A note on waiting times in systems with queues in parallel

Blanc, J.P.C.

Published in:
Journal of Applied Probability

Publication date:
1987

J. Appl. Prob. **24**, 540–546 (1987)
Printed in Israel

A NOTE ON WAITING TIMES IN SYSTEMS WITH QUEUES IN PARALLEL

J. P. C. BLANC,* *Delft University of Technology*

Abstract

Numerical data are presented concerning the mean and the standard deviation of the waiting-time distribution for multiserver systems with queues in parallel, in which customers choose one of the shortest queues upon arrival. Moreover, a new numerical method is outlined for calculating state probabilities and moments of queue-length distributions. This method is based on power series expansions and recursion. It is applicable to many systems with more than one waiting line.

POWER SERIES EXPANSIONS

1. Introduction

Conolly [1] discusses a queueing model with a Poisson arrival stream, two identical exponential servers, each with a queue, and customers choosing the shorter queue at the instant of their arrival (or joining either with equal probability in case of equal queue lengths). He proposes studying the system with a finite waiting room, and calculates the state probabilities by means of inversion of tridiagonal matrices. Then he compares the distribution of the sum of the lengths of these two parallel queues with the queue-length distribution for the $M/M/2$ system, and, based on the evidence given, he suggests that there is hardly any difference in the performance of the two queueing systems. However, it is difficult to draw conclusions on the basis of a few probabilities for the total number of customers in the systems. Roughly speaking, there are two features which make multiserver systems with parallel queues inferior to multiserver systems with a single queue. The first one is inefficiency, i.e. the possibility that a server is idle while customers are waiting in another queue. The second one is unfairness to the customers in not necessarily serving them

Received 9 December 1985; revision received 20 March 1986.

* Present address: University of Limburg, Faculteit der Algemene Wetenschappen, P.O.Box 616, 6200 MD Maastricht, The Netherlands.

in order of arrival (although in each queue the service discipline is FCFS). The distribution of the total number of customers in the system does not completely reflect the first feature, and does not reflect the second one at all. The mean waiting time, which is related to the mean of the minimum of the queue lengths, reflects the first feature more strongly, while the standard deviation of the waiting times also reflects the second feature (throughout this paper waiting times do not include service times). Therefore, we propose to compare the means and the standard deviations of the waiting-time distributions for multiserver systems with queues in parallel and related systems with a common queue for all servers. In Section 3 the mean waiting times (Table 1) and the standard deviations of the waiting times (Table 2) will be presented for the two types of systems, not only in the case of two servers with equal service rates, but also in the cases of two servers with different service rates and of three servers with equal service rates (the waiting rooms are assumed to be unbounded in all cases). For this purpose a new numerical method for calculating state probabilities and moments of the joint queue-length distributions for systems with more than one waiting line will be discussed in Section 2. The calculations show that the absolute differences in mean and standard deviation of the waiting times for the two types of systems increase, and the relative differences decrease, with increasing traffic intensity. The tables indicate further that these differences increase when the service rates become unbalanced and when the number of servers increases (at fixed traffic intensity ρ). For example for $\rho = 0.7$, the increase in waiting for systems with queues in parallel with respect to related systems with a common queue is 15% (mean) and 18% (standard deviation) in the case of two servers and equal service rates; these percentages are 19% and 24% in the case of two servers and a service rate ratio of 3:2, and 30% and 36% in the case of three servers and equal service rates.

2. New method for calculating state probabilities and moments

This section is devoted to a discussion of a new numerical method for calculating state probabilities and moments of joint queue-length distributions for multiserver systems with queues in parallel, based on power series expansions. Let λ be the arrival rate of customers, s the number of servers, $r_j\mu$ the service rate of server $j, j = 1, \cdots, s$, $\sum_{j=1}^{s} r_j = 1$, and let ρ be the traffic intensity, $\rho = \lambda/\mu$. The waiting rooms are unbounded. The system is stable when $\rho < 1$. In the stationary situation, let N_j be the number of customers in front of server $j, j = 1, \cdots, s$, and let C be the queue which an arriving customer chooses. Because Poisson arrivals see time averages, the assumption that an arriving customer chooses with equal probability one of the shortest

queues implies that for $j = 1, \cdots, s$, $m = 1, \cdots, s$,

$$\Pr\{C = j \mid N_j \leqq N_i,\ i = 1, \cdots, s \wedge m = \#(i; N_i = N_j)\} = 1/m,$$

$$\Pr\{C = j \mid \exists i N_i < N_j\} = 0. \tag{1}$$

Let $\bar{n} = (n_1, \cdots, n_s)$ be a vector with non-negative entries, and let $\bar{e}_j = (e_{j1}, \cdots, e_{js})$ be vectors with $e_{ji} = 1$ if $i = j$ and $e_{ji} = 0$ otherwise, $i = 1, \cdots, s, j = 1, \cdots, s$. The equations for the state probabilities

$$p(\bar{n}) := \Pr\{N_j = n_j, j = 1, \cdots, s\}$$

read with the above notation and with $I(E)$ the indicator function of the event E:

$$\left[\rho + \sum_{j=1}^{s} r_j I(n_j > 0)\right] p(\bar{n})$$

$$= \sum_{j=1}^{s} r_j p(\bar{n} + \bar{e}_j) \tag{2}$$

$$+ \rho \sum_{j=1}^{s} I(n_j > 0) p(\bar{n} - \bar{e}_j) \Pr\{C = j \mid N_i = n_i - e_{ji},\ i = 1, \cdots, s\}.$$

These equations cannot be solved recursively in this form, but Keane, Hooghiemstra and Van de Ree (personal communication) propose the following method for recursive solution. They note that for many exponential queueing systems with one or more waiting lines the limit as $\rho \downarrow 0$ of

$$\rho^{-\|\bar{n}\|} p(\bar{n}), \qquad \|\bar{n}\| := \sum_{j=1}^{s} n_j, \tag{3}$$

exists, and that the state probabilities $p(\bar{n})$ are regular functions of the traffic intensity ρ in a neighborhood of the origin. Because the power series expansions of $p(\bar{n})$ as functions of ρ are not convergent on the whole interval $0 < \rho < 1$ for the present model, they use the following conformal mapping of the interval (0, 1) onto itself:

$$\theta = \frac{1 + G}{1 + G\rho}\,\rho, \qquad \rho = \frac{\theta}{1 + G - G\theta}, \qquad G \geqq 0, \tag{4}$$

and introduce power series expansions of $p(\bar{n})$ as functions of θ, based on (3),

$$p(\bar{n}) = \theta^{\|\bar{n}\|} \sum_{k=0}^{\infty} \theta^k b(k, \bar{n}). \tag{5}$$

Substitution of (4) and (5) into (2) leads to the following equations: for $k = 0, 1, \cdots,$

$$
\begin{aligned}
&(1+G)\Big[\sum_{j=1}^{s} r_j I(n_j>0)\Big] b(k,\bar{n}) \\
&\quad=\Big[-1+G\sum_{j=1}^{s} r_j I(n_j>0)\Big] I(k>0)b(k-1,\bar{n}) \\
&\quad\quad+(1+G)\sum_{j=1}^{s} r_j I(k>0)b(k-1,\bar{n}+\bar{e}_j) \\
&\quad\quad-G\sum_{j=1}^{s} r_j I(k>1)b(k-2,\bar{n}+\bar{e}_j) \\
&\quad\quad+\sum_{j=1}^{s} I(n_j>0)b(k,\bar{n}-\bar{e}_j)\Pr\{C=j \mid N_i=n_i-e_{ji},\ i=1,\cdots,s\}.
\end{aligned} \tag{6}
$$

The requirement that the state probabilities sum to 1 implies

$$
\begin{aligned}
&b(0,\bar{0})=1; \quad b(k,\bar{0})=-\sum_{n_1}\cdots\sum_{n_s} b(k-\|\bar{n}\|,\bar{n}), \\
&\qquad 0<\|\bar{n}\|\leqq k, \quad k=1,2,\cdots.
\end{aligned} \tag{7}
$$

To obtain the coefficients of the power series expansions of the state probabilities up to the Mth power of θ determine $b(k,\bar{0})$ from (7) and then calculate $b(k,\bar{n})$ recursively from (6) for increasing values of $\|\bar{n}\|$ up to $\|\bar{n}\| = M-k$, successively for $k=0,1,\cdots,M$. For all cases considered a conformal mapping with $G=1$ in (4) provided good results, although in individual cases other values of G provided slightly better results (faster convergence of the power series). Several theoretical questions about convergence of the procedure have not been answered yet, but in practice the method works well. The advantage of this method over others (see [1]) is that all coefficients $b(k,\bar{n})$ can be obtained recursively and that it requires little effort to calculate the state probabilities for different values of the traffic intensity ρ once these coefficients are available.

The moments of the queue lengths $N_1,\cdots,N_s$ can be calculated from the state probabilities, but they require a large number of coefficients to be calculated for ρ close to 1. However, it is also possible to derive the coefficients of the power series expansions of the moments of $N_1,\cdots,N_s$ from those of the state probabilities $p(\bar{n})$. As for many queueing systems the coefficients of the first moments converge to a constant (a first-order pole at $\rho=1/\theta=1$), while those of the second moments converge to a linear asymptote (a second-order pole at $\rho=1/\theta=1$). Therefore, we propose to use, for $i,j=1,\cdots,s$,

$$
\sum_{k=1}^{M} e_j(k)\theta^k + e_j(M)\frac{\theta^{M+1}}{1-\theta}, \tag{8}
$$

$$(9) \qquad \sum_{k=1}^{M} c_{ij}(k)\theta^k + \left[c_{ij}(M) + \frac{c_{ij}(M) - c_{ij}(M-1)}{1-\theta}\right] \frac{\theta^{M+1}}{1-\theta},$$

as successive approximations for $E\{N_j\}$ and $E\{N_iN_j\}$ respectively; here $e_j(k)$ are the calculated first M power series coefficients of $E\{N_j\}$, and $c_{ij}(k)$ those of $E\{N_iN_j\}$: for $k = 1, \cdots, M$, $0 \leqq \| \bar{n} \| \leqq k$, $i, j = 1, \cdots, s$,

$$e_j(k) = \sum_{n_1} \cdots \sum_{n_s} n_j b(k - \| \bar{n} \|, \bar{n}), \quad c_{ij}(k) = \sum_{n_1} \cdots \sum_{n_s} n_i n_j b(k - \| \bar{n} \|, \bar{n}).$$

Addition of the second terms in (8) and (9) strongly improves the rate of convergence, especially for $\theta(\rho)$ close to 1. Without these terms the approximations with $M = 60$, $G = 1$ are 8.560 (8.557) for the mean waiting time and 9.205 (9.727) for the standard deviation of the waiting times in a symmetric system with two (three) queues in parallel. In Table 3 the approximations including the second terms in (8) and (9) are given for the same cases and several values of M. The rate of convergence of the numerical procedure is the smallest of all quantities considered in Tables 1 and 2 for the standard deviation of the waiting times in the case of three similar queues in parallel and $\rho = 0.9$. For $G = 1$ the relative error for this quantity is about 9×10^{-3} when $M = 24$, and about 2×10^{-4} when $M = 36$.

Remark 1. Straightforward application of the above method requires $\binom{M+s+1}{s+1}$ coefficients $b(k, \bar{n})$ to be successively calculated according to (6) and (7) in order that the power series approximations of the state probabilities and the moments of the joint queue-length distribution can be evaluated up to the Mth power of θ. The number of calculations can be reduced by using symmetry properties if present and by exploring the fact that some coefficients $b(k, \bar{n})$ vanish for the particular model where customers choose the shortest of s queues upon arrival (for this model a result stronger than (3) holds; it is related to Theorem 1 in [1]).

Remark 2. The method can also be applied in the case of a waiting room of restricted size. If the total number of customers in the system is not allowed to exceed L, then (6) and (7) hold for $\| \bar{n} \| \leqq L$ when the coefficients $b(k, \bar{n})$ are set equal to 0 for $\| \bar{n} \| > L$. In this case the system is stable for all $\rho, \rho > 0$. Therefore, addition of the second terms in (8) and (9) does not make sense. The method may be impractical (too slow convergence) for large values of ρ ($\rho \gg 1$). Note that the coefficients $b(k, \bar{n})$ are the same for the cases of finite (L) and infinite waiting rooms when $k + \| \bar{n} \| \leqq L$.

3. Waiting-time distribution

In this section the means and the standard deviations of the waiting times (excluding service) will be compared between multiserver systems with queues

TABLE 1
Comparison of mean waiting times ($\mu = 1$) for multiserver systems with queues in parallel and for those with a common queue

	$s = 2, r_1 = r_2 = \frac{1}{2}$		$s = 2, r_1 = 0.6, r_2 = 0.4$		$s = 3, r_1 = r_2 = r_3 = \frac{1}{3}$	
ρ	parallel	common	parallel	common	parallel	common
0.1	0.03543	0.02020	0.03814	0.02089	0.01049	0.004115
0.3	0.2881	0.1978	0.3059	0.2022	0.1929	0.1000
0.5	0.8526	0.6667	0.8950	0.6757	0.7345	0.4737
0.7	2.216	1.922	2.297	1.935	2.127	1.641
0.9	8.950	8.526	9.096	8.544	8.947	8.171

TABLE 2
Comparison of standard deviations of waiting times ($\mu = 1$) for multiserver systems with queues in parallel and for those with a common queue

	$s = 2, r_1 = r_2 = \frac{1}{2}$		$s = 2, r_1 = 0.6, r_2 = 0.4$		$s = 3, r_1 = r_2 = r_3 = \frac{1}{3}$	
ρ	parallel	common	parallel	common	parallel	common
0.1	0.3768	0.2109	0.4067	0.2144	0.2508	0.09554
0.3	1.090	0.7253	1.168	0.7326	1.075	0.5252
0.5	1.965	1.491	2.091	1.499	2.133	1.292
0.7	3.553	3.020	3.755	3.026	3.896	2.872
0.9	10.43	9.891	10.91	9.893	10.91	9.831

TABLE 3
Convergence of the successive approximations of the mean and the standard deviation of the waiting-time distribution by using (12) and increasing values of M in (8) and (9), for $\rho = 0.9$ and $G = 1$ ($\theta = 0.9474$), $\mu = 1$

	Two similar queues		Three similar queues	
M	mean	st. dev.	mean	st. dev.
12	9.0103092	10.258091	8.4745820	2.482531
24	8.9508779	10.422550	8.9353209	10.810719
36	8.9498431	10.427095	8.9470076	10.907775
48	8.9498259	10.427212	8.9472638	10.910459
60	8.9498256	10.427215	8.9472658	10.910489

in parallel and with a common queue. The waiting time ($W_{p/s}$) distribution for the system with s queues in parallel is related to the joint queue-length distribution of this system in the following way:

$$\text{(10)} \qquad \Pr\{W_{p/s} < t\} = \sum_{n=0}^{\infty} \sum_{j=0}^{s} [1 - \exp(-r_j \mu t)]^n * \Pr\{C = j, N_j = n\}, \quad t > 0.$$

By means of induction it can be shown that (2) implies the following relations:

$$\text{(11)} \qquad r_j \Pr\{N_j = n+1\} = \rho \Pr\{C = j, N_j = n\}, \quad j = 1, \cdots, s, \quad n = 0, 1, \cdots.$$

Note that Kingman [2], Formula (37), relates the waiting time erroneously to the maximum of the queue lengths.

From (10) and (11) we obtain the relations which have been used to derive the mean and the standard deviation of the waiting-time distribution from the moments of the queue-length distribution (Tables 1, 2, 3):

$$\text{(12)} \qquad \begin{aligned} E\{W_{p/s}\} &= \frac{1}{\rho\mu} \sum_{j=1}^{s} [E\{N_j\} - \Pr\{N_j > 0\}], \\ E\{W_{p/s}^2\} &= \frac{1}{\rho\mu^2} \sum_{j=1}^{s} \frac{1}{r_j} [E\{N_j^2\} - E\{N_j\}]. \end{aligned}$$

For the $M/M/s$ system we have, with the same notation as for the system with queues in parallel,

$$\text{(13)} \qquad E\{W_{M/M/s}\} = \frac{w_s}{\mu(1-\rho)}, \qquad E\{W_{M/M/s}^2\} = \frac{2w_s}{\mu^2(1-\rho)^2};$$

here w_s denotes the probability that a customer has to wait in an $M/M/s$ system:

$$\text{(14)} \qquad w_2 = \frac{\rho^2}{2r_1 r_2(1-\rho) + \rho}; \qquad w_3 = \frac{\frac{9}{2}\rho^3}{1 + 2\rho + \frac{3}{2}\rho^2}, \quad r_1 = r_2 = r_3 = \tfrac{1}{3}.$$

Finally, we remark that calculations for ρ up to 0.98 indicate that $E\{W_{p/s}\} - E\{W_{M/M/s}\}$ increases to a finite constant, while $\text{var}\{W_{p/s}\} - \text{var}\{W_{M/M/s}\}$ increases at least as $O((1-\rho)^{-1})$ as ρ increases to 1, in all cases examined.

References

[1] CONOLLY, B. W. (1984) The autostrada queueing problem. *J. Appl. Prob.* **21**, 394–403.

[2] KINGMAN, J. F. C. (1961) Two similar queues in parallel. *Ann. Math. Statist.* **32**, 1314–1323.