ARO 13179.6-M

ORC 77-34
DECEMBER 1977

LEVE

(12)

APPROXIMATIONS IN FINITE CAPACITY MULTI-SERVER QUEUES WITH POISSON ARRIVALS

by

SHIRLEY A. NOZAKI

and

SHELDON M. ROSS

OPERATIONS RESEARCH CENTER

UNIVERSITY OF CALIFORNIA · BERKELEY

APPROXIMATIONS IN FINITE CAPACITY MULTI-SERVER QUEUES WITH POISSON ARRIVALS[†]

Operations Research Center Research Report No. 77-34

Shirley A. Nozaki [redacted] Sheldon M. Ross

Dec[redacted]977

U. S. Army Research Office - Research Triangle Park

N00014-77-C-0299
DAAG29-76-G-0042

Operations Research Center
University of California, Berkeley

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

[†]Partially supported by the Office of Naval Research under Contract N00014-77-C-0299 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ORC 77-[illegible]	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) APPROXIMATIONS IN FINITE CAPACITY MULTI-SERVER QUEUES WITH POISSON ARRIVALS		5. TYPE OF REPORT & PERIOD COVERED Research Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Shirley A. Nozaki and Sheldon M. Ross		8. CONTRACT OR GRANT NUMBER(s) DAAG29-76-G-0042
9. PERFORMING ORGANIZATION NAME AND ADDRESS Operations Research Center University of California Berkeley, California 94720		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P-13179-M
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE December 1977
		13. NUMBER OF PAGES 17
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also supported by the Office of Naval Research under Contract N00014-77-C-0299.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Queueing
Multi Server
Finite Capacity
Average Delay
Approximations

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(SEE ABSTRACT)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

In this paper, ~~we~~ consider an M/G/k queueing model having finite capacity N . That is, a model in which customers, arriving in accordance with a Poisson process having rate λ , enter the system if there are less than N others present when they arrive, and are then serviced by one of k servers, each of whom has service distribution G . Upon entering, a customer will either immediately enter service if at least one server is free or else join the queue if all servers are busy. Our results will be independent of the order of service of those waiting in queue as long as it is supposed that a server will never remain idle if customers are waiting. To facilitate the analysis, however, ~~we will~~ ~~suppose~~ a service discipline of "first come first to enter service."

APPROXIMATIONS IN FINITE CAPACITY
MULTI-SERVER QUEUES WITH POISSON ARRIVALS

by

Shirley A. Nozaki and Sheldon M. Ross

0. INTRODUCTION

In this paper, we consider an M/G/k queueing model having finite capacity N. That is, a model in which customers, arriving in accordance with a Poisson process having rate λ, enter the system if there are less than N others present when they arrive, and are then serviced by one of k servers, each of whom has service distribution G. Upon entering, a customer will either immediately enter service if at least one server is free or else join the queue if all servers are busy. Our results will be independent of the order of service of those waiting in queue as long as it is supposed that a server will never remain idle if customers are waiting. To facilitate the analysis, however, we will suppose a service discipline of "first come first to enter service."

Our objective is to obtain an approximation for the average time spent waiting in queue by an entering customer. This is done mainly by means of an approximation assumption, presented in Section 2, and used in Section 3 to derive the approximation. In Section 4, we let $N = \infty$ and relate the approximation to the existing literature.

1. BASIC DEFINITIONS AND FUNDAMENTAL EQUATION

We shall need the following notation:

P_i : the steady state probability that there are i people in the system.

S : a service time random variable, i.e., $P\{S \leq x\} = G(x)$.

W_Q : the average amount of time that an entering customer spends waiting in queue (does not include service time).

L_Q : the (time) average number of customers waiting in queue.

V : the (time) average amount of work in the system, where the work in the system at any time is defined to be the total (of all servers) amount of service time necessary to empty the system of all those presently either being served or waiting in queue.

V_e : the average amount of work as seen by an entering arrival.

We will make use of the following idea (previously exploited in such papers as [1], [2] and [8]) that if a (possibly fictitional) cost structure is imposed, so that entering customers are forced to pay money (according to some rule) to the system, then the following identity holds--namely,

(1) time average rate at which the system earns = average arrival rate of entering customers × average amount paid by an entering customer.

A heuristic proof of the above is that both sides of (1) times T is approximately equal to the total amount of money paid to the system by time T , and the result follows by dividing by T and then letting $T \to \infty$.[+]

[+] A rigorous proof along these lines can easily be established in the models we consider since all have regeneration points. More general conditions under which it is true are presented in [1].

By choosing appropriate cost rules, many useful formulae can be obtained as special cases of (1). For instance, by supposing that each customer pays \$1 per unit time while in service, Equation (1) yields that

$$\text{average number in service} = \lambda(1 - P_N)E[S] \ .$$

Similarly, by supposing that each customer pays \$1 per unit time while waiting in queue, we obtain from (1) that

$$L_Q = \lambda(1 - P_N)W_Q \ .$$

Also, if we suppose that each customer in the system pays \$x per unit time whenever its remaining service times is x , then (1) yields that

$$(2) \qquad V = \lambda(1 - P_N)E\left[SW_Q^* + \int_0^S (S - x)dx\right] = \lambda(1 - P_N)\left[E[S]W_Q + E[S^2]/2\right]$$

where W_Q^* is a random variable representing the (limiting) amount of time that the n^{th} entering customer spends waiting in queue.

Another important fact which we shall use is that, since our arrival stream of customers is a Poisson process, the probability structure of what an arrival observes is identical to the steady state probability structure of the system.

2. THE APPROXIMATION ASSUMPTION

Let G_e denote the equilibrium distribution of G. That is,

$$G_e(x) = \int_0^x \frac{(1 - G(y))}{E[S]}\, dy ,$$

also let

$$\delta(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} .$$

We assume throughout that $\int x dG_e(x) = E[S^2]/2E[S]$ is finite. We make the following approximation assumption.

<u>Approximation Assumption:</u>

Given that a customer arrives to find i busy servers, $i > 0$, then at the time that he enters service, the remaining service times of the other $i - \delta(i,k)$ customers being served has a joint distribution that is approximately that of independent random variables each having distribution G_e.

<u>Heuristic Remarks Concerning the A.A.:</u>

1. In the infinite capacity case, the A.A. appears to be approximately true either in heavy traffic (that is, as $\lambda E[S] \to k$) or in light traffic (that is, as $\lambda E[S] \to 0$). This is so in heavy traffic since the great majority of arrivals will encounter a large queue and as a result the k departure processes (one for each server) they observe will be approximately independent delayed renewal pro-

cesses. Hence, considering those customers served by server i, it follows that when they enter service they would have been observing $k - 1$ independent delayed renewal processes for a large time, and the A.A. follows since the limiting distribution of excess in a renewal process is just G_e.

In extremely light traffic, the great majority of arrivals will find either 0 or 1 busy servers. Now, since Poisson arrivals see the system as it is (averaged over all time), it follows that arrivals finding 1 server busy would encounter the same additional service time (for the busy server) as would random (and uniform) time sampling of the excess of a renewal process. Hence, the A.A. follows in light traffic from the renewal process (excess) result.

2. Additional heuristics for the A.A. follows from the fact that it is known to be (exactly) true when no queue is allowed (see [9]).

3. THE APPROXIMATION

Since V is the average amount of work as seen by an arrival, it follows by conditioning upon whether or not an arrival enters the system that

$$V = (1 - P_N)V_e + P_N \times \text{(average work as seen by a lost customer)}.$$

In accordance with our basic A.A., it seems reasonable to approximate the average work as seen by a lost customer by $k\,\frac{E[S^2]}{2E[S]} + (N - k)E[S]$. Hence, we have that

$$(3) \qquad V \doteq (1 - P_N)V_e + P_N\left(k\,\frac{E[S^2]}{2E[S]} + (N - k)E[S]\right) .$$

Now for any arbitrary customer that enters the system we have the following identity

work as seen by the entering customer =

$k \times$ time entering customer spends waiting in queue + R

where R is defined to be the sum of the remaining service times of those being served when the customer enters service. Taking expectations yields that

$$(4) \qquad V_e = kW_Q + E[R] .$$

To obtain $E[R]$, we condition on B_e , the number of servers that are busy when the customer enters the system:

$$E[R] = E(E[R \mid B_e]) \tag{5}$$

$$\doteq E[B_e - \delta(B_e,k)]\frac{E[S^2]}{2E[S]} \quad \text{by the A.A.}$$

Now,

$$\lambda(1 - P_N)E[S] = \text{average number of busy servers as seen by an arrival} \tag{6}$$

$$= (1 - P_N)E[B_e] + kP_N \ .$$

Also,

$$E[\delta(B_e,k)] = \left(1 - P_N - \sum_{j=0}^{k-1} P_j\right)\Big/(1 - P_N) \tag{7}$$

and so from (3)-(7) and Equation (2) we obtain that

$$W_Q \doteq \frac{\frac{E[S^2]}{2E[S]}\left(1 - P_N - \sum_{j=0}^{k-1} P_j\right) - (N - k)P_N E[S]}{(1 - P_N)(k - \lambda E[S])} \ . \tag{8}$$

Therefore, it remains to obtain P_N and P_j , $0 \le j \le k - 1$. To do so, we impose the following fictitional cost structure--namely, that the i oldest customers in the system pay \$1 per unit time, $i = 1, 2, \ldots, k$, where the age of a customer is measured from the moment it enters the system. Hence, letting $S_1^e, S_2^e, \ldots, S_{k-1}^e$ denote $k - 1$ independent random variables each having distribution G_e , we obtain from Equation (1) that

$$P_1 + 2P_2 + \dots + (i-1)P_{i-1} + i(1 - P_0 - \dots - P_{i-1})$$

$$\doteq \lambda(P_0 + \dots + P_{i-1})E[S] + \lambda P_i E\left[\left(S - \min\left(S_1^e, S_2^e, \dots, S_i^e\right)\right)^+\right]$$

$$+ \lambda P_{i+1} E\left[\left(S - \text{2nd smallest of } \left(S_1^e, \dots, S_{i+1}^e\right)\right)^+\right]$$

$$+$$

$$(9) \quad \vdots$$

$$+ \lambda P_{k-2} E\left[\left(S - (k-1-i)\text{th smallest of } \left(S_1^e, \dots, S_{k-2}^e\right)\right)^+\right]$$

$$+ \lambda(1 - P_N - P_0 - \dots - P_{k-2})E\left[\left(S - (k-i)\text{th smallest of } \left(S_1^e, \dots, S_{k-1}^e\right)\right)^+\right]$$

$$i = 1, \dots, k-1$$

$$P_1 + 2P_2 + \dots + (k-1)P_{k-1} + k(1 - P_0 - \dots - P_{k-1}) = \lambda(1 - P_N)E[S]$$

$\left(\text{where } x^+ = \begin{cases} x & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}\right)$. To understand the above equations, suppose first that $i < k$. Now, as only the i oldest pay, it follows that when j customers are present the system earns at a rate j when $j \le i$ and at a rate i when $j > i$. Hence, the left side of Equation (9) represents the average rate at which the system earns. On the other hand, an arrival finding fewer than i customers already in the system will immediately go into service and will pay a total amount equal to his service time; while an arrival finding j present, $k - 1 \ge j \ge i$ will also go immediately into service but will only begin paying when $j - i + 1$ of the j others in service leave. Thus, in this latter case, under the A.A., the arrival would expect to pay a total of $E\left[\left(S - (j+1-i)\text{th smallest of } \left(S_1^e, S_2^e, \dots, S_j^e\right)\right)^+\right]$. Finally, if the arrival found more than $k - 2$ busy servers, then he will begin paying after $k - 1$ of those customers in service when he enters service leave the system. This explains the first $k - 1$ of the set of Equation (9). The last equation (when $i = k$) easily follows since in this case each customer will pay a total equal to his time in service.

To simplify the set of Equation (9), we will need the following lemma.

<u>Lemma 1</u>:

If $S, S_1^e, \ldots, S_r^e$ are independent random variables such that S has distribution G and the others G_e, then

$$E\left[\left(S - j\text{th smallest of } \left(S_1^e, \ldots, S_r^e\right)\right)^+\right] = \frac{r + 1 - j}{r + 1} E[S] .$$

<u>Proof</u>:

Using the identity $(x - y)^+ = x - \min(x,y)$, we have that

$$E\left[\left(S - j\text{th smallest of } S_1^e, \ldots, S_r^e\right)^+\right] =$$

$$E[S] - E\left[\min\left(S , j\text{th smallest of } S_1^e, \ldots, S_r^e\right)\right] .$$

Now,

$$E\left[\min\left(S , j\text{th smallest of } S_1^e, \ldots, S_r^e\right)\right]$$

$$= \int_0^\infty P\{S > a\} P\left\{j\text{th smallest of } \left(S_1^e, \ldots, S_r^e\right) > a\right\} da$$

$$= \int_0^\infty (1 - G(a)) \sum_{i=0}^{j-1} \binom{r}{i} (G_e(a))^i (1 - G_e(a))^{r-i} da$$

$$= E[S] \sum_{i=0}^{j-1} \binom{r}{i} \int_0^1 y^i (1 - y)^{r-i} dy$$

$$= E[S] \sum_{i=0}^{j-1} \binom{r}{i} \frac{i!(r - i)!}{(r + 1)!}$$

$$= E[S] \frac{j}{r + 1}$$

which proves the lemma.■

It follows from Lemma 1 that the equations for P_N and P_j, $0 \le j \le k - 1$ depend on G only through $E[S]$. Hence, as the equations are exactly true when G is exponential, it follows (since for fixed P_N, it can be shown that the set of equations has at most one solution) that the P_j, $0 \le j \le k - 1$ have the same relationship to P_N as when G is exponential. Thus, we are left to determine P_N, which we will approximate by the answer in the exponential case. In other words, we shall use the exact result for P_j, $0 \le j \le k - 1$, P_N when G is exponential as our approximation. This yields, from Equation (8), that

$$(10) \qquad W_Q \doteq \frac{\dfrac{E[S^2]}{2E[S]} \displaystyle\sum_{j=k}^{N-1} \frac{(\lambda E[S])^j}{k!k^{j-k}} - (N - k) \frac{E[S](\lambda E[S])^N}{k!k^{N-k}}}{\left[\displaystyle\sum_{j=0}^{k-1} \frac{(\lambda E[S])^j}{j!} + \sum_{j=k}^{N-1} \frac{(\lambda E[S])^j}{k!k^{j-k}}\right](k - \lambda E[S])} .$$

4. THE INFINITE CAPACITY CASE

In the infinite capacity case $N = \infty$, the approximation (10) reduces, when $\lambda E[S] < k$, to

$$(11) \quad W_Q \doteq \frac{\lambda^k E[S^2](E[S])^{k-1}}{2(k-1)!(k-\lambda E[S])^2\left[\sum_{j=0}^{k-1}\frac{(\lambda E[S])^j}{j!} + \frac{(\lambda E[S])^k}{(k-1)!(k-\lambda E[S])}\right]} .$$

Some remarks are in order:

1. In [4], Kingman obtained bounds on W_Q for the general queueing system GI/G/k. When adapted to the M/G/k case of Poisson arrivals, his inequalities are

$$\frac{E[S^2]}{2E[S](k-\lambda E[S])} - \frac{[E[S^2] + k/\lambda^2 - (E[S])^2/k]}{2E[S]} \leq W_Q \leq \frac{\lambda[E[S^2] - (E[S])^2] + k/\lambda}{2(k-\lambda E[S])} .$$

It is easily verified that our approximation for W_Q is consistent with Kingman's upper and lower bounds.

2. In [5], Kingman conjectured a heavy traffic approximation for W_Q in GI/G/k models. In the special case of Poisson arrivals, his conjecture is that

$$W_Q \approx \frac{\lambda^2 E[S^2] - \lambda^2 (E[S])^2 + k^2}{2\lambda k(k-\lambda E[S])} \quad \text{when} \quad \lambda E[S] \approx k .$$

Calling the right side of the above K and our approximation, as given by (8), $N - R$, we have

$$\frac{K}{N - R} = \frac{\lambda E[S]}{k\bar{P}_k} + \frac{E[S]}{\bar{P}_k E[S^2]}\left(k^2 - \frac{(\lambda E[S])^2}{\lambda k}\right)$$

where $\bar{P}_k = 1 - \sum_0^{k-1} P_j$. Hence, since in heavy traffic, $E[S] \approx k/\lambda$, $\bar{P}_k \approx 1$, we see that our approximation is consistent with Kingman's heavy traffic conjecture.

3. Numerical tables for L_Q have been published by Hillier and Lo in the special case $M/E_r/k$, where E_r represents an Erlang distribution with r phases. Table 1 compares our approximate formula for L_Q ($= \lambda W_Q$) with the Hillier-Lo tables.
4. Another heavy traffic conjecture was given by Maaloe who in [6] conjectured that for the model $M/E_r/k$

$$W_Q \approx \frac{\lambda E[S^2]}{2k(k - \lambda E[S])} \quad \text{when} \quad \lambda E[S] \approx k \ .$$

As the ratio between our approximation and the above approaches unity in heavy traffic, we see that our approximation is also consistent with this conjecture.

TABLE 1

L_Q for $M/E_r/k$, $L_Q \doteq \dfrac{(\rho k)^{k+1}\,\frac{r+1}{r}}{2(k!)k(1-\rho)^2\left[\sum_{n=0}^{k-1}(\rho k)^n/n! + \frac{(\rho k)^k}{k!(1-\rho)}\right]}$, $\rho = \frac{\lambda E[S]}{k}$

ρ \ k	3			4		5	6	7	8	9	10
	r=2	r=3	r=4	r=2	r=3	r=2	r=2	r=2	r=2	r=2	r=2
.1	.000309	.000274	.000257	.000066	.000059	.000015	.000003	.000001	0	0	0
	.000341	.000319	.000310	.000075	.000071	.000016	.000003	.000001	0	0	0
.3	.022509	.020008	.018758	.011909	.010586	.006473	.003583	.002009	.001137	.000648	.000372
	.023993	.022024	.0210483	.0129386	.0119909	.007142	.004004	.002269	.001297	.000745	.000431
.5	.177632	.157895	.148026	.130435	.115942	.097778	.074357	.057149	.044283	.034537	.027079
	.184420	.16697	.158227	.1370512	.124807	.103778	.079616	.0616708	.048126	.03778	.029798
.7	.861603	.765869	.718002	.750145	.666796	.661217	.587961	.526293	.473555	.427911	.388030
	.877853	.787394	.7420435	.768936	.6916764	.681288	.608581	.547007	.494070	.44803	.407627
.9	5.515162	4.902366	4.595968	5.317335	4.726520	5.146829	4.995848	4.859712	4.735348	4.620612	4.51394
	5.54407	4.9405	4.638435	5.354183	4.775027	5.189889	5.043990	4.912132	4.79145	4.6799	4.57607
.95	12.924872	11.488775	10.770726	12.702721	11.291307	12.508636	12.334640	12.175944	12.029403	11.892822	11.74459
	12.95731	11.53154	10.81833	12.74479	11.34665	12.55858	12.39130	12.23848	12.09718	11.96535	11.84147
.99	72.851736	64.757099	60.709780	72.609459	64.541742	72.395734	72.202451	72.024744	71.859417	71.704239	71.5545
	72.88707	64.80367	60.76161	72.65587	64.60278	72.45146	72.26631	72.09590	71.93724	71.7862	71.64728

top number in box = approximation for L_Q

bottom number in box = exact value as given by Hillier-Lo

REFERENCES

[1] Brumelle, S. L., "On the Relation Between Customer and Time Averages in Queues," Journal of Applied Probability, Vol. 8, pp. 508-520, (1971).

[2] Brumelle, S. L., "A Generalization of $L = \lambda W$ to Moments of Queue Length and Waiting Times," Operations Research, Vol. 20, No. 6, pp. 1127-1136, (December 1972).

[3] Hillier, F. S. and F. D. Lo, "Tables for Multiple-Server Queueing Systems Involving Erlang Distributions," Technical Report No. 31, Department of Operations Research, Stanford University, (December 1971).

[4] Kingman, J. F. C., "Inequalities in the Theory of Queues," Journal of the Royal Statistical Society, Series B, Vol. 32, pp. 102-110,

[5] Kingman, J. F. C., "The Heavy Traffic Approximation in the Theory of Queues," Proceedings of the Symposium on Congestion Theory, pp. 137-170, University of North Carolina, (1965).

[6] Maaloe, E., "Approximation Formulae for Estimation of Waiting-Time in Multiple-Channel Queueing System," Management Science, Vol. 19, No. 6, pp. 703-710, (February 1973).

[7] Newell, G., "Approximate Stochastic Behavior of n-Server Service Systems with Large n," Lecture Notes in Economics and Mathematical Systems, M. Beckmann, G. Goos and H. P. Künzi, eds., Springer-Verlag, (1970).

[8] Stidham, S., "Static Decision Models for Queueing Systems with Non-Linear Waiting Costs," Technical Report No. 9, Stanford University, (1968).

[9] Takács, L., "On Erlang's Formula," Annals of Mathematical Statistics, Vol. 40, No. 1, pp. 71-78, (1969).