accommodated. At more moderate loads, the combination of both server flexibility and customer flexibility and their relative impacts is an issue that we believe is worth study. The potential tradeoffs would require a more realistic model that captured economic considerations.

We also looked at the case where $p_1 = 0$ and p_2 varied. Note that we do not have analytic results for this case, as we do not have state-space collapse in the heavy traffic limit. As a result, we cannot expect performance to improve for the majority language customers, but we do find that for small values of p_2, that there is a very small increase in their mean waiting times, while there are dramatic improvements in the mean waiting times for the minority language customers, with magnitudes in line with those seen in Tables 1-4.

3 Need for Real-Time Information: Second Example

Here, we consider the problem of a two facility system. The rate at which the procedures can be performed at facility i is μ_i (one facility may have more capacity than another, for example). Suppose that currently, the demand (arrival rate) for facility i is λ_i. We wish to look at the situation when a proportion p_i of each of the demand streams is flexible, in that they are willing to go to the facility in which they expect to wait the least.

We immediately see that this is simply the bilingual caller model, if the queue lengths are used as the waiting time estimates for the flexible servers. So, as discussed in the previous section, the immediate conclusion is that it would be very useful to accommodate this flexibility. However, this is a very abstract model, and a key issue that we have conveniently overlooked is implementability (we will say more later). We should really recommend accommodating customer flexibility *as long as JSQ can be implemented*. For example, it may be possible in the call centre setting, but as of the writing of this article, waiting time estimates were being updated on a monthly basis at the Ontario Waiting Times Strategy site [21]. So, we now examine the importance of being able to implement JSQ, or in other words, the need for real-time state information.

To this end, we performed another simulation study. We chose a system with $\mu_1 = \mu_2 = 20.0$ and $\lambda_1 = \lambda_2 = 19.0$. This system is highly loaded (without adding flexibility, the load is 0.95 at each queue) and the load at each server is balanced. If $p_1 = p_2$, then we have the mean waiting time W, as given in Table 5. Note that entries for $p_1 = 0.0$ are not simulated; they are simply M/M/1 results. The remaining entries have an accuracy of no worse than 5 percent at a 95 percent confidence level.

p_1	0	0.05	0.10	0.20	1.00
W	1.00	0.72	0.63	0.57	0.52

Table 5: Base system mean waiting times

The improvement in heavy traffic is a factor of two, and as expected, most of the theoretical maximum improvement is achieved if only a small number of arrivals use the expected waiting time information (58 percent if 5 percent of arrivals are flexible, 76 percent if 10 percent are flexible).

We now turn to the question of how significantly does the situation degrade if we have alternate state estimates? In particular, we look at the case where the mean waiting time is updated periodically, with all arrivals using the most recent update.

To be precise, we performed a simulation study of the base system (results in Table 5), but with flexible arrivals who instead of joining the shortest queue, join the queue with the shortest average wait, where the average wait at each server is updated every T time units and is calculated as the average waiting time at each queue for all arrivals since the previous update. Simulations were run for 50,000 units of simulated time, with the results in Table 6, each entry being the simulated mean waiting time, except for $p_1 = 0.0$, which was obtained analytically.

Here, $T = 0.1$ corresponds to updating average waiting times every 3.8 arrivals, on average. We see here that there is some small degradation of the performance over that of the base system, with a small amount

T \ p_1	0.0	0.05	0.1	0.2	1.0
0.1	1.00	0.77	0.67	0.61	0.58
0.2	1.00	0.78	0.68	0.63	0.67
0.5	1.00	0.78	0.69	0.68	0.90
1.0	1.00	0.77	0.71	0.71	1.39
2.5	1.00	0.81	0.77	0.86	2.74
5.0	1.00	0.85	0.89	1.13	4.92
10.0	1.00	0.90	1.02	1.61	9.45
20.0	1.00	1.01	1.33	2.63	18.15

Table 6: Periodic wait time updates, partial flexibility

of flexibility still leading to significant gains. In fact, the proportion of overall improvement (over $p_1 = 1.0$) is mostly greater for low levels of flexibility. As T increases, the degradation increases, with the degree of degradation being magnified by increasing p_1. With a little thought, this is not particularly surprising - during each interval, all flexible customers join one queue, the one with the lower estimated mean wait, until the next update. If $p_1 = 1.0$ (all customers flexible), this creates very large oscillations in queue lengths (arrivals "herd" together to one queue for a period of time). In the last row of Table 6, corresponding to an update every 760 arrivals (on average), we see that this herding effect is prominent enough that even at low levels of flexibility, the system would be better off if no information were provided, and the flexible customers flipped a coin to choose their server.

3.1 Practical implications

We believe that the practical implications of the above model are clear, even if the observations are for a simple model. However, before we discuss these implications, we would first like to discuss limitations of our model.

1. Reneging is not included in the model. Certainly in any service system, including reneging is important to get precise performance results. However, we do not see that any of the qualitative effects that we have described would change if reneging were included, although the numbers themselves would change.

2. Arrival processes may be load dependent. Once again, we believe that the fundamental effects would still be present.

3. Service may not be FCFS. If the service time distributions were exponential, the mean waiting time would not change, so there would be no difference in the results. In general, we believe that the fundamental effects would still be present.

4. Interarrival and service times are not in general exponentially distributed. The magnitude of improvement would be more if more variable distributions were used (less if less variable distributions were used), but the relative effects would be in most cases unchanged (see He [11] for more details).

With this in mind, we believe the insight for system designers is clear. In a setting where one cannot add new resources, accommodating customer flexibility can have an improvement that is very dramatic at high loads, so as such, if this can be easily implemented, it can be quite worthwhile. However, there is the strong caveat that updates of waiting time estimates should happen at a frequency that is close to the arrival rate (at least within an order of magnitude), or else problematic effects could occur.

As a concrete example of these insights, we take the Ontario Wait Times Strategy [21]. As of the writing of this article, wait times are being updated on a monthly basis. We believe that in the light of our insights, the strategy should be carefully studied to see if the publication of wait time information could cause the wait times to actually increase over the case where they are not made public. In particular, if any significant proportion of people use the information to make choices, the update interval must be chosen in an appropriate manner. We are not saying that one month is definitely too long, but wonder if the effects discussed in this paper have been considered. We note that if updates are made at a sufficiently high frequency, significant benefits should arise, even if only a small fraction of the population uses the information to make choices.

The potentially detrimental effects of using periodically updated state information has been known for some time in the Computer Science literature. In particular, Mitzenmacher [20] looks at a system where actual queue length information is updated periodically for a system with N queues in parallel, where every arrival chooses one of d ($2 \leq d \leq N$) queues at random and joins the shortest one. The effects are similar to those described here. An instance of this phenomenon arising in practice is described in Fox et al. [8].

4 Larger Systems

One criticism of the work presented to this point could be that it is only looking at a two queue system. We have simply done this for ease of exposition. The effects that we discussed for two queues carry over to the many queue setting. In [12], there are theoretical results that show that for a system of N identical servers in parallel, if a proportion of arrivals $(1-p)/N$ is dedicated to each of the servers and with probability $p/(N-1)$, an arrival joins the shorter of queues j and $j+1$, $j = 1, \ldots, N-1$, then, for any $p > 0$, the heavy traffic limit of this system is the same as if *all* of the arrivals join the shortest of *all* of the queues. Note that this builds locality considerations into the model, such as may occur when customers make choices using geographic considerations, for example. Further extensions to more general settings, including multiple classes of arrivals, heterogeneous servers and different models for locality are presented in [11, 13]. We see no need to present details here, we simply want to make the point that there are theoretical results that suggest that at high loads, accommodating even a small amount of customer flexibility should have significant impact for very general systems. In addition, it should be clear that the issue of using periodically updated information will not go away in these larger systems, the herding phenomenon will still exist.

5 Conclusion

The results presented above suggest that if one has arriving customers that are flexible, there is the potential for significant benefit if this flexibility is accommodated.

In terms of implementation, in many cases we would expect that accommodating this flexibility would be easy and incur little cost. The issue that arises here (in particular with respect to the hospital example) is whether enough incentives can be built in to encourage enough customers to be flexible and thus allow *all* customers to reap the benefit. We think that this (which would require examining more complicated models) would be of great interest. There is the final caveat that the maximum performance improvement (in relative terms) increases as the overall system load increases. Even with customer flexibility accommodated, a highly loaded system may have unacceptably long queues and the only choice may to be add resources. However, on the positive side this work does suggest that in systems that are heavily loaded, if there are constraints that make adding resources difficult (due to cost and/or availability), then this simple mechanism can help to yield significant performance improvement.

6 Acknowledgements

This work is supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] O.Z. Aksin, M. Armony, and V. Mehrotra. The modern call-center: A multi-disciplinary perspective on Operations Management research. *Production and Operations Management*, 16:665-688, 2007.

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

[3] S. Andradóttir, H. Ayhan and D.G. Down. Dynamic server allocation for queueing networks with flexible servers. *Operations Research*, 51:952-968, 2003.

[4] Y. Azar, A.Z. Broder, A.R. Karlin and E. Upfal. Balanced allocations. *SIAM Journal on Computing*, 29:180-200, 1999.

[5] A. Bassamboo, J.M. Harrison and A. Zeevi. Design and control of a large call center: asymptotic analysis of an LP-based method. *Operations Research*, 54:419-435, 2006.

[6] R. Ferguson. Skeptics doubt wait-times website will help. *Toronto Star*. A8, October 25, 2005.

[7] R.D. Foley and D.R. McDonald. Join the shortest queue: stability and exact asymptotics. *Annals of Applied Probability*, 11:569-607, 2001.

[8] A. Fox, S.D. Gribble, Y. Chawathe, K.A. Brewer and P. Gauthier. Cluster-based scalable network services. *Proceedings of the 16th ACM Symposium on Operating Systems Principles*, 78-91, 1997.

[9] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5:79-141, 2003.

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

[11] Y-T. He. *Exploiting Limited Customer Choice and Server Flexibility.* Ph.D. thesis, McMaster University, 2008.

[12] Y-T. He and D.G. Down. Limited choice and locality considerations for load balancing. *Performance Evaluation*, 65:670-687, 2008.

[13] Y-T. He, I. Al-Azzoni, and D.G. Down. MARO - MinDrift affinity routing for resource management in heterogeneous computing systems. Proceedings of CASCON 2007, 71-85, 2007.

[14] W.J. Hopp, E. Tekin, and M.P. van Oyen. Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science*, 50:83-98, 2004.

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

[16] D.L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, 2:150-177, 1970.

[17] D.L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. *Advances in Applied Probability*, 2:355-369, 1970.

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

[19] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Computing*, 12:1094-1104, 2001.

[20] M. Mitzenmacher. How Useful is Old Information? *IEEE Transactions on Parallel and Distributed Systems*, 11:6-20, 2000.

[21] Ontario Ministry of Health and Long-Term Care. MOHLTC - Wait Times - Ontario Wait Times Strategy. http://www.health.gov.on.ca/transformation/wait_times/wait_mn.html (accessed June 20, 2008).

[22] M.I. Reiman. Some diffusion approximations with state space collapse. In Lecture Notes in Control and Information Sciences, volume 60, pages 209-240, Springer, Berlin-New York, 1984.

[23] D.A. Stanford and W.K. Grassmann. The bilingual server system: a queueing model featuring fully and partially qualified servers. *INFOR*, 31:261-277, 1993.

[24] D.A. Stanford and W.K. Grassmann. Bilingual server call centres. in *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D.R. McDonald and S.R.E. Turner, editors. Fields Institute Communications, 31-47, 2000.

[25] R.R. Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, 15:406-413, 1978.

[26] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14:181-189, 1977.