J. Appl. Prob. **43,** 274–281 (2006)
Printed in Israel

USER-OPTIMAL STATE-DEPENDENT ROUTEING IN PARALLEL TANDEM QUEUES WITH LOSS

SCOTT SPICER * AND
ILZE ZIEDINS,* ** *The University of Auckland*

Abstract

We consider a system of parallel, finite tandem queues with loss. Each tandem queue consists of two single-server queues in series, with capacities C_1 and C_2 and exponential service times with rates μ_1 and μ_2 for the first and second queues, respectively. Customers that arrive at a queue that is full are lost. Customers arriving at the system can choose which tandem queue to enter. We show that, for customers choosing a queue to maximise the probability of their reaching the destination (or minimise their individual loss probability), it will sometimes be optimal to choose queues with more customers already present and/or with greater residual service requirements (where preceding customers are further from their final destination).

Keywords: Parallel queues; tandem queues; loss probability; routeing; user optimality

2000 Mathematics Subject Classification: Primary 90B15
Secondary 60K25

1. Introduction

We consider a collection of parallel, finite tandem queues with loss. Each tandem queue consists of two single-server first-in–first-out queues in series, with capacities C_1 and C_2 and service rates μ_1 and μ_2 for the first and second queues, respectively. The results below hold for C_1 both finite and infinite, but we always have $C_2 < \infty$. Customers that arrive at a queue that is full are lost. Customers arrive at the system at rate λ, and on arrival can choose which tandem queue to join. Once they have joined a tandem queue, they progress through it, obtaining an exponentially distributed service at each stage. The interarrival and service times are all independent of one another. If, when a customer completes service, the next queue in the tandem series is full, then the customer is lost to the system. We will assume that, at the point when the customer chooses which tandem queue to join, it has full knowledge of the number of customers in each queue and chooses a route that will minimise its individual loss probability over the whole route. Figure 1 illustrates a system with three tandem queues. Although, to fix our ideas, we think of customers choosing a tandem queue upon arrival to the system, the results below also apply to situations in which jockeying is permitted, that is, customers are free to switch queues after arrival.

Two natural heuristics for parallel tandem queues, when trying to minimise sojourn times, might be that (i) a customer should choose a tandem queue with the fewest total number of customers in it; and that (ii) if there is a choice of two routes, with the same number of customers on both but with one or more customers on the second route being closer to their destination than any of those on the first, then the new arrival should choose the second route. We show

Received 28 February 2005; revision received 10 November 2005.
* Postal address: Department of Statistics, The University of Auckland, Private Bag 92019, Auckland, New Zealand.
** Email address: ilze@stat.auckland.ac.nz

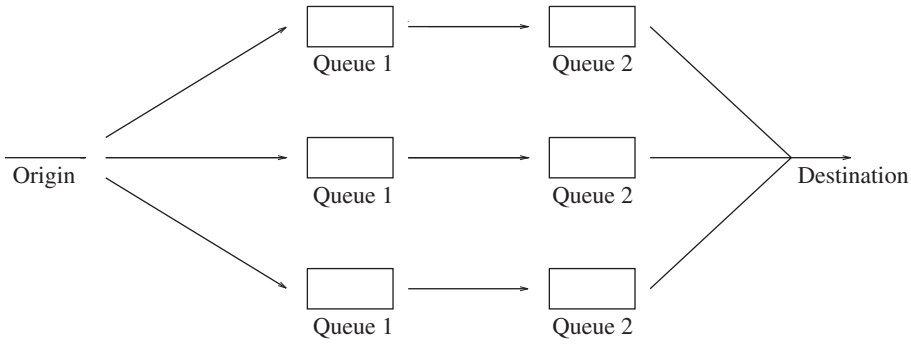

FIGURE 1: Three parallel tandem queues.

that, in general, neither of these heuristics gives the lowest individual loss probabilities. When minimising loss probabilitites, it may be optimal for customers to choose routes with a greater number of customers, or on which preceding customers are further from their final destination (i.e. have greater residual service requirements). Indeed, we show that, given a route with i customers ahead of the marked customer in queue 1, and j customers in queue 2, a lower loss probability is always obtained by travelling via a route with $i + l$ customers in the first queue and $j - l$ in the second, for $l \leq \min(j, C_1 - i - 1)$. Thus, paradoxically, beyond a certain point, increased congestion can actually decrease loss for the marked customer, by delaying its exposure to it. If there is the potential to be lost, then it is best to delay service (see, e.g. [9] and [16] for examples of other situations in which it may be optimal to delay customers).

Tandem queues, with both infinite and finite capacity, have received considerable attention in the literature (see, e.g. [1], [9], [14], and the references therein). Queues with finite capacity have the problem of how to treat customers that finish service at one queue but find the next queue full. The common assumption is that customers are never lost within the system, but block the server at their current queue until they are ready to move – see [5] and the more recent work [14] for a general framework for such blocking. However, for networks such as the Internet, a model that assumes customers to be lost if the receiving queue is full is more appropriate (see, e.g. [3]). Files or messages sent over the Internet are broken up into smaller components, called packets, before being transmitted. Each packet travels separately (possibly on a different route) and they are then reassembled into the complete file or message at the destination. As transmission proceeds, the destination node sends confirmation that packets have arrived. If the source fails to receive such a confirmation then packets are re-sent from the source. Thus, in the Internet, it is the source node that re-sends packets if they are lost, or delayed for too long, en route – not any intermediate node. Very little work has been done on tandem queues with loss – below we discuss [9], in which a two-stage tandem queue with parallel servers was studied.

In this paper we characterise some properties of the individually optimal route for an arriving customer presented with a choice of disjoint tandem routes, such as might be present in the Internet. This problem is an extension of the classical 'choice of parallel queues' problem, which has also received considerable attention in the literature, beginning with Winston [20], who showed that the 'join the shorter queue' policy is globally (or socially) optimal. Walrand [17, pp. 260–264] gave a nice introduction to this area. The 'join the shortest queue' policy has been found to be globally optimal under various criteria and in various settings for both infinite and finite queues ([7], [8], [10], [12], [13], and [18] are just a small selection of the many papers

https://doi.org/10.1239/jap/1143936259 Published online by Cambridge University Press

that have appeared). There are, however, interesting exceptions to the optimality of the shortest queue policy – for instance, Whitt [19] showed that it may not be optimal if service times are not exponential.

Optimal routeing to separate tandem queues has not been previously considered. Hordijk and Koole [9] considered a two-stage tandem queue with multiple servers at each stage and finite buffers with loss. In their model, routeing decisions for customers are made at both the first and the second stages (whereas our model only allows a routeing decision to be made at the first stage). They found that the shortest queue policy is optimal at the second stage, and showed that it is optimal or close to optimal for the first stage.

In this paper we consider only individually optimal routeing policies; there are many examples of such policies not being socially (globally) optimal. Bell and Stidham [2] gave a nice discussion of this, and other examples can be found in, e.g. [6], [4], and [19]. The unusual feature of the individually optimal policies in this paper is that individuals may find it optimal to choose routes that are more congested (whereas usually the individually optimal policies select routes that are *less* congested, and may thereby induce worse overall performance in the system).

We state and prove our main result in Section 2. In Section 3 we give some illustrative examples and a brief concluding discussion.

2. Optimal routeing

Consider a single tandem queue. Let n_i, $i = 1, 2$, be the number of customers in the ith queue, including any customer being served. Let $\boldsymbol{n} = (n_1, n_2)$ and let $\boldsymbol{e}_1 = (1, 0)$ and $\boldsymbol{e}_2 = (0, 1)$ be the unit vectors.

We wish to characterise the probability that a marked customer entering a tandem queue when it is in a given state is lost en route. Instead of studying this directly, we will instead consider the probability that a marked customer reaches the destination – we call this the *success probability*. More precisely, let $p_d(\boldsymbol{n})$ (the success probability) be the probability that a marked customer reaches the destination when there are $n_1 + n_2 - 1$ customers ahead of it in the tandem queue. If $n_1 > 0$ then there are $n_1 - 1$ customers ahead of the marked customer in queue one. If $n_1 = 0$ then the marked customer is in queue two and there are $n_2 - 1$ customers ahead of it in queue two. Note that, once the marked customer has entered queue two, its successful arrival at the destination is assured; hence, we need only concern ourselves with the situation in which the marked customer is still in queue one. Since the queues are first-in–first-out, whether a marked customer reaches the destination depends only on the customers already in the system when it arrives, and is not affected by arrivals after the marked customer. The latter arrivals are thus not included in the notation for the success probability. This also means that the calculations below for the success probability hold for any arrival process that is independent of the service times. The success probabilities, $p_d(\boldsymbol{n})$, satisfy recursion equations similar to those for the hitting or reaching probabilities of a Markov chain (see, e.g. [15, p. 13]).

Lemma 2.1. *When* $n_1 \geq 1$,

$$p_d(\boldsymbol{n}) = \alpha_1 p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2 1_{\{n_2 < C_2\}})(1 - 1_{\{n_1 = 1,\, n_2 = C_2\}}) + \alpha_2 p_d(\boldsymbol{n} - \boldsymbol{e}_2 1_{\{n_2 > 0\}}), \qquad (2.1)$$

where

$$\alpha_1 = \frac{\mu_1}{\mu_1 + \mu_2}, \qquad \alpha_2 = \frac{\mu_2}{\mu_1 + \mu_2},$$

and $1_{\{\cdot\}}$ *is an indicator function. The initial conditions are* $p_d(\boldsymbol{n}) = 1,\ 0 < n_1 + n_2 \leq C_2$.

https://doi.org/10.1239/jap/1143936259 Published online by Cambridge University Press

Proof. These equations are found, similarly to those for the hitting/reaching probabilities of a Markov chain, by conditioning on the next service transition. If both queues contain customers, then the next service transition is a departure from queue i with probability $\mu_i/(\mu_1+\mu_2)$. Whether a departure from queue one is accepted at queue two depends on whether or not queue two is full. If queue two is empty then no departure from that queue can occur, and there is a null transition into the same state, $\boldsymbol{n}$. Particular attention needs to be given to queue one, which contains the marked customer. If $n_1 = 1$ then the marked customer is currently in service. If also $n_2 = C_2$, and the next transition is a departure from queue one, then the marked customer is lost and its success probability is 0. Since arrivals after the marked customer do not affect its success probability, λ does not appear in these equations. Observe also that, for $\boldsymbol{n}$ such that $0 < n_1 + n_2 \le C_2$, we trivially have $p_d(\boldsymbol{n}) = 1$, since in this case there is no possibility of the marked customer being lost in the transition from queue one to queue two.

The inclusion of a possible null transition in the recursion is a device useful in the proofs of the results below. We note that it of course makes no difference to the values calculated for the success probabilities.

In the results below we will use the following order relation.

Definition 2.1. We write $\boldsymbol{m} \prec \boldsymbol{n}$ if either

(a) $m_1 + m_2 < n_1 + n_2$ or

(b) $m_1 + m_2 = n_1 + n_2$ and $m_1 < n_1$.

We will also write $\boldsymbol{m} \preceq \boldsymbol{n}$ if either $\boldsymbol{m} \prec \boldsymbol{n}$ or $\boldsymbol{m} = \boldsymbol{n}$. The relation '$\prec$' uniquely determines an ordering on all possible nonnegative vectors $\boldsymbol{n} \in \mathbb{Z}_+^2$.

Before stating and proving our main theorem, we need a preliminary lemma.

Lemma 2.2. *When $n_1 \ge 1$, $0 \le n_2 < C_2$, and $n_1 + n_2 \ge C_2$, we have $p_d(\boldsymbol{n}) \ge p_d(\boldsymbol{n} + \boldsymbol{e}_2)$ with the inequality being strict if $C_2 > 1$.*

Proof. We begin by showing that $p_d(1, C_2 - 1) > p_d(1, C_2)$. This follows from the fact that $p_d(1, C_2 - 1) = 1$, while $p_d(1, C_2) = \alpha_2 p_d(1, C_2 - 1) = \alpha_2 < 1$. We now use a proof by induction on the order relation in Definition 2.1. We have shown that the hypothesis holds for the 'lowest' element in the set of states $\{\boldsymbol{n} : p_d(\boldsymbol{n} + \boldsymbol{e}_2) < 1\}$. Now fix $\boldsymbol{n}$ such that $n_1 \ge 1$ and $0 \le n_2 < C_2$, and assume the hypothesis to hold for all $\boldsymbol{m} \prec \boldsymbol{n}$ with $\boldsymbol{m}$ satisfying $m_1 \ge 1$ and $0 \le m_2 < C_2$. Then, by applying the recursion (2.1), we obtain

$$\begin{aligned} p_d(\boldsymbol{n}) &\ge p_d(\boldsymbol{n} + \boldsymbol{e}_2) \\ &\Leftrightarrow \alpha_1 p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2) + \alpha_2 p_d(\boldsymbol{n} - \boldsymbol{e}_2 1_{\{n_2 > 0\}}) \\ &\qquad \ge \alpha_1 p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_2 1_{\{n_2+1<C_2\}}) + \alpha_2 p_d(\boldsymbol{n}). \end{aligned}$$

Now, since $\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2 \prec \boldsymbol{n}$, we have $p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2) > p_d(\boldsymbol{n} - \boldsymbol{e}_1 + 2\boldsymbol{e}_2)$ when $n_2 + 1 < C_2$, by the induction hypothesis. Similarly, when $n_2 > 0$ we have $\boldsymbol{n} - \boldsymbol{e}_2 \prec \boldsymbol{n}$ and, again by the induction hypothesis, $p_d(\boldsymbol{n} - \boldsymbol{e}_2) > p_d(\boldsymbol{n})$. If $n_2 + 1 = C_2$ or $n_2 = 0$ then, respectively, the coefficients of α_1 or α_2 in the inequality are equal. If $C_2 > 1$ then at least one of the conditions $n_2 + 1 < C_2$ and $n_2 > 0$ must be satisfied; in this case the inequality is thus strict.

Corollary 2.1. *For $n_1 \ge 1$, $0 \le n_2 < C_2$, and $n_1 + n_2 \ge C_2$, with $C_2 \ge 2$ and $2 \le i \le C_2 - n_2$, $i \in \mathbb{Z}_+$, we have $p_d(\boldsymbol{n}) > p_d(\boldsymbol{n} + i\boldsymbol{e}_2)$.*

https://doi.org/10.1239/jap/1143936259 Published online by Cambridge University Press

Proof. The statement follows immediately by repeated application of Lemma 2.2. The inequality is now strict, since $C_2 \geq 2$.

We are now ready to state and prove the main result.

Theorem 2.1. *Consider a tandem queue consisting of two single-server queues with capacities C_1 and $C_2 < \infty$ and service rates μ_1 and μ_2. For $(2, C_2 - 1) \preceq \boldsymbol{n} \preceq (C_1, C_2)$ exactly one of the following relations holds:*

(a) $p_d(\boldsymbol{n}) > p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2 1_{\{n_2 < C_2\}})$ *if* $n_1 > 1$ *and* $0 < n_2 \leq C_2$.

(b) $p_d(\boldsymbol{n}) = p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2)$ *if* $n_1 \geq 1$ *and* $n_2 = 0$.

Proof. As in Lemma 2.2, we use an inductive proof on $\boldsymbol{n}$ under the ordering '$\prec$'. Observe that (b) follows immediately from (2.1), so that we need only consider (a). We begin by checking that $\boldsymbol{n} = (2, C_2 - 1)$ satisfies the relationships given in the theorem. If $C_2 = 1$ then $\boldsymbol{n}$ satisfies (b). If $C_2 > 1$ then we need to check that $\boldsymbol{n}$ satisfies (a). This holds because

$$\begin{aligned} p_d(2, C_2 - 1) > p_d(1, C_2) &\Leftrightarrow \alpha_1 p_d(1, C_2) + \alpha_2 p_d(2, C_2 - 2) > p_d(1, C_2) \\ &\Leftrightarrow p_d(2, C_2 - 2) > p_d(1, C_2), \end{aligned}$$

while $p_d(1, C_2) = \alpha_2 < 1 = p_d(2, C_2 - 2)$. Now consider $\boldsymbol{n} = (2, C_2)$. We need to check that this satisfies (a). We have

$$\begin{aligned} p_d(2, C_2) > p_d(1, C_2) &\Leftrightarrow \alpha_1 p_d(1, C_2) + \alpha_2 p_d(2, C_2 - 1) > p_d(1, C_2) \\ &\Leftrightarrow p_d(2, C_2 - 1) > p_d(1, C_2), \end{aligned}$$

since $\alpha_1 + \alpha_2 = 1$, and we have already shown (immediately above) that this holds. Thus, the hypothesis holds for the lowest $\boldsymbol{n}$ satisfying the conditions in (a) and for all $\boldsymbol{n}$ satisfying the conditions in (b). Therefore, the result holds for any $\boldsymbol{n} \succeq (2, C_2 - 1)$ such that $n_1 = 2$. In particular, the theorem holds for all cases with $C_1 = 2$.

Now consider any $\boldsymbol{n}$ such that $\boldsymbol{n} \succ (2, C_2)$ with $n_1 > 2$, $n_2 > 0$, and $C_1 > 2$. Assume that the induction hypothesis holds for all $\boldsymbol{m}$ such that $(2, C_2 - 1) \prec \boldsymbol{m} \prec \boldsymbol{n}$ (any such $\boldsymbol{m}$ satisfies one and only one of the relationships given in (a) and (b) – whichever one of them holds). If $n_2 < C_2$ then (a) holds for $\boldsymbol{n}$ if $p_d(\boldsymbol{n}) > p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2)$, which holds if and only if

$$\alpha_1 p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2) + \alpha_2 p_d(\boldsymbol{n} - \boldsymbol{e}_2) > \alpha_1 p_d(\boldsymbol{n} - 2\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_2 1_{\{n_2+1<C\}}) + \alpha_2 p_d(\boldsymbol{n} - \boldsymbol{e}_1).$$

Since $\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2 \prec \boldsymbol{n}$, by the induction hypothesis we have $p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2) > p_d(\boldsymbol{n} - 2\boldsymbol{e}_1 + 2\boldsymbol{e}_2)$ if $n_2 + 1 < C_2$ and $p_d(\boldsymbol{n} - \boldsymbol{e}_1 + \boldsymbol{e}_2) > p_d(\boldsymbol{n} - 2\boldsymbol{e}_1 + \boldsymbol{e}_2)$ if $n_2 + 1 = C_2$. Similarly, since $\boldsymbol{n} - \boldsymbol{e}_2 \prec \boldsymbol{n}$, by the induction hypothesis we also have $p_d(\boldsymbol{n} - \boldsymbol{e}_2) > p_d(\boldsymbol{n} - \boldsymbol{e}_1)$. If $n_2 = C_2$ then (a) holds for $\boldsymbol{n}$ if

$$\begin{aligned} &p_d(n_1, C_2) > p_d(n_1 - 1, C_2) \\ &\quad \Leftrightarrow \alpha_1 p_d(n_1 - 1, C_2) + \alpha_2 p_d(n_1, C_2 - 1) > p_d(n_1 - 1, C_2) \\ &\quad \Leftrightarrow p_d(n_1, C_2 - 1) > p_d(n_1 - 1, C_2), \end{aligned}$$

since $\alpha_1 + \alpha_2 = 1$. However, the final inequality follows from (a), by the induction hypothesis, since $(n_1, C_2 - 1) \prec (n_1, C_2)$.

Thus, the hypothesis holds for $\boldsymbol{n}$ and, so, by induction, the theorem holds.

https://doi.org/10.1239/jap/1143936259 Published online by Cambridge University Press

Corollary 2.2. *Consider a tandem queue consisting of two single-server queues with capacities* C_1 *and* $C_2 < \infty$ *and service rates* μ_1 *and* μ_2. *The following inequalities hold for* $\boldsymbol{n} \succeq (2, C_2 - 1)$, $i \in \mathbb{Z}_+$:

(a) $p_d(\boldsymbol{n}) > p_d(\boldsymbol{n} - i\boldsymbol{e}_1 + i\boldsymbol{e}_2)$ *if* $n_1 > 1$, $0 < n_2 < C_2$, *and* $1 < i \leq \min(n_1 - 1, C_2 - n_2)$,

(b) $p_d(\boldsymbol{n}) > p_d(\boldsymbol{n} - i\boldsymbol{e}_1)$ *if* $n_1 > 1$, $n_2 = C_2$, *and* $1 < i < n_1$.

Proof. The statements follow by repeated application of the inequalities of Theorem 2.1.

A natural heuristic when choosing a route, given two routes with the same total number of customers in each, might be to choose that route on which the customers ahead of you are as close to the destination as possible. However, this theorem shows that a customer may do better by choosing a route on which other customers ahead of them are closer to the source and further from the destination. The reason for this is clear once the phenomenon has been observed: the presence of other customers in a queue delays the expected completion of service of the marked customer, and its subsequent loss, and so has a protective effect for the marked customer. Hordijk and Koole [9] gave an example showing that it may be best to delay the *choice* of routes in a parallel system when they are both equally busy.

It is tempting to conjecture that a more general result holds for longer series of queues. Let n_i, $1 \leq i \leq K$, now be the number of customers in the ith queue, where queue one is the first queue after leaving the source, and queue K is the last queue before reaching the destination. Let $\boldsymbol{n} = (n_1, n_2, \ldots, n_K)$ and let the operators

$$T_{ij}\boldsymbol{n} = (n_1, n_2, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_K),$$
$$T_{i\cdot}\boldsymbol{n} = (n_1, n_2, \ldots, n_i - 1, \ldots, n_K),$$
$$T_{\cdot j}\boldsymbol{n} = (n_1, n_2, \ldots, n_j + 1, \ldots, n_K)$$

respectively denote a customer moving from queue i to queue j, moving from queue i out of the system, and moving into queue j from outside the system (see [11, p. 40, p. 48]). Then we might conjecture that $p_d(\boldsymbol{n}) > p_d(T_{1j}\boldsymbol{n})$ or even $p_d(\boldsymbol{n}) > p_d(T_{ij}\boldsymbol{n})$ might hold. Somewhat surprisingly, there are many choices of i, j, C, and $\boldsymbol{n}$ for which these inequalities *do* hold, but counterexamples show that neither holds in general. For instance, with $C_i = 2$ and $\mu_i = 1$, $i = 1, 2, 3$, we have

$$p_d(2, 1, 0) = 0.667 < 0.833 = p_d(1, 1, 1),$$
$$p_d(2, 1, 0) = 0.667 < 0.917 = p_d(2, 0, 1).$$

3. Examples and conclusion

In this section we give two examples that illustrate the results derived above.

Example 3.1. $C_1 = C_2 = 10$ and $\mu_1 = \mu_2$.

The success probabilities, calculated using the recursion formula of Lemma 2.1, are plotted in Figure 2. They satisfy the inequalities of Theorem 2.1. Note, for instance, that $p_d(1, 10) = 0.5$, while $p_d(10, 1) = 0.99$; the success probability is almost doubled, while $n_1 + n_2$ stays constant. Note also that $p_d(1, 10)$ is considerably less than $p_d(\boldsymbol{n})$ for any other choice of $\boldsymbol{n}$; thus, a route in any other state will be better for the customer, including those on which there are more customers in total.

https://doi.org/10.1239/jap/1143936259 Published online by Cambridge University Press

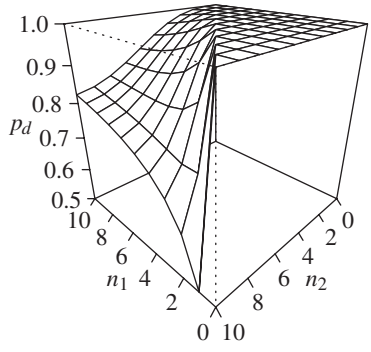

FIGURE 2: The success probabilities, $p_d(\boldsymbol{n})$, for a marked customer when $C_1 = C_2 = 10$ and $\mu_1 = \mu_2$.

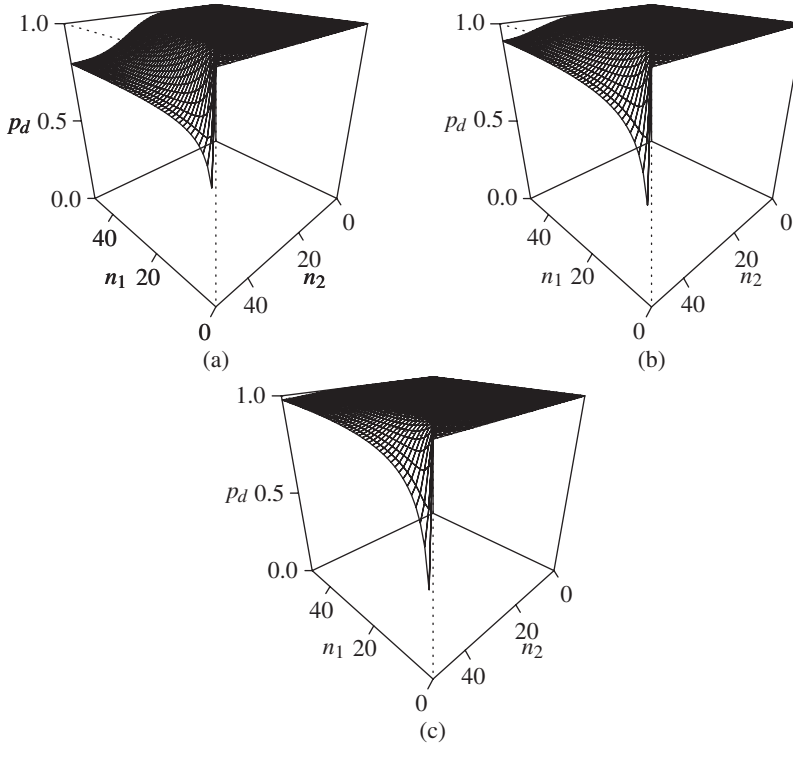

FIGURE 3: The success probabilities for a marked customer when $C_1 = C_2 = 50$, $\mu_1 = 1.0$, and (a) $\mu_2 = 0.75$, (b) $\mu_2 = 1.00$, (c) $\mu_2 = 1.25$.

Example 3.2. $C_1 = C_2 = 50$ and $\mu_1 = 1.0$, with μ_2 varying.

In this example, $\mu_1 = 1.0$ is fixed and μ_2 is allowed to vary. In Figure 3, the plots show the success probabilities for $\mu_2 = 0.75, 1.0, 1.25$. The most marked increase in the success probability occurs as n_1 increases, close to its minimum value.

https://doi.org/10.1239/jap/1143936259 Published online by Cambridge University Press

User-optimal routeing in parallel tandem queues can lead to apparently paradoxical behaviour, with users choosing routes that are busier, in the sense that either more customers are already present in the tandem queue or customers have greater remaining service requirements. This could have practical implications for routeing strategies in networks such as the Internet. In future work, we intend to examine the socially optimal policy and consider longer series of queues, queues with cross-traffic, and other service disciplines, such as processor sharing.

Acknowledgement

We thank the referee for a careful reading of the paper and helpful comments that increased the brevity and improved the clarity of the presentation.

References

[1] Avi-Itzhak, B. and Levy, H. (2001). Buffer requirements and server ordering in a tandem queue with correlated service times. *Math. Operat. Res.* **26,** 358–374.
[2] Bell, C. E. and Stidham, S. (1983). Individual versus social optimization in the allocation of customers to alternative servers. *Manag. Sci.* **29,** 831–839.
[3] Bertsekas, D. and Gallager, R. (1992). *Data Networks*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ.
[4] Calvert, B., Solomon, W. and Ziedins, I. (1997). Braess's paradox in a queueing network with state-dependent routing. *J. Appl. Prob.* **34,** 134–154.
[5] Cheng, D. W. and Yao, D. D. (1993). Tandem queues with general blocking: a unified model and comparison results. *J. Discrete Event Dyn. Systems Theory Appl.* **2,** 207–234.
[6] Cohen, J. E. and Kelly, F. P. (1990). A paradox of congestion in a queueing network. *J. Appl. Prob.* **27,** 730–734.
[7] Ephremides, A., Varaiya, P. and Walrand, J. (1980). A simple dynamic routing problem. *IEEE Trans. Automatic Control* **25,** 690–693.
[8] Hordijk, A. and Koole, G. (1990). On the optimality of the generalized shortest queue policy. *Prob. Eng. Inf. Sci.* **4,** 477–487.
[9] Hordijk, A. and Koole, G. (1992). On the shortest queue policy for the tandem parallel queue. *Prob. Eng. Inf. Sci.* **6,** 63–79.
[10] Hordijk, A. and Koole, G. (1992). On the assignment of customers to parallel queues. *Prob. Eng. Inf. Sci.* **6,** 495–511.
[11] Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. John Wiley, Chichester.
[12] Koole, G., Sparaggis, P. D. and Towsley, D. (1999). Minimizing response times and queue lengths in systems of parallel queues. *J. Appl. Prob.* **36,** 1185–1193.
[13] Kumar, P. R. and Walrand, J. (1985). Individually optimal routing in parallel servers. *J. Appl. Prob.* **22,** 989–995.
[14] Martin, J. B. (2002). Large tandem queueing networks with blocking. *Queueing Systems* **41,** 45–72.
[15] Norris, J. R. (1997). *Markov Chains*. Cambridge University Press.
[16] Towsley, D., Sparaggis, P. D. and Cassandras, C. G. (1992). Optimal routing and buffer allocation for a class of finite capacity queueing systems. *IEEE Trans. Automatic Control* **37,** 1446–1451.
[17] Walrand, J. (1988). *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, NJ.
[18] Weber, R. R. (1978). On the optimal assignment of customers to parallel servers. *J. Appl. Prob.* **15,** 406–413.
[19] Whitt, W. (1986). Deciding which queue to join: some counterexamples. *Operat. Res.* **34,** 55–62.
[20] Winston, W. (1977). Optimality of the shortest line discipline. *J. Appl. Prob.* **14,** 181–189.