Available online at www.sciencedirect.com

ScienceDirect

European Journal of Operational Research 195 (2009) 167–173

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Stochastics and Statistics

Bad luck when joining the shortest queue

J.P.C. Blanc *

Tilburg University, Department Econometrics and Operations Research, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Received 24 July 2007; accepted 24 January 2008
Available online 5 February 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.

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

1. Introduction

Consider a service system with $c \geqslant 2$ 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 $n \geqslant 1$ 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,b, 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

* Tel.: +31 13 4663039; fax: +31 13 4663280.
E-mail address: blanc@uvt.nl

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2008.01.037

(1989), Adan et al. (1990, 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

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

where $W_{n_i},\ i = 1, \ldots, c$ 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 $n_1 \geqslant 1$. By conditioning on the length y of the n_1 services in queue 1 this conditional probability becomes, for $n_2, \ldots, n_c \geqslant 1$,

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

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

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

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

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

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

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

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

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

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

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

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 $\phi_1(n+m, n)$, $n \geqslant 1,\ m \geqslant 1$, are irrelevant since an arriving customer will join the shorter queue, and, hence, not queue 1 in these states.