Performance Evaluation 17 (1993) 123–139
North-Holland

An approximation for the mean response time for shortest queue routing with general interarrival and service times

Randolph D. Nelson
IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Thomas K. Philips
IBM Retirement Fund, 262 Harbor Drive, Stanford, CT 06904, USA

Received 21 August 1990
Revised 15 October 1991

Abstract

Nelson, R.D. and T.K. Philips, An approximation for the mean response time for shortest queue routing with general interarrival and service times, Performance Evaluation 17 (1993) 123–139.

In this paper we derive an approximation for the mean respone time of a multiple queue system in which shortest queue routing is used. We assume there are K identical queues with infinite capacity. Interarrival and service times are generally distributed, and an arriving job is routed to a queue of minimal length. Our approximation is a simple closed form equation that requires only the mean and coefficient of variation of job's interarrival and service times. The approximation is extensively compared to simulated values for values of $K \leqslant 8$, and has small relative errors, typically less than 5%, for systems where the coefficient of variation of interarrival and serivce times are both $\leqslant 1$. For the system consisting of Poisson arrivals and exponential service times, we extend the approximation so that the error is less than one half of one percent for $K \leqslant 8$.

Keywords: shortest queue routing; load balancing; queueing theory.

1. Introduction

In this paper we consider a queueing system consisting of K, for $K \geqslant 1$, identical, infinite capacity queues, that have independent and identically distributed servers with generally distributed service times with mean $1/\mu$ and coefficient of variation C_a. Jobs arrive at the system with an average rate of λ, with interarrival times that are generally distributed with coefficient of variation C_s. Jobs are scheduled according to a shortest queue discipline. In case of a tie, we randomly select one queue from those of minimal length. Customers leave the system after being served once. We define the utilization of the servers to be $\rho = \lambda/K\mu$ and assume that $\rho < 1$, which assures that the system is ergodic. We develop a closed form approximation for the mean response time for a system of this type and compare it with simulation. Our results show that for systems where $C_a \leqslant 1$ and $C_s \leqslant 1$, relative errors of less than 5% are typically found over all values of ρ. When the coefficient of variations of either the interarrival time or the service time are greater than 1, larger relative errors are obtained and can be as large as 20%. Extensive validation of the approximation is provided in the paper and in the case of Poisson arrivals with exponential service times, an improved approximation is presented that yields an error of less than one half of one percent for $K \leqslant 8$.

Correspondence to: Dr. R. Nelson, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY 10598, USA.

Shortest queue routing is a natural way to balance the load of a system across several processors and has been used as a load balancing mechanism as well as a scheduling mechanism in an effort to minimize job response time. Besides the intuitive appeal of the policy, the attractiveness of shortest queue routing is supported by work that shows the policy is optimal under a variety of metrics when the servers are exponential (see the literature survey below). The difficulty in analyzing the performance of such systems lies in the fact that the queues of such a system are not independent and that the arrival process of jobs to each queue depends upon the state of the entire system. One state space description of a system with K queues is given by a K dimensional vector consisting of the number of jobs in each of the queues. Because the system is symmetric, state reductions can be obtained which can lead to results for $K = 2$ by numerical techniques that rely on state truncation or by solving a Reimann Hilbert boundary value problem. Because the state space is intrinsically multidimensional these techniques do not generalize to higher values of K. As a result of these inherent mathematical difficulties, it is not surprising that most results are derived for the $K = 2$ case with exponential interarrival and service times. A brief literature review of some of this work now follows.

Winston [21] showed shortest queue routing to be optimal in that it maximized the discounted number of jobs to complete their service within any specified interval of time. Ephremides et al. [4] showed it to be optimal in that it minimized the expected total time for the completion of service of all jobs that arrive by some fixed time T. For the case of $K = 2$, Kingman [10] and Flatto and McKean [5] studied the problem via transform methods and found expressions for the mean number in the system and the occupancy probabilities. The mean number of jobs in the system is expressed as an infinite sum, and simplifies only under a heavy traffic assumption. Cohen and Boxma [2] reduce this problem to a Reimann Hilbert boundary value problem and obtain a functional representation for the expected number of jobs in the system. Approaches leading to heavy traffic approximations for the the two queue case can be found in Foschini and Salz [6] and Knessl et al. [12]. The latter is significant in that it allows generally distributed service times. Rao and Posner [17] develop a computational algorithm for a two queue system with asymmetric

Randolph Nelson received his Ph.D. degree in computer science in 1982 from the University of California, Los Angeles, CA. Since then he has been employed as a Research Staff Member at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. He currently manages a modeling methodology group at the research center and is generally interested in fields associated with mathematical modeling of computer systems.

Thomas K. Philips received his B.S. in electrical engineering from the Benares Hindu University, Benares, India in 1980, and his M.S. and Ph.D. degrees in electrical and computer engineering from the University of Massachusetts at Amherst in 1983 and 1986, respectively. He then joined the IBM T.J. Watson Research Center in Yorktown Heights, NY, where he worked on queueing theory, graph theory and the design and analysis of algorithms. He is currently a Program Manager in the Investment Research department at the IBM Retirement Fund in Stamford, CT.

servers. Another computational algorithm that is valid for an arbitrary number of servers has been developed by Blanc [1]. These approaches require substantial computational effort, especially at high utilizations, and in the case of Blanc's algorithm, for large numbers of queues. Halfin [9] examines the two queue problem with equal rate servers and uses a linear programming technique to compute bounds on the expectation and the distribution of the number of jobs in the system. The above work does not, in most cases, generalize to systems with more than two queues and when it does is typically limited by practical considerations to very small values of K (typically $K = 3$).

In a previous paper [15] we examined the shortest queue problem for Poisson arrivals and exponential service times. The form of this approximation is an algebraic equation that uses certain observations about the structure of shortest queue routing. We were able to obtain a closed form expression for this equation that suggested a possible extension to general arrival and service distributions. In this work we present this extension. The proposed approximation is easily calculated, depends only upon the first two moments of interarrival and service times and is accurate when the coefficients of variation of the interarrival and service times are less than 1. To keep the paper self contained, we outline an extended version of the basic approximation found in [15] for the special case of Poisson arrivals and exponential service times. This improved approximation has an error in comparison to simulated values that is less than one half of one percent for $K \leqslant 8$.

An outline of this paper follows. In Section 2 we first derive an approximation for the case with exponential servers and Poisson arrivals. This approximation is based on two key insights and is compared to simulation in this section. In section 3 we generalize the approximation to general interarrival and service time distributions and also compare its accuracy to simulation. Section 4 outlines the extension of the basic approximation for Poisson arrivals and exponential service times that has improved accuracy. The improved approximation is a scaled version of the basic approximation and is also in closed form. In Section 5, we discuss two applications of these results. Our conclusions and suggestions for further research are contained in Section 6.

2. Derivation of the approximation

In this section we develop our approximation for exponential servers with Poisson arrivals. The approximation is based on two key insights regarding properties of the shortest queue scheduling algorithm. The first insight is that the total queue length probability for a system of this type does not differ greatly from that of a $M/M/K$ queue. The second insight is that the shortest queue scheduling algorithm results in a small variation in queue lengths between the different queues of the system. In fact, Gubner et al. [8] have shown that the expectation of the difference of the queue lengths for any two queues is uniformly bounded for all $\rho > 0$. Using these insights we propose a form for the approximation and derive an expression for the mean response time. We now motivate the approximation, present notation and state some preliminary results.

2.1. Preliminary results

We first discuss the queue length process for the shortest queue system. The state of the system, denoted by $\boldsymbol{n} = (n_1, n_2, \ldots, n_K)$ is the number of jobs in each of the K queues. We define $\beta(\boldsymbol{n})$ to be the stationary probabilities and $\overline{W}$ and $\overline{T}$ to be the mean job waiting and response time, respectively. Let $S(\boldsymbol{n})$ be the length of the shortest queue for state $\boldsymbol{n}$. Using this we can write

$$\overline{W} = \frac{1}{\mu} \sum_{\boldsymbol{n}} \beta(\boldsymbol{n}) S(\boldsymbol{n}) \tag{1}$$

and $\overline{T} = \overline{W} + 1/\mu$. The difficulty in evaluating (1) lies in the evaluation of $\beta(\boldsymbol{n})$. For $K > 2$, such an evaluation appears to be mathematically intractable. An approximation, however, can be based on the observation that the shortest queue discipline attempts to distribute work evenly among the queues. Consequently, the function $S(\boldsymbol{n})$ depends strongly upon the total number of jobs in the system, and only weakly on the individual components of $\boldsymbol{n}$. Equivalently, certain configurations of jobs in queues are substantially more probable than others. If we define the total number of jobs to be $|\boldsymbol{n}| = \sum_{k=1}^{K} n_k$, then this implies that $S(\boldsymbol{n})$ can be approximated by a function S_n where $n = |\boldsymbol{n}|$. S_n is the expected length of the shortest queue for a given total number of

jobs in the system and since this quantity is dependent on the utilization of the system we will write S_n as $S_n(\rho)$. Let N_{sq} be a random variable that denotes the number of jobs in the system and let $\beta_n(\rho)$, for $n > 0$, be its distribution, i.e. $\beta_n(\rho)$ is the stationary probability that there are n jobs in the shortest queue system having a utilization of ρ. Our approximation is of the form

$$\overline{W} \simeq \frac{1}{\mu} \sum_{n=0}^{\infty} \beta_n(\rho) S_n(\rho). \tag{2}$$

We next establish approximations for the above two functions. We first concentrate on $\beta_n(\rho)$.

Consider the probability of finding a given total number of jobs in the K queues. Let $N_{m/m/k}$ be a random variable denoting the number of jobs in an $M/M/K$ system and let its distribution be given by $\pi_n(\rho)$, $n \geqslant 0$. A $M/M/K$ queue may be thought of as K single server queues which use shortest queue routing *and* allow jobs to jockey between queues. Jockeying, in this case, is done so as to equalize the queue lengths, and to ensure that jobs are served in first come first served order. This ensures that a server cannot remain idle when there are jobs in the system. It is easy to see that a shortest queue discipline which allows this type of jockeying minimizes the number of jobs in the system. This implies that $N_{sq} \geqslant_{st} N_{M/M/K}$ and consequently that the mean response time for the $M/M/K$ queue provides a lower bound to the mean response time for the shortest queue system.

These are circumstances under which the stationary probabilities for the total number of jobs in the shortest queue system and a $M/M/K$ system converge. For example, for sufficiently small utilization the probability of finding more than one job in either of the systems becomes negligible. Under these conditions, jockeying occurs infrequently and $\lim_{\rho \to 0} (\beta_n(\rho) - \pi_n(\rho)) \to 0$, $n \geqslant 0$. This property has also been observed by Blanc [1] and Conolly [3]. On the other hand, for large utilizations the servers are hardly ever idle, and jockeying does not significantly alter the number of jobs in the system. Consequently, the steady state probabilities must once again be similar. Foschini and Salz [6] among others have shown that for $K = 2$, $\lim_{\rho \to 1} (\beta_n(\rho) - \pi_n(\rho)) \to 0$, $n \geqslant 0$. In our approximation we assume that the two probabilities are identical for all utilizations and write

$$\beta_n(\rho) \simeq \pi_n(\rho), \quad n = 0, 1, \ldots, \quad 0 \leqslant \rho < 1. \tag{3}$$

Previous results allow us to write [13]

$$\pi_n = \begin{cases} \dfrac{(K\rho)^n}{n!A(\rho)} & n = 0, 1, \ldots, K-1, \\ \dfrac{K^K \rho^n}{K!A(\rho)} & n \geqslant K, \end{cases} \tag{4}$$

where $A(\rho) = \sum_{n=0}^{K-1} (K\rho)^n/n! + (K\rho)^K/K!(1-\rho)$. For $n \geqslant K$, π_n can also be written as

$$\pi_n = P_K(\rho)\rho^{n-K}(1-\rho), \tag{5}$$

where $P_K(\rho)$ is the Erlang delay function [11], and is the probability that an arriving job has to wait for a server. $P_K(\rho)$ is given by

$$P_K(\rho) = (K\rho)^K/K!(1-\rho)A(\rho). \tag{6}$$

Expressing the occupancy probabilities in this manner allows us to write the mean response time for the shortest queue system in a compact form. We record here the average waiting time for the $M/M/K$ queue, denoted by $\overline{W}_{M/M/K}(\rho)$, and given by

$$\overline{W}_{M/M/K}(\rho) = \frac{1}{\mu} \left\{ \frac{\rho(K\rho)^{K-1}}{K!A(\rho)(1-\rho)^2} \right\} = \frac{1}{\mu} \frac{P_K(\rho)}{K(1-\rho)}. \tag{7}$$

It is important to note that the assumption that the stationary probabilities of the shortest queue and of the $M/M/K$ system are equal is used only within the context of the approximation given in eq. (2). Since, as noted before, $N_{sq} \geqslant_{st} N_{M/M/K}$, the function $S_n(\rho)$, which must also be determined in our approximation, must correct the underestimation of response time that results from our assumption that N_{sq} and $N_{M/M/K}$ are equal in distribution. In Section 2.2 we determine an equation for $S_n(\rho)$ which yields the basic approximation.

2.2. Basic approximation

Consider the arrangement of the jobs in a shortest queue system. Since the shortest queue policy attempts to equalize the queue lengths, we

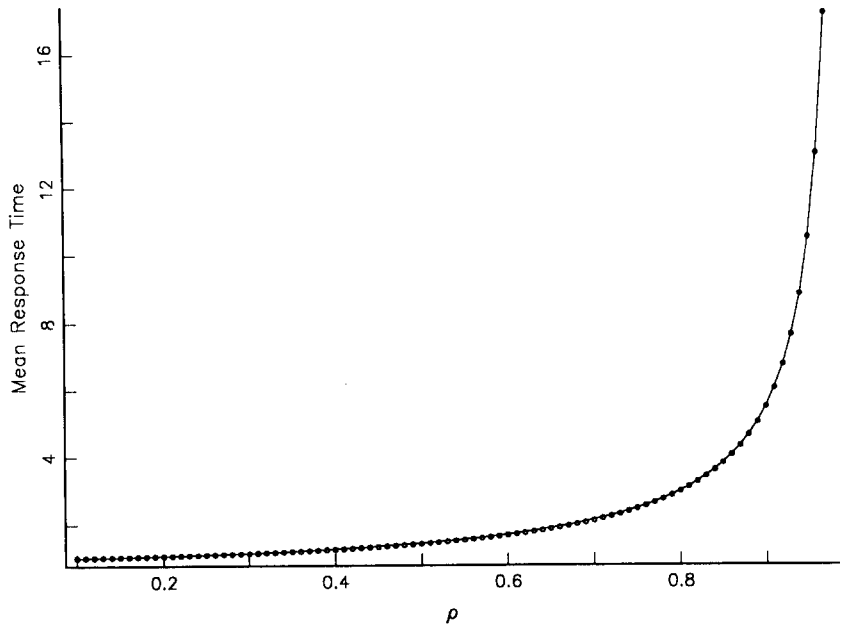

Fig. 1. Comparison of mean response time for $K = 2$. Basic approximation.

assume that the queue lengths differ by at most one job. This assumption is supported by the fact that as $\rho \to 1$, the absolute value of the difference between the queue lengths converges to a birth death process with mean 1 [9]. This implies that we can approximate $S_n(\rho)$ by $\lfloor n/K \rfloor$. Using this and eq. (3) in (2) yields

$$\overline{W}_K(\rho) \simeq \frac{1}{\mu}\left\{\sum_{i=1}^{\infty} i \sum_{j=0}^{K-1} \pi_{iK+j}(\rho)\right\}. \tag{8}$$

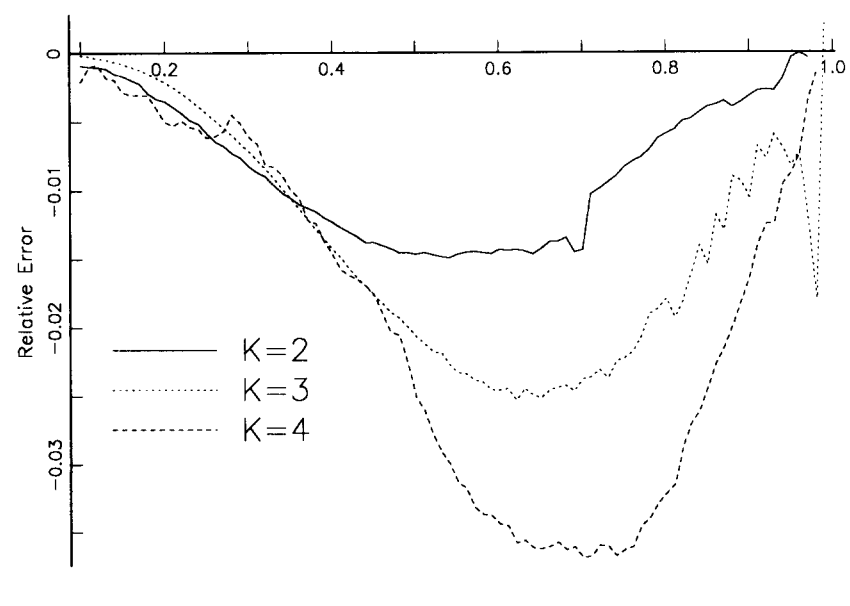

Fig. 2. Relative error of mean response time. Basic approximation.

Substituting (5) into (8) and simplifying yields

$$\overline{W}_K(\rho) \simeq \frac{1}{\mu}\frac{P_K(\rho)}{1-\rho^K}. \tag{9}$$

Equation (9) is our basic approximation. We record here expressions for the cases where $K \leqslant 4$.

$$\overline{W}_2(\rho) \simeq \frac{1}{\mu}\left\{\frac{2\rho^2}{(1+\rho)(1-\rho^2)}\right\}, \tag{10}$$

$$\overline{W}_3(\rho) \simeq \frac{1}{\mu}\left\{\frac{9\rho^3}{(1-\rho^3)(2+4\rho+3\rho^2)}\right\} \tag{11}$$

and

$$\overline{W}_4(\rho) \simeq \frac{1}{\mu}\left\{\frac{32\rho^4}{(1-\rho^4)(3+9\rho+12\rho^2+8\rho^3)}\right\}. \tag{12}$$

To verify this approximation we wrote a simulation for the shortest queue policy. In Fig. 1 we plot the approximation against simulation points for $K=2$. The points in this figure, as in all figures in this paper, were obtained from a regenerative simulation in which the number of regenerative cycles was determined so that the 99% confidence intervals for the mean response time were less than 1% of the sample mean. Confidence intervals are not shown in the figures as they are smaller than the points used to indicate the expected values. As an example of the computational requirement for this precision, for $\rho = 0.97$, 2×10^7 cycles were necessary. As can be seen from this figure, the approximation has a maximum relative error of 7% for $K=8$. In Fig. 2 we plot the relative error for $K \leqslant 8$. As can be seen from the figure, the maximum relative error, though small, increases with K. In the next section, we generalize the approximation to the case of general interarrival and service times.

3. General interarrival and service times

In this section we examine the response time when the interarrival and service time distributions are generally distributed. Our approach is identical to that outlined in Sections 2.1 and 2.2 in that we will assume that the statistics for the number of jobs in the system is the same as in a $G/G/K$ queueing system and that jobs are uniformly distributed among the queues. For the $G/G/K$ queue there is no exact analysis for the number of jobs in the system. We thus must approximate this statistic in our approximation. Our approach is to assume that the number of jobs in a $G/G/K$ system is geometrically distributed starting at K, i.e.,

$$\pi_n \simeq P_K(\hat{\rho})\hat{\rho}^{n-K}(1-\hat{\rho}), \quad n \geqslant K, \tag{13}$$

where $\hat{\rho}$ is a parameter whose choice is next discussed. We note that there is no need to approximate the occupancy probabilities for $n < K$ for in that case the length of the shortest queue is 0. Also note that this expression has the same form as in the exponential case given by eq. (5). Shore [20] proposed a $G/G/K$ approximation in which he assumed that the queue length was geometrically distributed past K, and choose $\hat{\rho}$ so that the mean number of jobs in the waiting room equalled the mean number in the waiting room of an $G/G/1$ queue with a mean service time of $1/K\mu$ and an identical coefficient of variation. If the arrival process is Poisson, the mean number in the $M/G/1$ system is given by the Pollaczek–Khinchin formula, and equating the two means gives

$$\hat{\rho} = \frac{\rho(1+C_s^2)}{2+\rho(C_s^2-1)}. \tag{14}$$

Since $C_s^2 \geqslant 0$, it is clear that $\hat{\rho}$ is bounded between 0 and 1. Observe that if the coefficient of variation of the service time is 1, as is the case when the servers are exponential, $\hat{\rho}=\rho$, and our approximation returns the $M/M/K$ occupancy probabilities. When the arrival process is not Poisson, no solution for the mean number of jobs is known in general. A number of approximations are, however, available [14,18,19]; we use one due to Marchal [14] that is also cited in Gross and Harris [7]. The expression is simple, and extensive testing has shown it to be quite accurate when the coefficients of variation of the interarrival and service times do not exceed 1. Marchal's approximation for the mean queue length, $\overline{N}$, of a $G/G/1$ queue is

$$\overline{N} \simeq \rho + \frac{\rho^2(C_a^2+C_s^2)}{2(1-\rho)}, \tag{15}$$

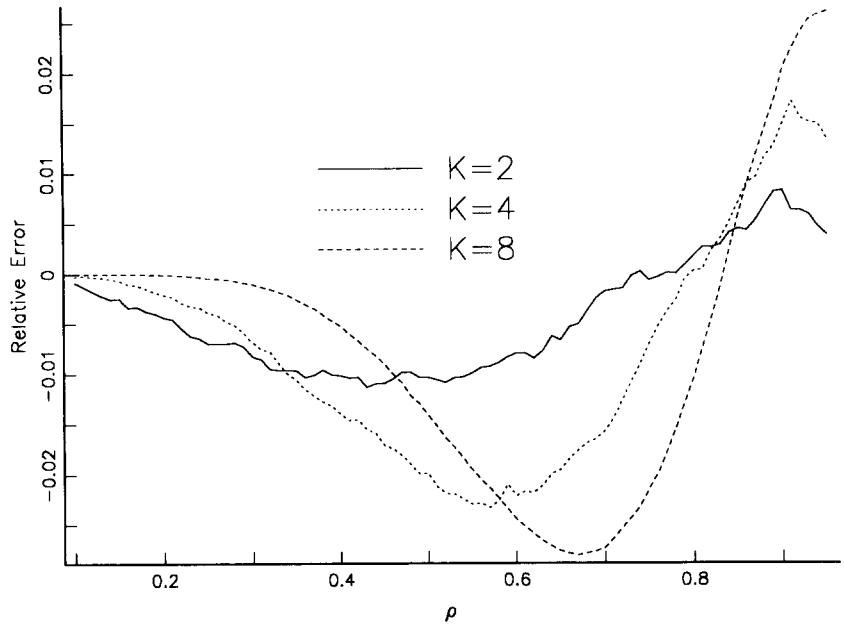

Fig. 3. Relative error: Poisson arrivals, deterministic servers. Basic approximation.

implying that

$$\hat{\rho} = \frac{\rho(C_a^2 + C_s^2)}{2 + \rho(C_a^2 + C_s^2 - 2)}. \tag{16}$$

We have used this approximation throughout this study, even when the coefficients of variation lie outside the suggested ranges. We believe that if a more accurate expression for the mean number of jobs in a $G/G/1$ queue can be found, (especially for very high coefficients of variation), a more accurate approximation would result.

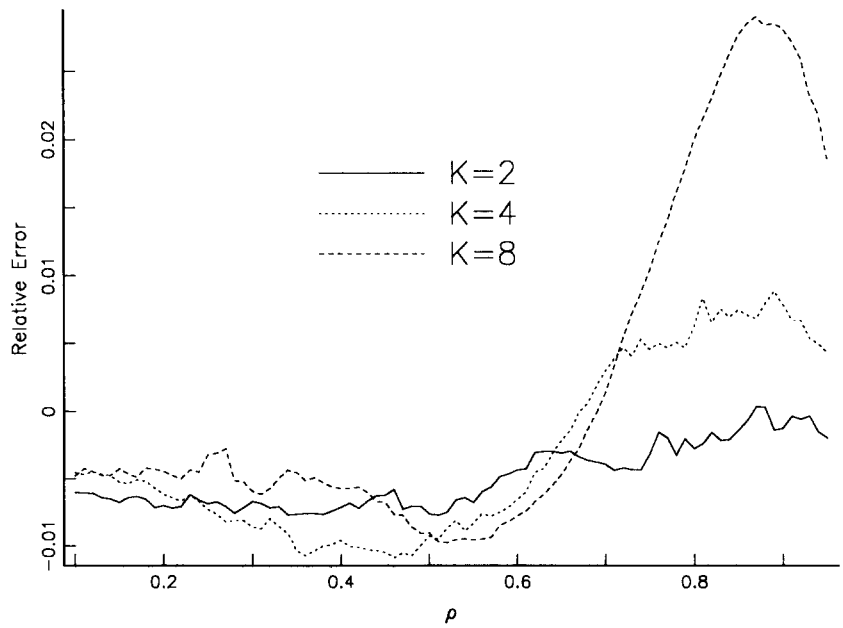

Fig. 4. Relative error: Poisson arrivals, uniform servers. Basic approximation.

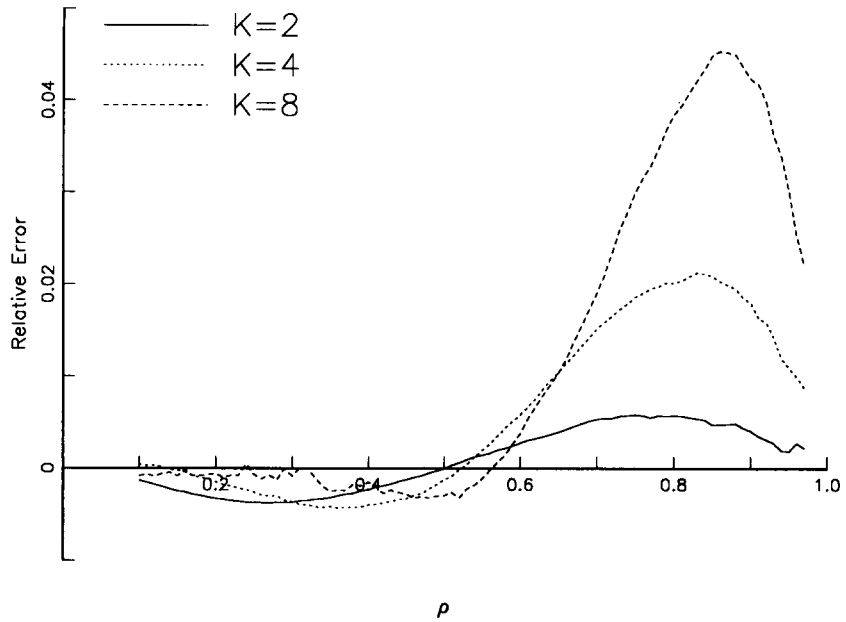

Fig. 5. Relative error: Poisson arrivals, Erlang 2 servers. Basic approximation.

The mean waiting time is obtained from arguments identical to those leading to eq. (9) and yields

$$\overline{W}_K(\rho) \simeq \frac{1}{\mu}\frac{P_K(\hat{\rho})}{1-\hat{\rho}^K}. \tag{17}$$

The equation has exactly the same form as (9), differing from it only in the use of $\hat{\rho}$ instead of ρ. We have evaluated our approximation for the following interarrival and service time distributions:

- Exponential interarrival times, deterministic service times (Fig. 3)

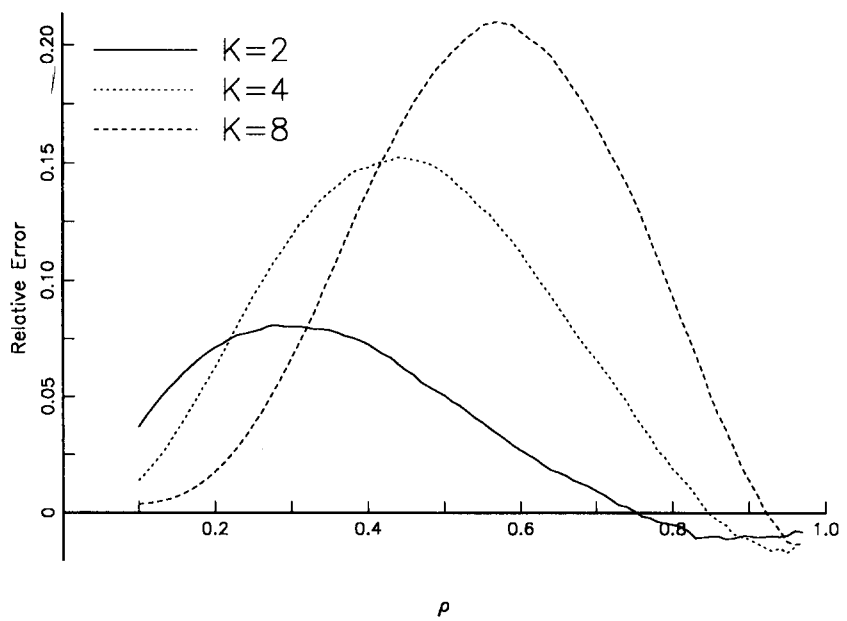

Fig. 6. Relative error: Poisson arrivals, hyperexponential servers. Basic approximation.

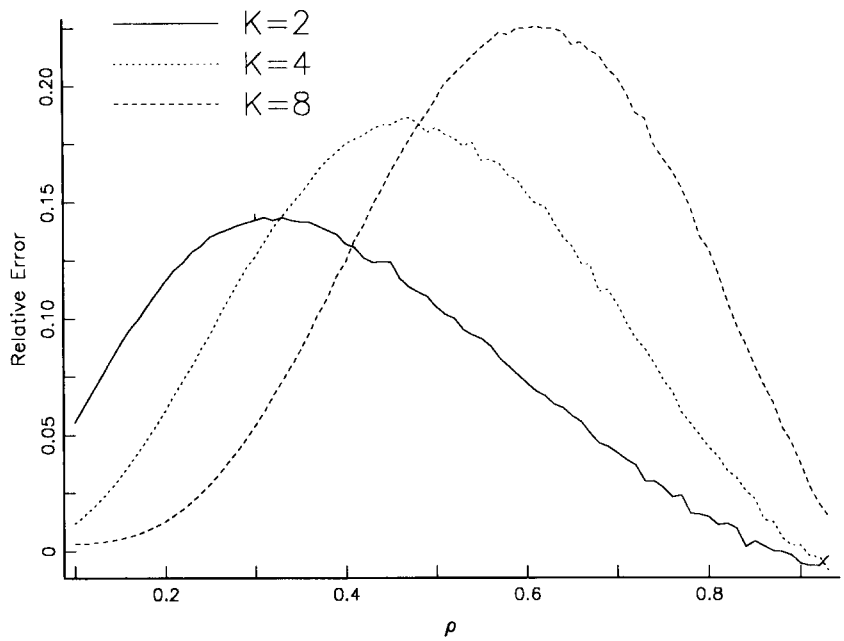

Fig. 7. Relative error: Erlang arrivals, hyperexponential servers. Basic approximation.

- Exponential interarrival times, uniform service times (Fig. 4)
- Exponential interarrival times, E_2 service times (Fig. 5)
- Exponential interarrival times, H_2 service times (Fig. 6)
- E_2 interarrival times, H_2 service times (Fig. 7)
- E_2 interarrival times, E_2 service times (Fig. 8)

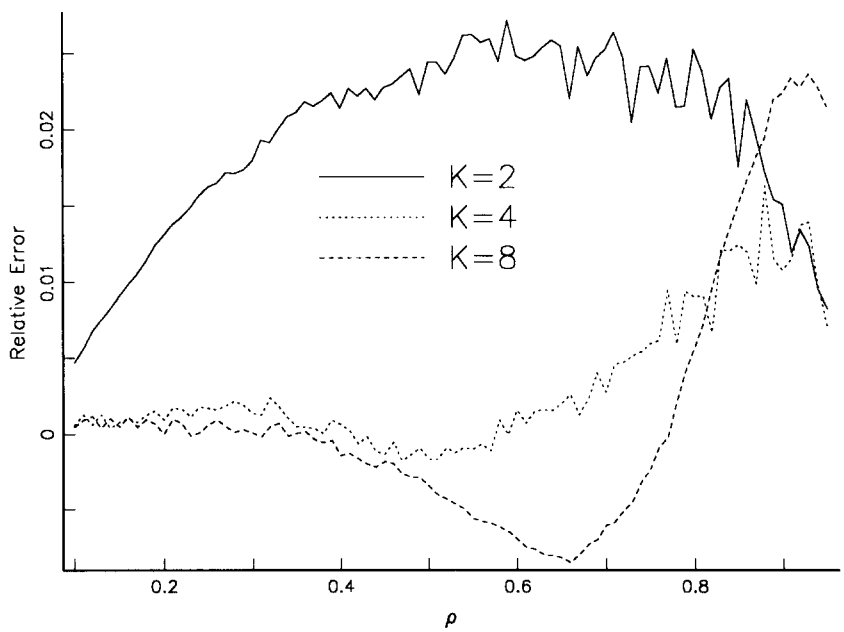

Fig. 8. Relative error: Erlang 2 arrivals, Erlang 2 servers. Basic approximation.

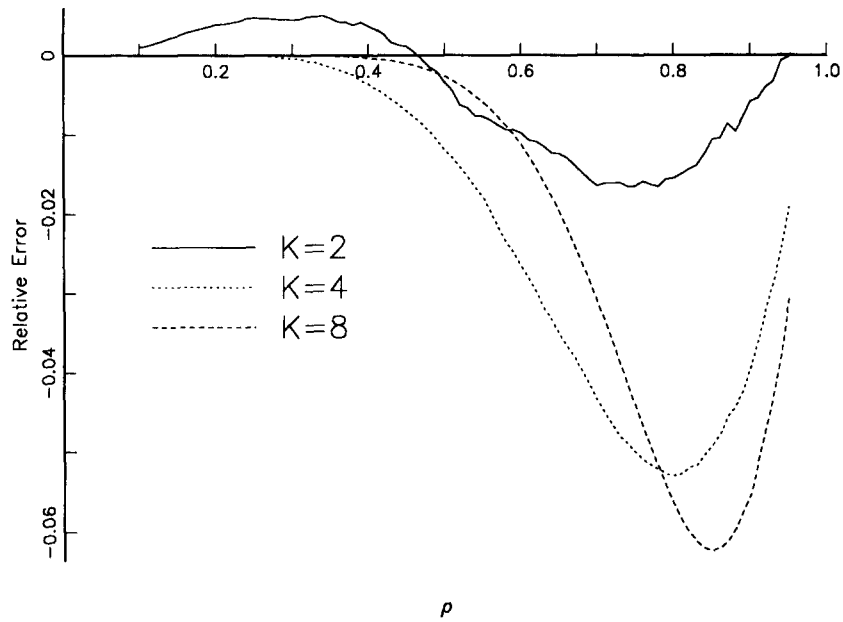

Fig. 9. Relative error: Erlang 2 arrivals, deterministic servers. Basic approximation.

- E_2 interarrival times, deterministic service times (Fig. 9)
- E_5 interarrival times, E_5 service times (Fig. 10)
- Deterministic interarrival times, exponential service times (Fig. 11)
- H_2 interarrival times, E_2 service times (Fig. 12)
- H_2 interarrival times, H_2 service times (Fig. 13)

The service time distributions were scaled to have a mean of 1. The 2-stage Erlang distribution has a coefficient of variation squared of 0.5, and the hyperexponential has a coefficient of varia-

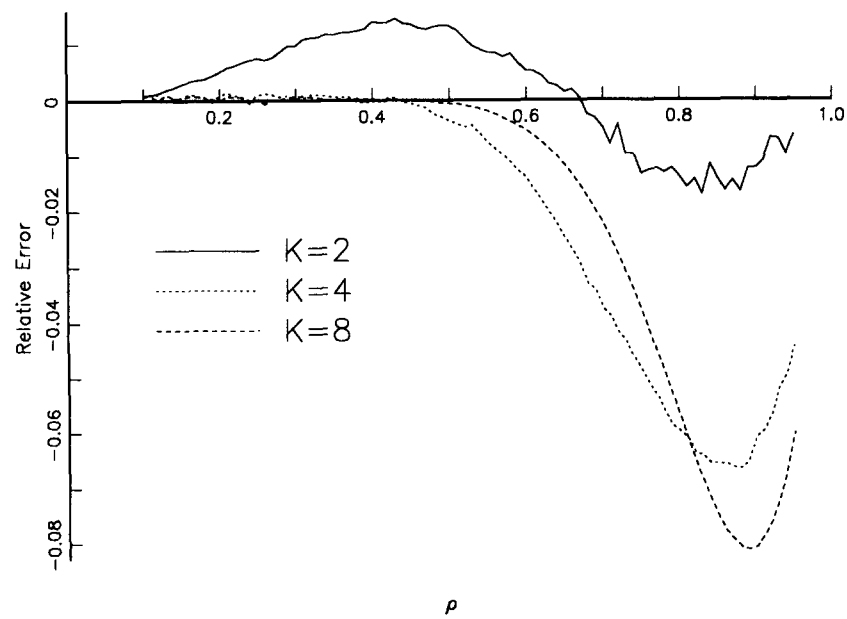

Fig. 10. Relative error: Erlang 5 arrivals, Erlang 5 servers. Basic approximation.

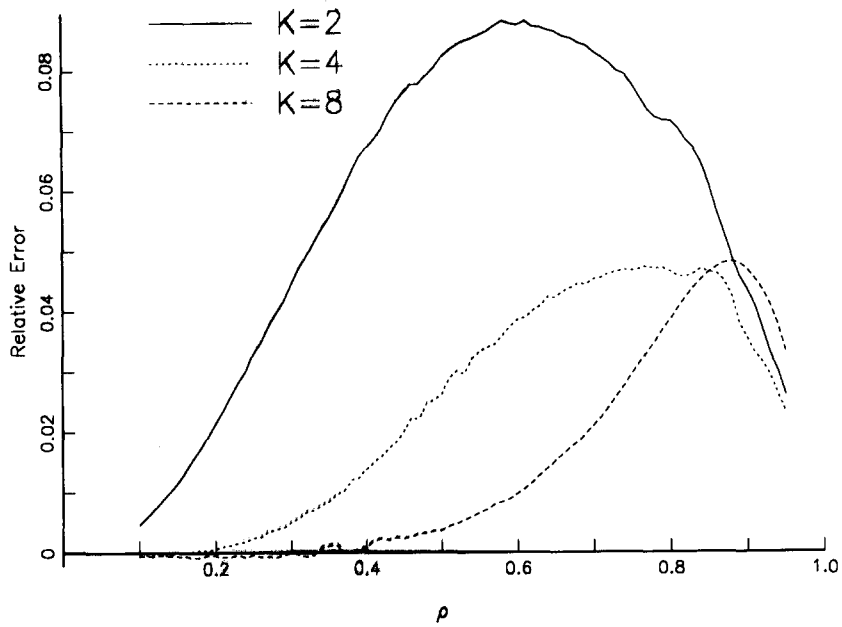

Fig. 11. Relative error: deterministic arrivals, exponential servers. Basic approximation.

tion squared of 4. It was generated by sampling an exponential distribution with parameter 0.375 with probability 0.351, and an exponential distribution with parameter 10 with probability 0.649. The coefficient of variation of the hyperexponential distribution lies well outside the range of validity of eq. (15). In spite of this, the errors for hyperexponential interarrival and service times are reasonable. Our results for these distributions are shown in Figs. 3–13, where we plot the relative errors for $K = 2$, 4 and 8. The simulated points again have 99% confidence intervals that

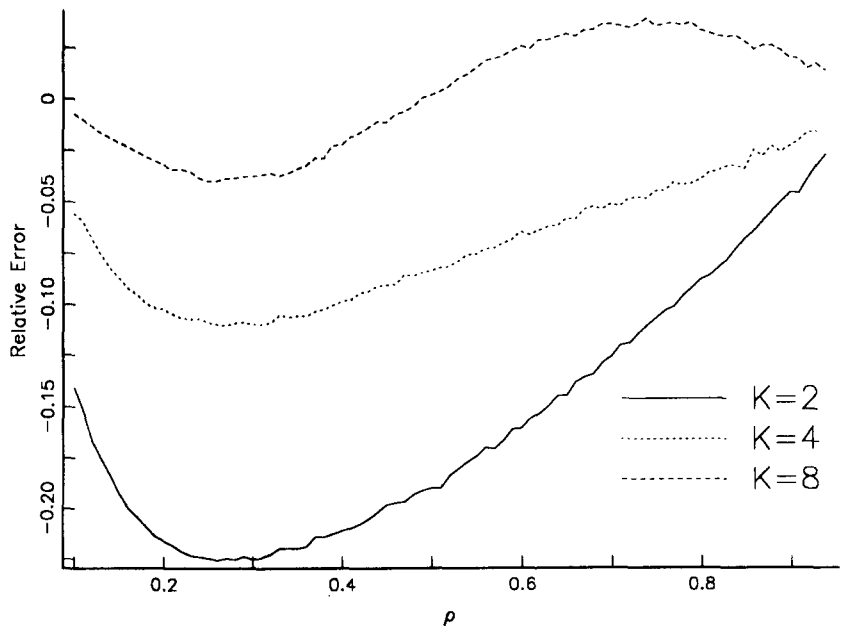

Fig. 12. Relative error: hyperexponential arrivals, Erlang servers. Basic approximation.

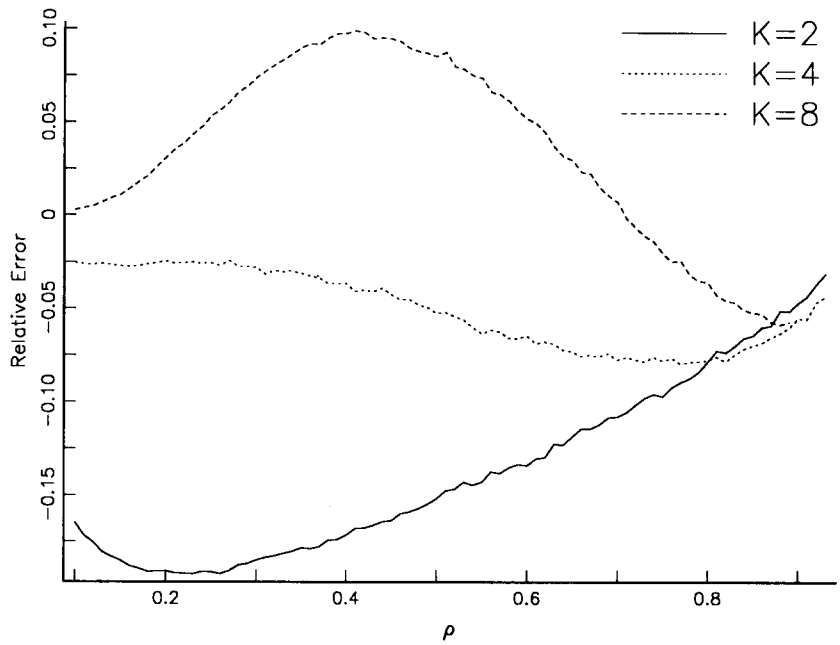

Fig. 13. Relative error: hyperexponential arrivals, hyperexponential servers. Basic approximation.

are 1% of the simulated mean. The curves reveal some interesting behavior. When the coefficients of variation of both the interarrival and service times are small, the approximation is extremely good, with errors of 2% or less. When the coefficient of variation of either the interarrival time or the service time rises to 2 (the hyperexponential case), the errors are substantially larger, often approaching 20%. There seem to be three separate sources of error:

(1) As the coefficient of variation increases, we conjecture that the queue length distribution

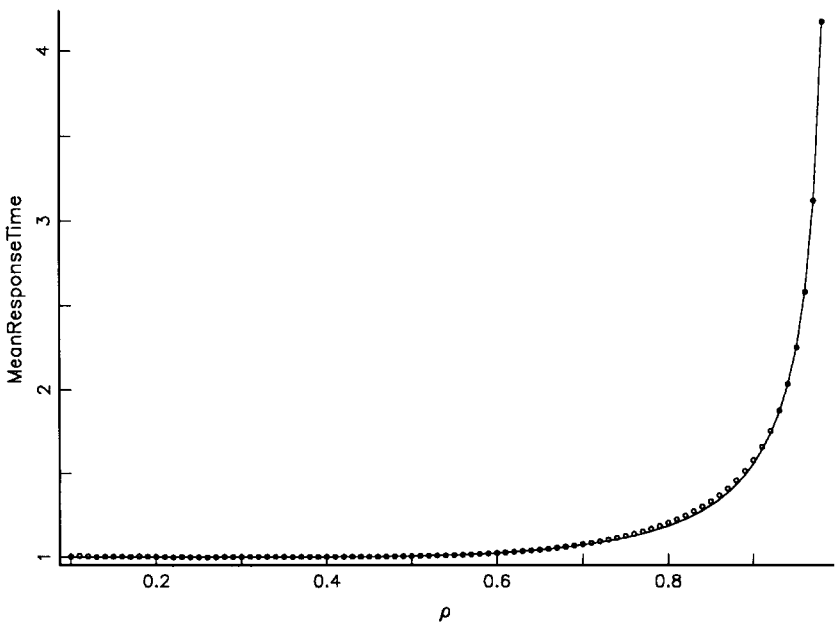

Fig. 14. Comparison of simulation and approximation for $K = 16$. Improved approximation.

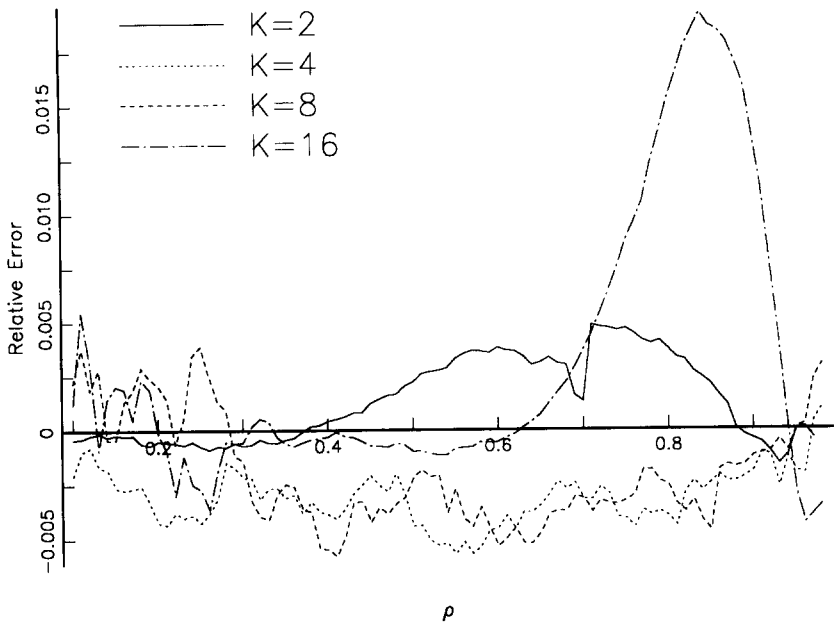

Fig. 15. Relative error: Poisson arrivals, exponential servers. Improved approximation.

is inaccurately modelled by a geometric random variable, especially at intermediate utilizations, and for n comparable to K. For large utilizations, however, Kingman's upper bound [11] on the tail of the distribution assures that the probabilities eventually do fall off geometrically. For very small utilizations the approximation is accurate as the probability of finding an empty queue approaches 1.

(2) An arriving job that is routed to a non-empty queue has to wait for the residual lifetime of the job in service and for the service times of

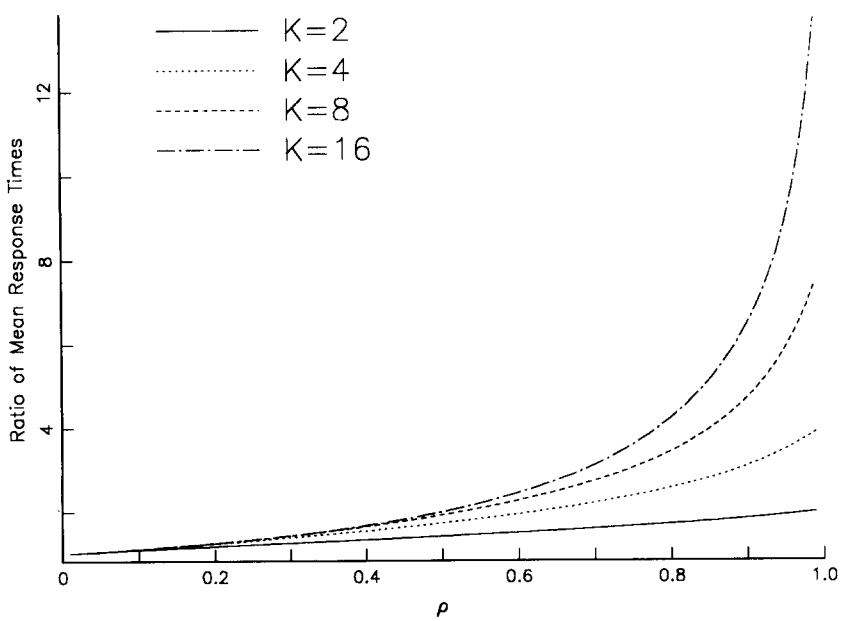

Fig. 16. Ratio of expected response times of Bernoulli splitting to shortest queue.

the jobs in the waiting room. The approximation assumes that the mean residual life and the mean service time of a job are equal. In light or heavy traffic this is a reasonable assumption. In light traffic the arrival sees an empty queue with probability close to 1, while in heavy traffic the residual lifetime of the job in service is a small fraction of the arriving job's waiting time. At intermediate utilizations, however, this can be a source of error.

(3) Equation (15) breaks down, and consequently our estimate of $\hat{\rho}$ is poor.

In the next section we improve the approximation for the case where there are Poisson arrivals and exponential service times.

4. Improvements for Poisson arrivals and exponential servers

In this section we outline an improved version of the basic approximation for Poisson arrivals and Exponential servers. For a derivation of this approximation, see [15]. The improved approximation is based on the observation that the errors obtained in the basic approximation are systematic and have a specific form. In order to present the approximation we first have to introduce some auxiliary equations. Let

$$a(\rho) = 1 - \frac{K\rho}{K+4}$$

$$b(\rho) = \frac{K\rho}{(K+4)(K-1)}$$

$$\bar{\xi} = \frac{\rho(1 - K\rho^{K-1} + (K-1)\rho^K)}{(1-\rho)(1-\rho^K)} \tag{18}$$

$$\bar{q} = a(\rho) + b(\rho)\bar{\xi}$$

$$S(\rho) \simeq \frac{K(1-\rho)}{1-\rho^K}\{\rho^K + \bar{q}(1-\rho^K)\}, \tag{19}$$

$\alpha_1 = 0.0455$, $\alpha_2 = 0.7678$, $\gamma_1 = 0.0216$ and $\gamma_2 = 0.0045$

$$r_K = \gamma_1 \log_2(K) + \gamma_2$$

$$i_K = -1/\log_2\{\alpha_1 \log_2(K) + \alpha_2\}$$

$$R(\rho) = \frac{1}{1 - 4r_K\rho^{i_K}(1-\rho^{i_K})}. \tag{20}$$

Using the above, the final approximation can be written as

$$\begin{aligned}\bar{W}_K^{\text{final}}(\rho) &\simeq \bar{W}_{M/M/K}(\rho)S(\rho)R(\rho)\\ &= \bar{W}_K(\rho)\{\rho^K + \bar{q}(1-\rho^K)\}R(\rho)\\ &= \frac{P_K(\rho)}{1-\rho^K}\{\rho^K + \bar{q}(1-\rho^K)\}R(\rho).\end{aligned} \tag{21}$$

Finally, we compare the approximation given by eq. (21) to simulation. A comparison of the simulated mean response time and the approximation for $K = 16$ is shown in Fig. 14. It is clear from this graph that the approximation is very close to the simulated values. In Fig. 15 we plot the relative error for the approximation for K in the range of 2 to 16. The maximum relative error is seen to be less than one half of one percent for $K \leqslant 8$ and achieves a maximum of just under 2 percent for $K = 16$. The precision we required in our simulations made simulating the $K = 32$ case prohibitively expensive. In fact, for $K = 32$ and $\rho = 0.96$, several days of continuous running on an IBM 3090 yielded only 5 regenerative cycles, which was too few for our purposes.

5. Applications

In this section we use the approximation of Section 4 to compare a static load balancing scheme with that of shortest queue routing and also to determine the relative performance of two multiprocessor architectures.

5.1. Static versus dynamic load balancing

Static load balancing does not use state information to determine where to route jobs. It is clear that an optimal dynamic policy must have a lower expected response time than an optimal static policy. In this section we compare the performance of two static policies, Bernoulli splitting and round robin, to that of shortest queue routing. In the Bernoulli splitting policy, jobs are routed to queues randomly with a uniform probability of $1/K$. Such a policy forms K independent $M/M/1$ queues and thus the expected response time is given by

$$T_b = \frac{1}{\mu - \lambda/K}. \tag{22}$$

In Fig. 16 we plot the ratio of the expected response times of Bernoulli splitting, as given by eq. (22), to that of shortest queue routing, as given by (21). This ratio is always greater than 1, indicating that the dynamic policy is always better than the static policy and inspection of the figures shows that the improvement in performance of the dynamic policy increases with ρ and with K. Since, as $\rho \to 1$, the response time of the shortest queue system approaches that of a $M/M/K$ system, the ratio of expected response times can be as large as K times. One should note, however, that the response time for the shortest queue system for large values of K is small, as with large probability, an arriving customer will be routed to an empty queue.

It is well known that a static policy which simply sends jobs in a round robin fashion to the processors has a lower expected response time than Bernoulli splitting. In round robin, new arrivals are routed in a cyclic fashion among the processors, i.e. the queues are labeled $0, 1, \ldots, K-1$ and the nth arrival is routed to queue n mod K. Such a policy also leads to independent queues which are $E_K/M/1$ queues. The response time for this system can be written as

$$T_{rr} = \frac{1}{\mu - \mu\sigma} \tag{23}$$

where σ is the unique solution in the range of (0, 1) of

$$\sigma = \left(\frac{K\lambda}{\mu - \mu\sigma + K\lambda} \right)^K . \tag{24}$$

The solution for σ can be found using the bisection method and in Fig. 17 we plot the ratio of the mean response time of the round robin scheme to that of shortest queue routing. Here again, the dynamic policy always has a better performance than the static policy and this improvement increases with ρ and K. It is interesting to note that the mean response time of round robin is substantially better than that of Bernoulli scheduling. These results unequivocally demonstrate the benefits of a dynamic load balancing strategy over static load balancing.

5.2. Multiprocessor architectures

In this section we compare two different multiprocessor architectures, one in which jobs are executed concurrently by all processors and the other where jobs are executed sequentially and shortest queue routing scheduler is used. The parallel model corresponds to a fork/join configuration. In this model we assume that jobs can be split into subtasks. Specifically, we assume that

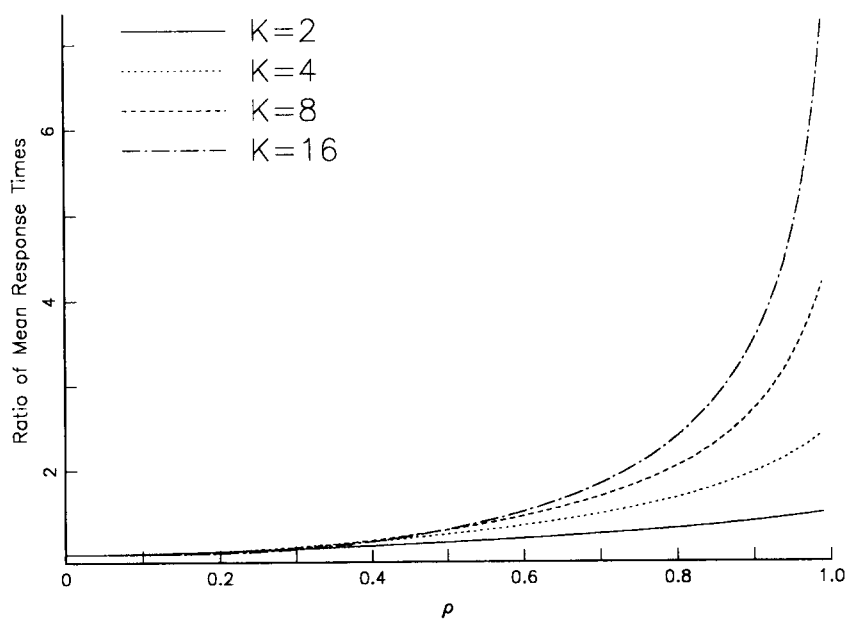

Fig. 17. Ratio of expected response times of Round Robin to shortest queue.

each job consists of K subtasks which can be executed concurrently. Upon a job arrival, the K subtasks are scheduled so that one goes to each of the queues (assumed to be of infinite capacity) of the K processors. This corresponds to the fork operation. These subtasks are assumed to have independent and identically distributed service requirements. The job is considered to be completed when all of its subtasks have finished execution. This corresponds to the join operation and the time that the first completed task spends waiting for the remainder of its sibling tasks to complete is called the *synchronization* time. The expected job response time is thus the elapsed time between job arrival and the finish of execution of all of its subtasks. In [16] an approximation is given for such a system where arrivals come from a Poisson point source and task service times are assumed to be exponentially distributed. An expression for the expected job response time is given by

$$T_{FJ} = \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\rho\right]\frac{(12-\rho)}{8(\mu-\lambda)} \qquad (25)$$

where $H_j = \sum_{j=1}^{j} 1/i$. It can be shown that as the load of the system increases, a proportionately larger amount of job response time is spent in waiting for sibling tasks to complete execution, i.e., in the synchronization stage. This suggests that a possibly better architecture, in terms of expected response time, might be to avoid the joining operation by executing all the subtasks of a job only on one processor. One approach to scheduling jobs in this way would be to use Bernoulli scheduling with a uniform probability of $1/K$. The total time needed to complete all of the K exponential subtasks of a job is then distributed as an Erlang distribution and the resultant queueing system is a set of independent $M/E_K/1$ queues. In [16] the fork/join system and Bernoulli systems were compared and it was shown that the fork/join system always had a lower expected response time.

Since shortest queue routing has better response time characteristics than Bernoulli routing, it is natural to investigate how shortest queue routing compares with fork/join scheduling. This comparison is shown in Fig. 18 and reveals an interesting trade-off between the two systems. From the figure one sees that the parallel fork/join architecture has a lower expected response time for low to medium utilizations but that for high utilizations, shortest queue routing is better. The reason for this can be best explained by considering the case where $\rho \to 0$. For such a system, there is little queueing and the

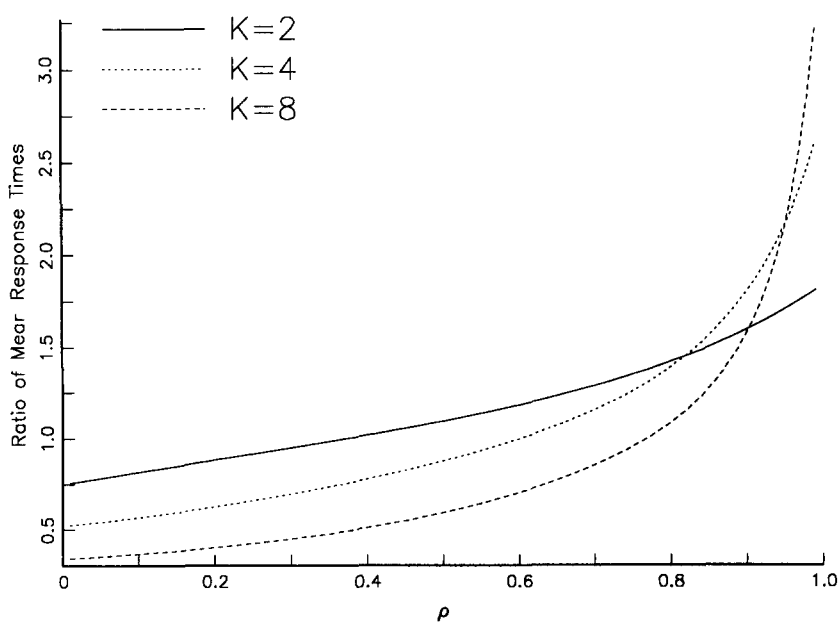

Fig. 18. Ratio of expected response times of fork/join to shortest queue.

mean response time is equal to the time needed to complete the service of all subtasks of a job. In the fork/join system this is distributed as the maximum of K exponential random variables given by H_K/μ. This is clearly less than K/μ which is obtained from shortest queue routing. Thus for low utilizations executing jobs in parallel is beneficial. As pointed out above, however, for large utilizations a large percentage of a jobs response time in the fork/join system is spent in the synchronization stage. This arises from the fact that in such a system the queue lengths between the K queues are very staggered. Shortest queue routing, which tends to minimize the variation between the different queue lengths and also avoids the joining operation overcomes the benefits of parallelism and achieves a lower expected response time.

6. Conclusions and future research

In this paper we have derived an approximation for the expected response time for a queueing system that uses shortest queue routing. The approximation is based on both theoretical and experimental results and is very accurate. We presented some results that compared the relative performance of shortest queue routing to that of two different static load balancing schemes as well as two multiprocessor architectures.

We are currently pursuing many extensions and applications of the basic approximation. We are also investigating more general models, such as having a mixture of traffic consisting of exogeneous arrival processes dedicated to each processor and common traffic that is routed using shortest queue routing. Such a system could model a dynamic load balancing scheme where each processor had some dedicated user traffic. If one combines traffic types of shortest queue routing and fork/join traffic, general models of replicated data bases can be analyzed.

References

[1] J.P.C. Blanc, A note on waiting times in systems with queues in parallel, *J. Appl. Probab.* **24** (1987) 540–546.

[2] J.W. Cohen and O.J. Boxma, *Boundary Value Problems in Queueing System Analysis* (North-Holland, Amsterdam, 1983).

[3] B.W. Conolly, The autostrada queueing problem, *J. Appl. Probab.* **21** (1984) 394–403.

[4] A. Ephremides, P. Varaiya and J. Walrand, A simple dynamic routing problem, *IEEE Trans. Autom. Control* **25** (4) (1980) 690–693.

[5] L. Flatto and H.P. McKean, Two queues in parallel, *Comm. Pure Appl. Math.* **30** (1977) 255–263.

[6] G. Foschini and J. Salz, A basic dynamic routing problem and diffusion, *IEEE Trans. Comm.* **26** (3) (1978) 320–328.

[7] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory* (Wiley, New York, NY, 1985, 2nd ed.).

[8] J.A. Gubner, B. Gopinath and S.R.S. Varadhan, Bounding functions of Markov processes and the shortest queue problem, Preprint, University of Maryland, 1988.

[9] S. Halfin, The shortest queue problem, *J. Appl. Probab.* **22** (1985) 865–878.

[10] J.F.C. Kingman, Two similar queues in parallel, *Biometrika* **48** (1961) 1316–1323.

[11] L. Kleinrock, *Queueing Systems, Vol. 1: Theory* (Wiley, New York, NY, 1975).

[12] C. Knessl, B.J. Matkowsky, Z. Schuss and C. Tier, Two parallel $M/G/1$ queues where arrivals join the system with the smaller buffer content, *IEEE Trans. Comm.* **35** (11) (1987) 1153–1158.

[13] S.S. Lavenberg, *Computer Performance Modeling Handbook* (Academic Press, New York, NY, 1983).

[14] W.G. Marchal, An approximate formula for waiting times in single server queues, *AIIE Trans.* **8** (1976) 473.

[15] R. Nelson and T.K. Philips, An approximation to the response time for shortest queue routing, *Performance Evaluation Rev.* **7** (1) (1989) 181–189.

[16] R. Nelson and A. Tantawi, Approximate analysis of fork/join synchronization in parallel queues, *IEEE Trans. Comput.* **37** (6) (1988) 739–743.

[17] B.M. Rao and M.J.M. Posner, Algorithmic and approximate analysis of the shorter queue model, *Naval Res. Logist.* **34** (1987) 381–398.

[18] Shantikumar, J.G., Bounds and an approximation for single server queues, *J. Oper. Res. Soc. Japan* **26** (1988) 118–134.

[19] Shantikumar, J.G., On the approximations to the single server queue, *Int. J. Prod. Res.* **18** (1980) 761–773.

[20] Shore, H., Simple approximations for the $GI/G/c$ queue, I: the steady state probabilities, *J. Oper. Res. Soc.* **39** (3) (1988) 279–284.

[21] Winston, W., Optimality of the shortest line discipline, *J. Appl. Probab.* **14** (1977) 181–189.