No. 2008–54

BAD LUCK WHEN JOINING THE SHORTEST QUEUE

By Hans Blanc

June 2008

ISSN 0924-7815

BAD LUCK WHEN JOINING THE SHORTEST QUEUE

Hans (J.P.C.) Blanc
Tilburg University, Dept. Econometrics & Operations Research,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands.
E-mail: blanc@uvt.nl

June 9, 2008

Abstract

A frequent observation in service systems with queues in parallel is that customers in other queues tend to be served faster than those in one's own queue. This paper quantifies the probability that one's service would have started earlier if one had joined another queue than the queue that was actually chosen, for exponential multiserver systems with queues in parallel in which customers join one of the shortest queues upon arrival and in which jockeying is not possible.

Jel codes: C44, C60

Keywords: Queueing, Join-the-shortest-queue; Probability of bad luck; Power-series algorithm; Overtaking customers; Dedicated customers.

1 Introduction

Consider a service system with $c \geq 2$ parallel servers. Separate queues are formed in front of each server. Throughout, queues are defined as including the customer in service, if there is one. Each queue is served in a FIFO order. Customers arrive according to a Poisson process at rate λ. They join one of the shortest queues upon arrival and stay in the queue of their choice until they have been served. Then, they leave the system. This means that jockeying (see Zhao and Grassmann, 1990) is not considered. An example of a parallel service system in which jockeying is hardly possible is a toll booth at an autostrada (see Conolly, 1984). Services performed by server j have an exponentially distributed duration with a mean of $1/\mu_j$, $j = 1, \ldots, c$. Customers in such systems often notice that customers in other queues are being served faster than those in their own queue, and that they are overtaken by customers that arrived later. Of course, this phenomenon may be due to different skills, and hence different service rates, among the servers. If customers are aware of such differences, joining the shortest queue may not be the optimal decision. But even if the service rates of all servers are equal, this phenomenon frequently occurs. A simple explanation is found by considering the situation that a customer meets an equal number of customers $n \geq 1$ in each of the queues upon arrival. Then, by the lack of memory of the exponential service time distributions and the symmetry of the system, each queue has the same probability of becoming the queue that is soonest exempted of its n customers. Hence, the arriving customer has in this situation a probability of $(c-1)/c$ of bad luck, in the sense that he does not join the queue in which his service would have started earliest.
The aim of the present paper is to quantify the probability of bad luck for systems in which customers join one of the shortest queues upon arrival. For the computations reported in this paper we have used

the power-series algorithm to compute the stationary queue length distribution as described in Blanc (1987a, 1987b, 1992) for the shortest-queue system. The efficiency of the algorithm is further enhanced in Blanc (1993). Other approaches to shortest-queue systems can be found, among others, in Haight (1958), Flatto and McKean (1977), Halfin (1985), Rao and Posner (1987), Hanqin and Rongxin (1989), Adan, Wessels and Zijm (1990), Adan, Van Houtum and Van der Wal (1994) and Wu and Posner (1997). Winston (1977), Johri (1989) and Hordijk and Koole (1990) consider the optimality of the shortest queue discipline.
The organization of the rest of this paper is as follows. Section 2 considers the probability of bad luck for symmetric shortest-queue systems. Section 3 contains a discussion of this probability for asymmetric shortest-queue systems with different service rates among the servers. Section 4 is devoted to systems with both customers who join a shortest queue and customers who are dedicated to specific servers. A conclusion can be found in Section 5.

2 Symmetric systems

Consider a symmetric system in the sense that the service rates of all servers are equal, $\mu_j = \mu$, $j = 1, \ldots, c$, and that an arriving customer joins one of the shortest queues with equal probabilities. The load of this system is defined as $\rho \doteq \lambda/(c\mu)$, and for stability it is assumed that $\rho < 1$. Given that a customer joins a queue in which n customers were already present, the waiting time W_n of this new customer has an Erlang distribution with mean n/μ and consisting of n phases, $n = 1, 2, \ldots$, by the assumption of exponential service times. The conditional probabilities of bad luck given the state of the system upon arrival of a customer and the queue that is joined by this customer are defined as follows. Suppose the system is in state $(n_1, \ldots, n_c)$, with n_k the length of queue k, $k = 1, \ldots, c$, and the arriving customer joins queue j, then $\phi_j(n_1, \ldots, n_c)$ is the probability that some server i, $i \neq j$, will be the first to complete service of its current n_i customers. This probability can be determined from the relation

$$\phi_j(n_1, \ldots, n_c) \doteq \Pr\{\min_{i=1,\ldots,c} W_{n_i} < W_{n_j}\}, \quad j = 1, \ldots, c; \tag{2.1}$$

here, W_{n_i}, $i = 1, \ldots, c$, represent independent, Erlang distributed random variables with mean n_i/μ and consisting of n_i phases. To keep notation simple this probability will be evaluated for the case $j = 1$; the other cases follow by interchanging the indices. Clearly, if $n_1 = 0$ an arriving customer has zero waiting time, and, hence, for all $n_2, \ldots, n_c \in \mathbb{N}$,

$$\phi_1(0, n_2, \ldots, n_c) = 0. \tag{2.2}$$

Next, let $n_1 \geq 1$. By conditioning on the length y of the n_1 services in queue 1 this conditional probability becomes, for $n_2, \ldots, n_c \geq 1$,

$$\phi_1(n_1, \ldots, n_c) = 1 - \int_0^\infty \Pr\{W_{n_2} > y, \ldots, W_{n_c} > y\} \, d \, \Pr\{W_{n_1} \leq y\}. \tag{2.3}$$

By the independence of the services by the various servers this can be written as

$$\phi_1(n_1, \ldots, n_c) = 1 - \int_0^\infty \Pr\{W_{n_2} > y\} \cdots \Pr\{W_{n_c} > y\} \, d \, \Pr\{W_{n_1} \leq y\}. \tag{2.4}$$

Using the explicit expressions for the Erlang distribution and its density it follows that

$$\phi_1(n_1, \ldots, n_c) = 1 - \int_0^\infty \left[\prod_{j=2}^{c} \sum_{i_j=0}^{n_j-1} \frac{(\mu y)^{i_j}}{i_j!} \, e^{-\mu y} \right] \mu \, \frac{(\mu y)^{n_1-1}}{(n_1-1)!} \, e^{-\mu y} \, dy. \tag{2.5}$$

Table 1: Conditional probability of bad luck in queue 1 in the symmetric system with $c = 2$.

$n_2 \backslash n_1$	1	2	3	4	5	6
6	0.0156	0.0625	0.1445	0.2539	0.3770	0.5000
5	0.0313	0.1094	0.2266	0.3633	0.5000	0.6230
4	0.0625	0.1875	0.3438	0.5000	0.6367	0.7461
3	0.1250	0.3125	0.5000	0.6563	0.7734	0.8555
2	0.2500	0.5000	0.6875	0.8125	0.8906	0.9375
1	0.5000	0.7500	0.8750	0.9375	0.9688	0.9844

Table 2: Conditional probability of bad luck in queue 1 if $n_1 = 2$ in the symmetric system with $c = 3$.

$n_3 \backslash n_2$	2	3	4	5	6
6	0.5066	0.3271	0.2117	0.1431	0.1045
5	0.5158	0.3448	0.2379	0.1764	0.1431
4	0.5364	0.3813	0.2887	0.2379	0.2117
3	0.5802	0.4527	0.3813	0.3448	0.3271
2	0.6667	0.5802	0.5364	0.5158	0.5066

By interchanging the order of summation and integration this expression can be written as

$$\phi_1(n_1, \ldots, n_c) = 1 - \sum_{i_2=0}^{n_2-1} \cdots \sum_{i_c=0}^{n_c-1} \frac{1}{(n_1-1)! i_2! \cdots i_c!} \int_0^\infty \mu(\mu y)^{n_1+i_2+\cdots+i_c-1} \, e^{-c\mu y} \, dy. \tag{2.6}$$

This integral can be evaluated as, for $n_1, \ldots, n_c \geq 1$,

$$\phi_1(n_1, \ldots, n_c) = 1 - \sum_{i_2=0}^{n_2-1} \cdots \sum_{i_c=0}^{n_c-1} \frac{(n_1+i_2+\cdots+i_c-1)!}{(n_1-1)! i_2! \cdots i_c!} \frac{1}{c^{n_1+i_2+\cdots+i_c}}. \tag{2.7}$$

In the special case that all queues are equally short this probability becomes, for $n \geq 1$,

$$\phi_1(n, \ldots, n) = 1 - \sum_{i_2=0}^{n-1} \cdots \sum_{i_c=0}^{n-1} \frac{(n+i_2+\cdots+i_c-1)!}{(n-1)! i_2! \cdots i_c!} \frac{1}{c^{n+i_2+\cdots+i_c}} = 1 - \frac{1}{c} = \frac{c-1}{c}, \tag{2.8}$$

which is immediate for symmetrical systems, as noted in Section 1. Table 1 shows the conditional probability of bad luck $\phi_1(n_1, n_2)$ for customers joining queue 1 in the case $c = 2$, for $n_1, n_2 = 1, \ldots, 6$. Note that the values $\phi_1(n+m, n),\ n \geq 1,\ m \geq 1$, are irrelevant since an arriving customer will join the shorter queue, and, hence, not queue 1 in these states. Further, observe that $\phi_1(n, n+m) \to 0$ as $m \to \infty$ for fixed $n \geq 1$, but that $\phi_1(n, n+m)$ increases with increasing n for fixed $m \geq 1$. Moreover, using (2.7) it follows with the aid of Stirling's formula that for fixed $m \geq 1$, as $n \to \infty$,

$$\phi_1(n, n+m) = 1 - \sum_{i=0}^{n+m-1} \frac{(n+i-1)!}{(n-1)! i!} \frac{1}{2^{n+i}} = \frac{1}{2} - \sum_{k=0}^{m-1} \binom{2n+k-1}{n-1} \frac{1}{2^{2n+k}} \uparrow \frac{1}{2}. \tag{2.9}$$

Table 2 shows the conditional probability of bad luck $\phi_1(2, n_2, n_3)$ for customers joining queue 1 in the case $c = 3$, for $n_2, n_3 = 2, \ldots, 6$. Note that $\phi_1(2, 2, 2+m) = \phi_1(2, 2+m, 2) \to \frac{1}{2}$ as $m \to \infty$, which agrees with the value of $\phi_1(2, 2)$ for $c = 2$. More generally, as $m \to \infty$, $\phi_1(n, n+k, n+k+m) =$

$\phi_1(n, n+k+m, n+k)$ tends to the value of $\phi_1(n, n+k)$ for $c = 2$. For instance, for $n = 2$ and $k = 1$ the limit is $\phi_1(2, 3) = 0.3125$, see Tables 2 and 1. Hence, the limiting behavior of the conditional probabilities for $c = 3$ is more complex than that for $c = 2$. However, the most important property is that parallel to the main diagonal $n_1 = n_2 = n_3$ these probabilities tend to $\frac{2}{3}$, although rather slowly. For instance, $\phi_1(n, n, n+1) = \phi_1(n, n+1, n)$ equals 0.6527 for $n = 100$ and 0.6568 for $n = 200$, while $\phi_1(n, n+1, n+1)$ equals 0.6379 for $n = 100$ and 0.6464 for $n = 200$.
The (unconditional) probability of bad luck is defined as

$$P_{\mathrm{BL}} \doteq \sum_{n_1=1}^{\infty} \cdots \sum_{n_c=1}^{\infty} p(n_1, \ldots, n_c) \sum_{j=1}^{c} \gamma_j(n_1, \ldots, n_c) \phi_j(n_1, \ldots, n_c); \tag{2.10}$$

here, $\gamma_j(n_1, \ldots, n_c),\ j = 1, \ldots, c$, denotes the probability that a customer joins queue j when the system is in state $(n_1, \ldots, n_c)$. It is defined by, with $I_{\{.\}}$ the indicator function,

$$\gamma_j(n_1, \ldots, n_c) \doteq I_{\{\forall i\ n_i \geq n_j\}} \Big/ \sum_{i=1}^{c} I_{\{n_i = n_j\}}, \quad j = 1, \ldots, c,\ \ n_1, \ldots, n_c \in \mathbb{N}; \tag{2.11}$$

in particular, $\gamma_j(n_1, \ldots, n_c) = 0$ whenever $n_j > n_i$ for some $i \neq j,\ j = 1, \ldots, c$. For application of the power-series algorithm, the equilibrium state probabilities $p(n_1, \ldots, n_c)$ of the joint queue length process in (2.10) are represented as

$$p(n_1, \ldots, n_c) = \rho^{n_1 + \cdots + n_c} \sum_{k=0}^{\infty} \rho^k\, b(k; n_1, \ldots, n_c), \quad n_1, \ldots, n_c \in \mathbb{N}. \tag{2.12}$$

The coefficients $b(k; n_1, \ldots, n_c)$ can be recursively computed by a scheme (see Blanc 1987a, 1987b, 1992) that follows after substitution of (2.12) into the following global balance equations

$$\left[\lambda + \sum_{j=1}^{c} \mu_j\, I_{\{n_j \geq 1\}}\right] p(\mathbf{n}) = \lambda \sum_{j=1}^{c} \gamma_j(\mathbf{n} - \mathbf{e}_j) I_{\{n_j \geq 1\}}\, p(\mathbf{n} - \mathbf{e}_j) + \sum_{j=1}^{c} \mu_j\, p(\mathbf{n} + \mathbf{e}_j); \tag{2.13}$$

here, $\mathbf{n} \doteq (n_1, \ldots, n_c) \in \mathbb{N}^c$ denotes a state vector, and $\mathbf{e}_j$ are vectors of all zeros except a 1 at the jth coordinate, $j = 1, \ldots, c$.
Figure 1 shows the probability of bad luck in symmetric systems with $c = 2, 3, 4, 5$ servers, respectively, and a fixed service capacity of $c\mu = 1$, as a function of the load ρ. Recall that $\rho = \lambda < 1$ if $c\mu = 1$. It can be seen that at fixed, low values of ρ the probability of bad luck is decreasing with the number of servers. This can be explained by noting that in light traffic the probability that a customer finds an idle server upon arrival, and hence has zero probability of bad luck, increases with an increasing number of servers. In fact, it follows from the power-series expansions at $\rho = 0$ that in light traffic: for $c = 2, 3, \ldots,$

$$P_{\mathrm{BL}} \sim \frac{c^{c-2}\rho^c}{(c-2)!} - \frac{c^{c-2}\rho^{c+1}}{c!}\,(c^3 - c^2 - c + 2) + O(\rho^{c+2}), \quad \rho \downarrow 0. \tag{2.14}$$

On the other hand, the figure shows that at fixed values of ρ close to 1 the probability of bad luck is increasing with the number of servers. For these moderate numbers of servers the probability of bad luck seems to tend to $(c-1)/c$ as $\rho \to 1$. This is supported by (2.9) for the case $c = 2$. The general form of the heavy traffic asymptote can be written as: for $c = 2, 3, \ldots,$

$$P_{\mathrm{BL}} \sim \frac{c-1}{c} - A_c(1-\rho)^{q_c}, \quad \rho \uparrow 1. \tag{2.15}$$

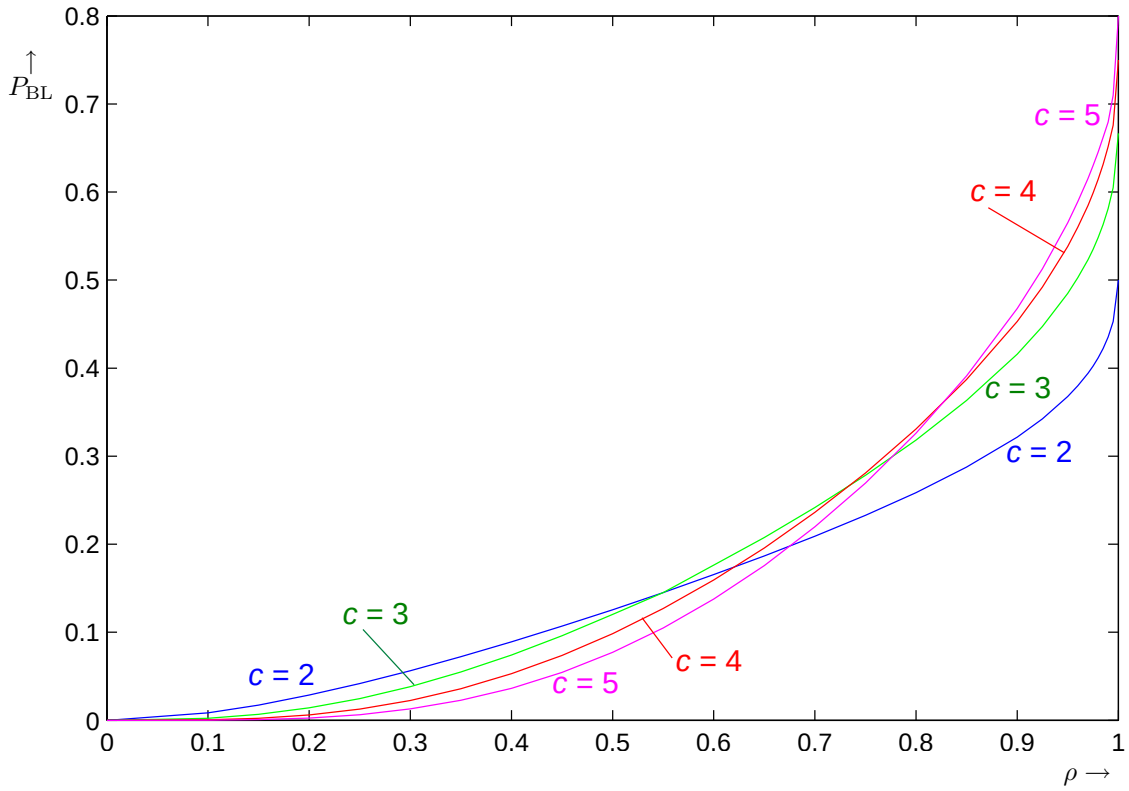

Figure 1: Probability of bad luck in symmetric systems, for $c = 2, 3, 4, 5$.

Based on values of this probability for ρ in the range 0.95–0.99, least square estimates of the constants are $A_2 \approx 0.51$, $q_2 \approx 0.45$, for $c = 2$. For $c = 3$, $A_3 \approx 0.75$ and $q_3 \approx 0.47$. And for $c = 4$, $A_4 \approx 0.88$ and $q_4 \approx 0.48$. This asymptotic behavior is illustrated by Figure 2 which displays the difference between $(c-1)/c$ and $P_{\rm BL}$ on a logarithmic scale for values of $1-\rho$ between 0.01 and 0.1, for $c = 2, 3, 4, 5$. Due to the rapidly changing behavior of $P_{\rm BL}$ for ρ close to 1 many coefficients of the power-series expansions in (2.12) are required to compute $P_{\rm BL}$ with sufficient accuracy in this area, much more than are needed for computing the mean and standard deviations of queue lengths and waiting times. We have used 80 or more terms for $c \leq 4$, and 58 terms for $c = 5$ because not only the power-series algorithm but also the evaluation of the conditional probabilities in (2.7) became very time consuming. As a consequence, the graphs for $c = 5$ in Figures 1 and 2 are less accurate for $\rho > 0.95$. Finally, we note that the probability of bad luck as defined in (2.10) is a rather crude performance measure for systems with three or more servers. Then, one can distinguish several degrees of bad luck. For instance, it is worse luck if all $c-1$ other servers work faster than the selected server than if only one of the other servers works faster. The expressions for such refined performance measures are more complicated than (2.7) but can be evaluated with the same techniques as discussed above.

3 Asymmetric systems

Next, consider an asymmetric system in which server j serves customers at rate μ_j, $j = 1, \ldots, c$. Customers are supposed to be not aware of these differences among the servers, and still join the shortest queue upon arrival. Hence, we will apply (2.11) unless stated otherwise. Expression (2.7) is generalized

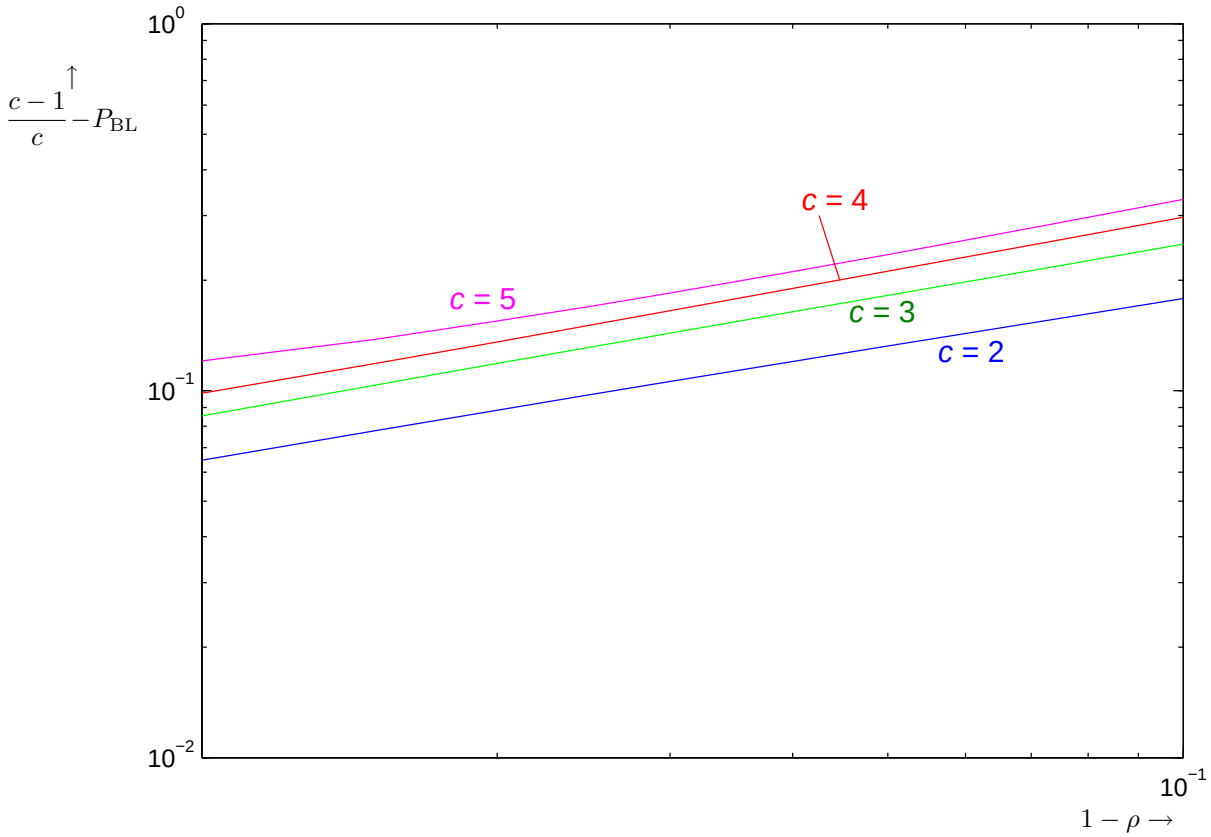

Figure 2: Heavy-traffic behavior of probability of bad luck in symmetric systems, for $c = 2, 3, 4, 5$.

for this case to, for $n_1, \ldots, n_c \geq 1$,

$$\phi_1(n_1, \ldots, n_c) = 1 - \sum_{i_2=0}^{n_2-1} \cdots \sum_{i_c=0}^{n_c-1} \frac{(n_1 + i_2 + \cdots + i_c - 1)!}{(n_1 - 1)! i_2! \cdots i_c!} \frac{\mu_1^{n_1} \mu_2^{i_2} \cdots \mu_c^{i_c}}{(\mu_1 + \cdots + \mu_c)^{n_1 + i_2 + \cdots + i_c}}. \tag{3.1}$$

Table 3 shows the conditional probability of bad luck $\phi_1(n_1, n_2)$ for customers joining queue 1 in the case $c = 2,\ \mu_1 = 1.2,\ \mu_2 = 0.8$ for $n_1, n_2 = 1, \ldots, 6$. The values $\phi_1(n + m, n)$, $n \geq 1,\ m \geq 1$, are again irrelevant as in Table 1, but they indicate that in some cases (when $\phi_1(n+m, n) \leq \frac{1}{2}$) arriving customers would be better off if they did not join the shorter queue. Further, note that $\phi_2(n_1, n_2) = 1 - \phi_1(n_1, n_2)$ for all $n_1, n_2 = 1, 2, \ldots$.

Table 4 shows the probability of bad luck for an arbitrary customer in systems with $c = 2$ servers for varying service rates. The load of the system is given by $\rho = \lambda/(\mu_1 + \mu_2)$. In the first three columns, $\gamma_1(n, n) = \gamma_2(n, n) = \frac{1}{2}$ for all $n \in \mathbb{N}$, according to (2.11). In the last column, the case $\gamma_1(n, n) = 1$, $\gamma_2(n, n) = 0$, $n \in \mathbb{N}$, that is, customers join queue 1 when they find both queues equally short upon arrival, is considered. It turns out that in lightly to moderately loaded systems, asymmetry in the service rates increases the probability of bad luck. This has more to do with an increase of congestion with increasing difference between the service rates than with the conditional probabilities of bad luck. For instance, $P_{\mathrm{BL}} \sim p(1,1)\,[\frac{1}{2}\phi_1(1,1) + \frac{1}{2}\phi_2(1,1)]$ $(\rho \downarrow 0)$, see (2.10), (2.12), and $p(1,1) \sim \frac{1}{2}\rho^2(\mu_1 + \mu_2)^2/(\mu_1\mu_2)$ $(\rho \downarrow 0)$ increases for fixed (small) load ρ as $\mu_1 = 2 - \mu_2$ increases, while $\frac{1}{2}\phi_1(1,1) + \frac{1}{2}\phi_2(1,1) = \frac{1}{2}$ remains constant. However, if customers join queue 1 when they find both queues equally short, both the congestion and the conditional probability of bad luck decrease with increasing difference between the service rates, since now $P_{\mathrm{BL}} \sim p(1,1)\phi_1(1,1)$ $(\rho \downarrow 0)$, and $p(1,1) \sim$

Table 3: Conditional probability of bad luck if queue 1 is joined, for $c = 2$, $\mu_1 = 1.2$, $\mu_2 = 0.8$.

$n_2 \backslash n_1$	1	2	3	4	5	6
6	0.0041	0.0188	0.0498	0.0994	0.1662	0.2465
5	0.0102	0.0410	0.0963	0.1737	0.2666	0.3669
4	0.0256	0.0870	0.1792	0.2898	0.4059	0.5174
3	0.0640	0.1792	0.3174	0.4557	0.5801	0.6846
2	0.1600	0.3520	0.5248	0.6630	0.7667	0.8414
1	0.4000	0.6400	0.7840	0.8704	0.9222	0.9533

Table 4: Probability of bad luck for arbitrary customer in asymmetric systems with $c = 2$.

ρ	$\mu_1 = \mu_2 = 1$	$\mu_1 = 1.2$, $\mu_2 = 0.8$	$\mu_1 = 1.5$, $\mu_2 = 0.5$	$\mu_1 = 1.5$, $\mu_2 = 0.5^\dagger$
0.25	0.0417	0.0427	0.0490	0.0262
0.50	0.1256	0.1265	0.1297	0.0950
0.75	0.2328	0.2285	0.2019	0.1779
0.90	0.3217	0.3061	0.2341	0.2238
0.95	0.3676	0.3421	0.2426	0.2374
0.99	0.4353	0.3843	0.2486	0.2475
$\uparrow 1$	0.5000	0.4000	0.2500	0.2500

† customers join queue 1 if queues are equally short upon arrival

$\rho^2(\mu_1+\mu_2)/\mu_1$ ($\rho \downarrow 0$), while $\phi_1(1,1) = \mu_2/(\mu_1+\mu_2)$. It further turns out that, on the contrary, in more heavily loaded systems, asymmetry in the service rates decreases the probability of bad luck. This can be explained by the features that if server 1 works faster ($\mu_1 > \mu_2$), the joint queue length process will tend to spend more time in the area $n_1 < n_2$ than in the area $n_1 > n_2$, while for $n_1 < n_2$, $\phi_1(n_1, n_2)$ is smaller than its opposite $\phi_2(n_2, n_1) = 1 - \phi_1(n_2, n_1)$, see Table 3. A further analysis indicates that $P_{\rm BL}$ approaches $\mu_2/(\mu_1 + \mu_2)$ as $\rho \uparrow 1$ if $\mu_1 > \mu_2$, while the approach of this limit is less steep with increasing value of $\mu_1 = 2 - \mu_2$, $1 \leq \mu_1 \leq 2$. This limit is obtained from numerical analysis. There is no simple generalization of (2.9) to the asymmetric case, since, e.g., $\phi_1(n, n) \downarrow 0$ as $n \to \infty$, see Table 3.

4 Systems with dedicated traffic

In this section we extend the foregoing analysis to shortest queue systems with dedicated customers to some or all of the servers. Let λ_0 denote the arrival rate of customers who have a simple service demand that can be dealt with by any server, and who join one of the shortest queues upon arrival. Further, let λ_j, $j = 1, \ldots, c$, denote the arrival rate of customers who have a specialized service demand that can only be dealt with by server j, and who join queue j whatever the state of the system upon their arrival. The conditional probabilities of bad luck given the state of the system upon arrival are the same as for systems without dedicated customers, see (3.1). But the dedicated customers do influence the equilibrium queue length probabilities $p(n_1, \ldots, n_c)$, cf. (2.10), which satisfy the global balance equations

$$\left[\lambda_0 + \sum_{j=1}^{c} \lambda_j + \sum_{j=1}^{c} \mu_j \, I_{\{n_j \geq 1\}} \right] p(\mathbf{n})$$

Table 5: Probability of bad luck in systems with dedicated customers.

ρ	$\lambda_1 = \lambda_2 = 0$	$\lambda_1 = \lambda_2 = \frac{1}{8}\lambda_0$	$\lambda_1 = \lambda_2 = \frac{3}{4}\lambda_0$	$\lambda_1 = \frac{2}{3}\lambda_0,\ \lambda_2 = 0$
0.25	0.0417	0.0387	0.0320	0.0353
0.50	0.1256	0.1179	0.0985	0.1042
0.75	0.2328	0.2190	0.1791	0.1802
0.90	0.3217	0.3041	0.2458	0.2318
0.95	0.3676	0.3503	0.2874	0.2621
0.99	0.4353	0.4232	0.3719	0.3359

$$= \sum_{j=1}^{c} [\lambda_j + \lambda_0\, \gamma_j(\mathbf{n} - \mathbf{e}_j)]\, I_{\{n_j \geq 1\}}\, p(\mathbf{n} - \mathbf{e}_j) + \sum_{j=1}^{c} \mu_j\, p(\mathbf{n} + \mathbf{e}_j); \tag{4.1}$$

here, we used the same notations as in (2.13). The above set of balance equations forms a straightforward extension of the set of equations (2.13) for systems without dedicated traffic. The probabilities $p(n_1, \ldots, n_c)$ again allow a power-series expansion of the form (2.12), and the corresponding coefficients $b(k; n_1, \ldots, n_c)$, which depend in this case on the service rates $\mu_j,\ j = 1, \ldots, c$, and the normalized arrival rates $\lambda_j/\rho,\ j = 0, \ldots, c$, can be recursively computed.

Table 5 shows the probability of bad luck for an arbitrary customer who joins the shorter queue in systems with $c = 2$ servers with equal service rates ($\mu_1 = \mu_2 = 1$). The load of the system is defined by $\rho = (\lambda_0 + \lambda_1 + \lambda_2)/(\mu_1 + \mu_2)$ and the system is stable for $\rho < 1$. In all cases, changing the load is performed by changing all arrival rates in fixed proportions. The first three columns concern symmetric systems, with equal shares of dedicated traffic for both servers. It turns out that the presence of dedicated customers decreases the probability of bad luck for customers who join the shorter queue. This is caused by the fact that the queue length process tends to move further away from the diagonal $n_1 = n_2$ due to the arrivals of dedicated customers, which is advantageous for customers joining the shorter queue, see Table 1. The last column concerns an asymmetric system, in which only server 1 receives dedicated traffic. Although the total proportion of dedicated traffic (40%) in this case is less than that in the case of column 3 (60%), it turns out that in heavy traffic the probability of bad luck is smaller due to the fact that the queue length process tends to move further away to one side of the diagonal ($n_1 > n_2$). Note that, as a consequence of (2.9), the probability of bad luck will tend to $\frac{1}{2}$ as $\rho \uparrow 1$ in all cases. The approach of this limit is even steeper the higher the fraction of dedicated traffic because the queue length process tends to reside further away from the diagonal $n_1 = n_2$ and $\phi_2(n + m, n) = \phi_1(n, n + m)$ approaches its limit $\frac{1}{2}$ as $n \to \infty$ slower the higher the value of m.

5 Conclusion

This paper has studied what we have called the probability of bad luck for shortest-queue systems. A customer is said to experience bad luck if he joined one of the shortest queues upon arrival, but his service would have started earlier if he had joined one of the other queues. In symmetric systems, the probability of bad luck may well exceed $\frac{1}{2}$ when there are three or more servers, but this only occurs if the load of the system is very close to 1. The approach of this probability to its heavy traffic limit is very steep, so that this limit, which is easily computable, will not be a good approximation for most values of the load. Asymmetry in the service rates tends to increase this probability in light traffic, but to decrease it in moderate to heavy traffic. Dedicated background traffic decreases this probability.

References

Adan, I.J.B.F., J. Wessels, W.H.M. Zijm, 1990. Analysis of the symmetric shortest queue problem, *Stochastic Models* 6: 691–713.

Adan, I.J.B.F., G.J. Van Houtum, J Van der Wal, 1994. Upper and lower bounds for the waiting times in the symmetric shortest queue, *Annals of Operations Research* 48: 197–217.

Blanc, J.P.C., 1987a. A note on waiting times in systems with queues in parallel, *Journal of Applied Probability* 24: 540–546.

Blanc, J.P.C., 1987b. On a numerical method for calculating state probabilities for queueing systems with more than one waiting line, *Journal of Computational and Applied Mathematics* 20: 119–125.

Blanc, J.P.C., 1992. The power-series algorithm applied to the shortest-queue model, *Operations Research* 40: 157–167.

Blanc, J.P.C., 1993. Performance analysis and optimization with the power-series algorithm. In: L. Donatiello, R. Nelson (eds.), *Performance Evaluation of Computer and Communication Systems*. Springer, Berlin, pp. 53–80.

Conolly, B.W., 1984. The autostrada queueing problem, *Journal of Applied Probability* 21: 394–403.

Flatto, L., H.P. McKean, 1977. Two queues in parallel, *Communications in Pure and Applied Mathematics* 30: 255–263.

Haight, F.A., 1958. Two queues in parallel, *Biometrika* 45: 401–410.

Halfin, S., 1985. The shortest queue problem, *Journal of Applied Probability* 22: 865–878.

Hanqin, Z., W. Rongxin, 1989. Heavy traffic limit theorems for a queueing system in which customers join the shortest line, *Advances in Applied Probability* 21: 451–469.

Hordijk, A., G. Koole, 1990. On the optimality of the generalized shortest queue policy, *Probability in the Engineering and Informational Sciences* 4: 477–487.

Johri, P.K., 1989. Optimality of the shortest line discipline with state-dependent service rates, *European Journal of Operational Research* 41: 157–161.

Rao, B.M., M.J.M. Posner, 1987. Algorithmic and approximation analyses of the shorter queue model, *Naval Research Logistics* 34: 381–398.

Winston, W. 1977. Optimality of the shortest line discipline, *Journal of Applied Probability* 14: 181–189.

Wu, P., M.J.M. Posner, 1997. A level-crossing approach to the solution of the shortest-queue problem, *Operations Research Letters* 21: 181–189.

Zhao, Y., W.K. Grassmann, 1990. A solution of the shortest queue model with jockeying - in terms of traffic intensity ρ, *Naval Research Logistics* 37: 773–787.