J. Appl. Prob. **33**, 256–266 (1996)
Printed in Israel

A SAMPLE PATH ANALYSIS OF THE DELAY IN THE *M/G/C* SYSTEM

SRIDHAR SESHADRI,* *New York University*

Abstract

Using sample path analysis we show that under the same load the mean delay in queue in the *M/G/2* system is smaller than that in the corresponding *M/G/1* system, when the service time has either the DMRL or NBU property and the service discipline is FCFS. The proof technique uses a new device that equalizes the work in a two server system with that in a single sterver system. Other interesting quantities such as the average difference in work between the two servers in the *GI/G/2* system and an exact alternate derivation of the mean delay in the *M/M/2* system from sample path analysis are presented. For the same load, we also show that the mean delay in the *M/G/C* system with general service time distribution is smaller than that in the *M/G/1* system when the traffic intensity is less than $1/c$.

SAMPLE PATH ANALYSIS; MULTI-SERVER QUEUE; BOUNDS FOR MEAN DELAY

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60K25

1. Introduction

Using sample path analysis, in this paper we show that under the same load the mean delay in the FCFS *M/G/2* system is smaller than that in the *M/G/1* system with a fast server when the service time distribution has the DMRL or NBU property. The result is new. Earlier known results of this form are for the *GI/M/C* queue (Cox and Smith 1961) and the *GI/D/C* queue (Mori 1975, Daley and Rolski 1984); see also the discussion on p. 497 of Wolff (1989). In doing the sample path analysis we use a new method, called the work equalization scheme, to compare the time average work in the single and two server systems. Using the same method we prove that, if the traffic intensity is less than $1/c$, the mean delay in the *M/G/C* system with general service time distribution is smaller than that in the *M/G/1* system. This is an improvement upon the result of Suzuki and Yoshida (1970). The method also yields interesting bounds for other quantities and leads to a new way of determining the mean delay in the *M/M/2* system without actually solving balance equations.

2. The *M/G/2* system

Consider a two server and a single server system, both operating under the FCFS discipline, which have common inter-arrival times but the service times in the two server

Received 6 July 1993; revision received 14 January 1994.

* Postal address: Department of Statistics and Operations Research, Leonard A. Stern School of Business, New York University, NY 10012, USA.

system are twice as large as those in the single server system. Whenever only one server is busy in the two server system, work is decreasing in that system at only half the rate at which it is decreasing in the single server system. Therefore, the work in the single server system is smaller than half the work in the two server system at every time and on every sample path. To equalize the work in the single server system to half that in the two server system, on all sample paths and at all times, we slow down the rate at which work is being done to half the normal rate in the single server system whenever there is only one customer in the two server system. This increases the time average work in the single server system, and we show that this increase in work can be computed for the $M/M/2$ case and bounded from above for the $M/G/2$ case when the service times have the DMRL or the NBU property. These estimates are then used to compute or bound the mean delay in the two server queue. We now make the work equalization argument precise. We shall call the two server system II. The following definitions are for II:

T_n = time between arrival of the nth and $(n+1)$th customers.

$\{T_n\}$ = i.i.d. sequence with T_1 distributed exponential with parameter λ.

$2S_n$ = service time for the nth customer.

$\{2S_n\}$ = i.i.d. sequence with S_1 having mean $1/\mu$.

We shall write S for S_1. Assume that S has a finite second moment, ES^2.

$\rho = \lambda/\mu$ = system load factor (assumed to be less than one).

B_n = arrival epoch of the nth customer.

$C_n(2)$ = nth departure epoch. By convention, quantities subscripted by zero or negative numbers are set to zero.

$D_n(2)$ = delay in queue of the nth customer, when FCFS scheduling discpline is used.

$d(2)$ = mean delay in queue $= \lim_{n\to\infty} (\Sigma_{i=1}^{n} D_n(2)/n)$.

$W(2, 1, t)$ = larger component of the work vector at time t (under FCFS).

$W(2, 2, t)$ = smaller component of the work vector at time t (under FCFS). Note that we have $D_n(2) = W(2, 2, B_n)$.

$W(2, 1) = \lim_{t\to\infty} \int_0^t W(2, 1, t)dt/t$ and $W(2, 2) = \lim_{t\to\infty} \int_0^t W(2, 2, t)dt/t$, the corresponding time averages of the two components.

The FCFS scheduling discipline will be used in the two server system. It will be assumed that sample paths are continuous from the left, and have right-hand limits. Thus the sample path quantities at arrival instants are what a customer finds on arrival. Consider now a single server queue with the same inter-arrival sequence of customers, but service time sequence $\{S_n\}$. This will be called the *fast server system*. We shall modify the dynamics of the fast server system such that the time average work in the modified system will be exactly half that in II. The modified system will be denoted by I. The systems I and II will be run concurrently. We assume that arrivals occur in pairs, one customer going to I and the other to II. On each arrival using the distribution of S_1 we randomly draw a service time and assign that sample value to be the service time of the customer going to I *and* twice the sampled value to the one going to II. Assume that both systems start off empty. The modifications for I are as follows. Let

$I(0, n)$ = indicator of the event the nth customer on arrival finds system II empty.
$I(1, n)$ = indicator of the event the nth customer on arrival finds a single customer in system II.

These two cases will be called generically cases (0) and (1). Define $C_n(1)$, $D_n(1)$, $d(1)$, and $W(1, 1, t)$ analogous to the quantities defined for system II. Here $W(1, 1, t)$ is the virtual delay in I. Let,

$$\Delta_n = \tfrac{1}{2}(W(2, 1, B_n) - W(2, 2, B_n)) \tag{1}$$
$$= \text{half the difference in work in II found by the } n\text{th arrival.}$$

$$A_n = \tfrac{1}{8} W(2, 1, C_{n-2}(2))^2 I(0, n) \tag{2}$$
$$+ [\tfrac{1}{4} W(2, 1, C_{n-2}(2)) - \tfrac{1}{8}(B_n - C_{n-2}(2))](B_n - C_{n-2}(2)) I(1, n).$$

We shall show that A_n is the extra area to be added to account for the different rates of working when only one server is occupied in II. This area is called the *padded area*. The dynamics of I are modified as follows:

$$D_n(1) = (D_{n-1}(1) + S_{n-1} - T_{n-1})^+ (1 - I(1, n)) + \tfrac{1}{2} I(1, n) W(2, 1, B_n). \tag{3}$$

The recursion (3) for the delay in I is as usual, but with one exception. As long as there is work for both servers in II, the work in I will decrease at the rate of one unit per unit time. However, as given by (3), whenever only one server is busy in II and a customer arrives before that server has completed service, the sample path of work in I will jump to increase the delay. A way of visualizing the dynamics is to imagine that, when only one server is busy in II, the server in I will serve at half the usual rate. This is depicted in Figures 1 and 2, and discussed in Lemma 2. With this modification, whenever an arrival finds a single customer in II, the customer going to I will be delayed by half the work found in system II. This ensures that the sample path of work in I is exactly centered between $W(2, 1, t)$ and $W(2, 2, t)$ when both servers are busy in II. We shall prove these statements below.

Lemma 1. $D_n(1) = D_n(2) + \tfrac{1}{2}(W(2, 1, B_n) - W(2, 2, B_n))$; $n = 1, 2, 3, \cdots$.

Proof. Let the lemma be true for n. If II has only one customer at time B_{n+1}, then the claim is true for $(n+1)$ by (3). Else if II has emptied then so has I by the induction hypothesis. Else,

$$D_{n+1}(1) = D_n(1) + S_n - T_n = D_n(2) + \tfrac{1}{2}(W(2, 1, B_n) - W(2, 2, B_n)) + S_n - T_n$$
$$= \tfrac{1}{2}((W(2, 1, B_n) - T_n) + (W(2, 2, B_n) + 2S_n - T_n))$$
$$= D_{n+1}(2) + \tfrac{1}{2}\Delta_{n+1}.$$

Lemma 2. When both servers are busy in II, $W(1, 1, t) = \tfrac{1}{2}(W(2, 1, t) + W(2, 2, t))$.

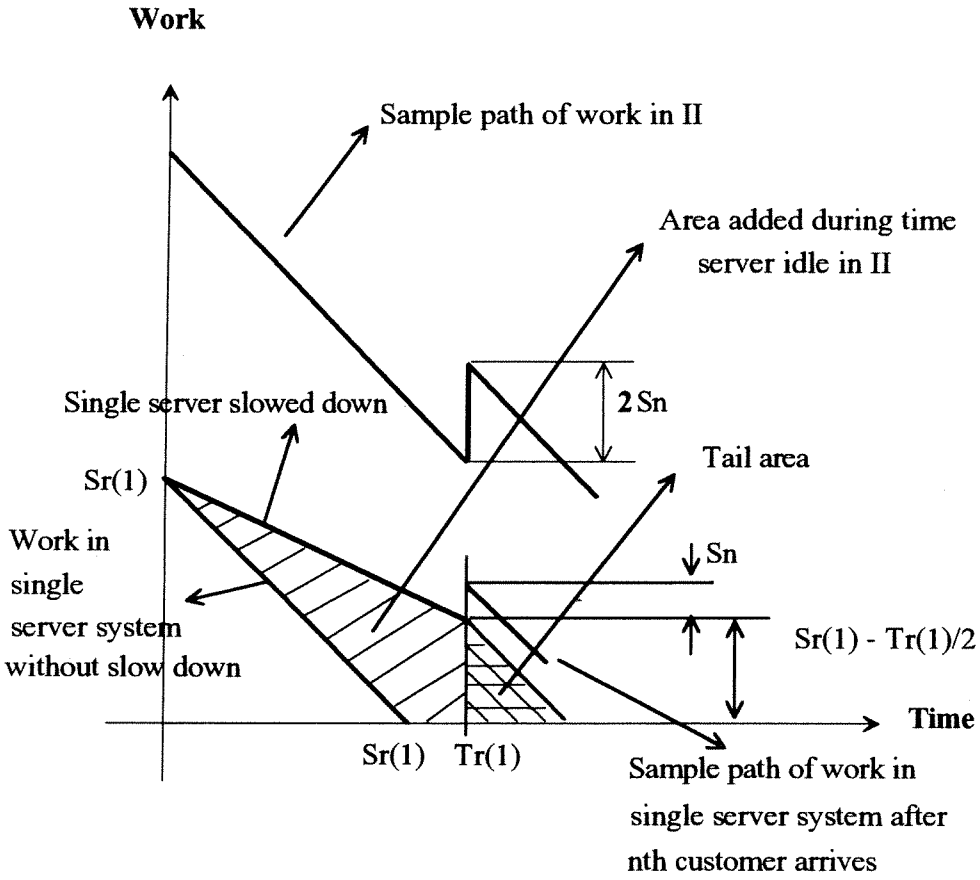

Figure 1. Case 2(a) $T_{R(1)} \geqq S_{R(1)}$

Proof. Consider the time interval between the instant the nth customer enters service in II until the next customer arrives. If both servers are busy in II during this time interval then $(D_n(2)+2S_n-T_n)$ is non-negative. This implies $(D_n(1)+S_n-T_n)$ is non-negative by Lemma 1. Therefore the sample path of work in both systems is a straight line during this interval. Again using Lemma 1, the assertion of Lemma 2 follows.

The *work equalization* construction adds the padded areas as defined in (2) to the nth customer. Let Ψ_n be the instant the nth customer commences service in I. With this padding of areas, we have:

Lemma 3. By adding $\Sigma_{i=1}^{n} A_i$ to the area under the work curve in I we guarantee

$$2\left(\int_0^{\Psi_n} W(1, 1, t)dt + \sum_{i=1}^{n} A_n\right) = \sum_{i=1}^{2} \int_0^{\Psi_n} W(2, i, t)dt.$$

Proof. Because of Lemma 2, we need only be concerned when some server falls idle in II. There are two cases to consider.

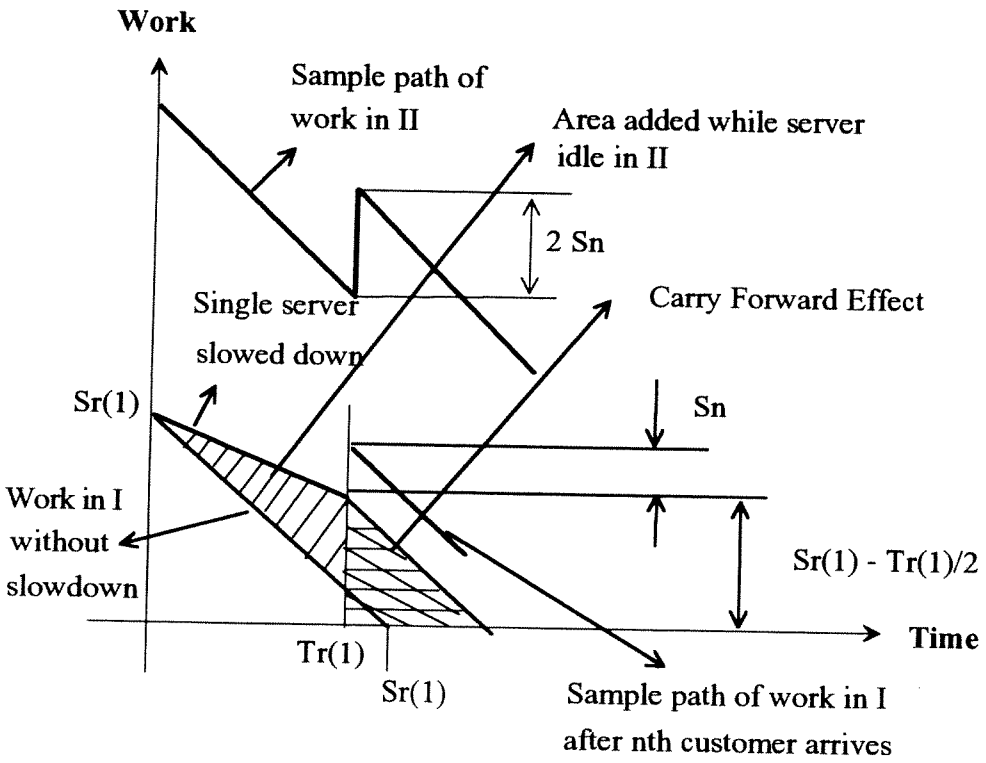

Figure 2. Case 2(b) $T_{R(1)} < S_{R(1)}$

Case (i) is when an arrival finds system II empty. In this case, at time $C_{n-2}(2)$, $W(2, 1, C_{n-2}(2))$ is twice as large as $W(1, 1, C_{n-2}(2))$. So by (2), the padded area is set to $\frac{1}{8}W(2, 1, C_{n-2}(2))^2$.

Case (ii), when only one customer is found by an arrival in II, is more complicated. First, for notational ease, define:

$2S_{R(1)} = W(2, 1, C_{n-2}(2)) =$ the residual work at the last departure instant in system II.
$T_{R(1)} = B_n - C_{n-2}(2) =$ the residual inter-arrival time since that last departure.

There are two subcases to consider, the first when $T_{R(1)}$ is greater than $S_{R(1)}$ and the other when this is reversed. Call these cases 2(a) and 2(b). In the Figures 1 and 2 we assume that the nth customer on arrival finds one customer in II.

Case 2(a). $T_{R(1)} \geqq S_{R(1)}$. Refer to Figure 1. Note that the area padded has a *tail* because the modified recursion in (3) creates extra work that has to be finished. The tail area is termed the *carry forward effect*. The *total area padded* equals the area added to account for the different service rate while only one server is busy in II, called the *area added when the server is idle* plus the tail area. These albeit loose terms help in fixing the cause for the two components. The padded area is equal to

$$\tfrac{1}{4}S_{R(1)}^2+\tfrac{1}{2}(\tfrac{1}{2}S_{R(1)}+S_{R(1)}-\tfrac{1}{2}T_{R(1)})(T_{R(1)}-S_{R(1)})+\underbrace{\tfrac{1}{2}(S_{R(1)}-\tfrac{1}{2}T_{R(1)})^2}_{\text{carry forward effect}}$$

(4)

$$=\tfrac{1}{2}S_{R(1)}T_{R(1)}-\tfrac{1}{8}T_{R(1)}^2.$$

Case 2(b). $T_{R(1)}<S_{R(1)}$. Refer to Figure 2. The area is now given by:

$$\tfrac{1}{4}T_{R(1)}^2+\underbrace{\tfrac{1}{2}[(S_{R(1)}-\tfrac{1}{2}T_{R(1)})^2-(S_{R(1)}-T_{R(1)})^2]}_{\text{carry forward effect}}=\tfrac{1}{2}S_{R(1)}T_{R(1)}-\tfrac{1}{8}T_{R(1)}^2. \tag{5}$$

The time average work in II is given by (see Wolff (1989), for example):

$$W(2,1)+W(2,2)=2(\lambda/\mu)d(2)+2\lambda ES^2. \tag{6}$$

Define $EA=\lim_{n\to\infty}\Sigma_{i=1}^{n}A_i/B_n$, where EA is the unconditional expectation of the padded area added as per (2). To compute the time average 'work' in I (in quotes because it includes the padded area), the nth customer's contribution to 'work' can be written as $S_n(D_n(2)+\Delta_n)+A_n+S_n^2/2$. But S_n and Δ_n are independent. Using this fact and applying $H=\lambda G$ (see Wolff (1989), for example) we can write the time average 'work' in I as

$$\lim_{n\to\infty}\left(\int_0^{B_n}W(1,1,t)dt+\sum_{i=1}^{n}A_n\right)\Big/B_n$$

(7)

$$=\lambda(ES(d(2)+\tfrac{1}{2}(W(2,1)-W(2,2))))+\lambda(EA+ES^2/2).$$

Equating 2 times (7) to (6) by using Lemma 2, and manipulating: $2\lambda ES^2=\lambda ES^2+2\lambda EA+\rho(W(2,1)-W(2,2))$, which gives

$$W(2,1)-W(2,2)=\mu ES^2-2\mu EA. \tag{8}$$

Remark. (8) holds for the *GI/G/2* queue also.

The next step is to derive an expression for $d(2)$ in terms of $d(1)$. Using (6), (7), (8), work equalization and PASTA we obtain:

$$2W(2,2)\underset{\text{PASTA}}{=}2d(2)=2\rho d(2)+2\lambda ES^2-\mu ES^2+2\mu EA,$$

which gives

$$d(2)=\frac{\lambda ES^2}{2(1-\rho)}-\frac{1}{2(1-\rho)}[(\mu-\lambda)ES^2+2\mu EA]$$

(9)

$$=d(1)-\left(\frac{1-\rho}{\rho}d(1)-\mu EA/(1-\rho)\right).$$

Let $p_0=\lim_{n\to\infty}\Sigma_{i=1}^{n}I(i,i)/n$ and $p_1=\lim_{n\to\infty}\Sigma_{i=1}^{n}I(i,2)/n$ be the fraction of arrivals (equal to the fraction of time) that see an empty II and a single customer in II respectively. By equating the fraction of idle servers to the sum of these probabilities we get

$$2p_0+p_1=2(1-\rho). \tag{10}$$

EA can be written, using these probabilities, as

$$EA=p_0E(A_n \mid \text{case}(0))+p_1E(A_n \mid \text{case}(1)). \tag{11}$$

Digression. Let us check (9) for the $M/M/2$ case. In case (ii) the value of S_R can be written as $(\frac{1}{2}T_{R(ii)}+S)$. This allows us to write the area in case (ii) as $\frac{1}{8}T^2_{R(ii)}+\frac{1}{2}ST_{R(ii)}$. Therefore, the compensating factor for the delay in the $M/M/2$ queue can be written using (9) as

$$\begin{aligned}&\frac{-(1-\rho)}{\rho}\frac{\lambda}{2(1-\rho)}\frac{2}{\mu^2}+\frac{\mu}{(1-\rho)}\left[\left(\frac{1}{4(\lambda+\mu/2)^2}+\frac{1}{2\mu(\lambda+\mu/2)}\right)\frac{2\lambda}{\mu}+\frac{1}{4(\lambda+\mu/2)^2}\right]\frac{1-\rho}{1+\rho}\\ &=-\frac{1}{\mu}+\left(\frac{\mu\lambda/2+\lambda(\lambda+\mu/2)+\mu^2/4}{\mu(\lambda+\mu/2)^2}\right)\frac{1}{1+\rho}=-\frac{\rho}{1+\rho}\frac{1}{\mu}\end{aligned} \tag{12}$$

which checks out correctly.

Next we show that the mean delay in queue in the $M/G/2$ system is smaller than that in the $M/G/1$ system when the service time distribution has either the *decreasing mean residual life property* (DMRL) property (i.e. $E(S-t/S>t)$ decreases in t) or the NBU property (i.e. $(S-t/S>t)\leqq_{\text{st}} S$). Define $S_{R(0)}$ and $T_{R(0)}$, analogous to the definitions used for case (1). Consider the terms other than the delay in the $M/G/1$ system in (9). After simplification, we need to show that

$$-\tfrac{1}{2}(1-\rho)ES^2+(\tfrac{1}{2}p_1ES_{R(1)}T_{R(1)}+\tfrac{1}{2}p_0ES^2_{R(0)}-\tfrac{1}{8}p_1ET^2_{R(0)})\leqq 0. \tag{13}$$

For this purpose a lemma is required.

Lemma 4. Let X be a non-negative random variable with finite second moment and distribution function G. Let Y be a random variable independent of X and distributed exponential with parameter a. Then $E[(XY-\frac{1}{4}Y^2)I\{2X\geqq Y\}]+\frac{1}{2}E[X^2I\{2X<Y\}]\leqq\frac{1}{2}EX^2$.

Proof. After manipulation we must show that $E[((X-Y)^2-\frac{1}{2}Y^2)I\{2X\geqq Y\}]\geqq 0$. However standard integration yields

$$\begin{aligned}&\int_0^{2x}[(x-y)^2-\tfrac{1}{2}y^2]ae^{-ay}dy\\ &=(1-e^{-2ax})(x^2-2x/a+2/a^2)-[-2x^2e^{-2ax}-2x/ae^{-2ax}+1/a^2(1-e^{-2ax})]\\ &\geqq(1-e^{-2ax})(x-1/a)^2\geqq 0.\end{aligned}$$

Proposition 1. The mean delay in the queue in the $M/G/2$ system is smaller than that in the equivalent fast single server system under DMRL or NBU service times.

Proof. Jointly condition on whether case (0) or (1) occurs, i.e. isolate the last customer in service in II. To be specific, if case (0) or (1) occurred when customer n arrived, then

the last customer in service in II is the one who departs at time $C_{n-1}(2)$. Denote the residual service time of this customer when a server in system II was last idle as $2S_R$. To be specific, in the example above this is $W(2, 1, C_{n-2})$. Let T_R (which is independent of S_R) be distributed exponentially with parameter λ. Using Lemma 4, we obtain

$$E((\tfrac{1}{2}S_R T_R - \tfrac{1}{8}T_R^2)I(T_R < 2S_R) + \tfrac{1}{4}S_R^2 I(T_R \geqq 2S_R)) \leqq \tfrac{1}{4}ES_R^2. \tag{14}$$

Using the DMRL property and Proposition 1.6.1 in Stoyan (1983), or by definition of NBU, we can bound the right-hand side of (14) by $\frac{1}{4}ES^2$. Similarly

$$E(\tfrac{1}{4}S_R^2 \mid T_R \geqq 2S_R) \leqq \tfrac{1}{4}ES^2. \tag{15}$$

Using (14) and the fact that $P(T_R \geqq 2S_R)$ and $P(T_R < 2S_R)$ are in the ratio of p_0 and p_1 gives

$$(\tfrac{1}{2}p_1 ES_{R(1)}T_{R(1)} + \tfrac{1}{4}p_0 ES_{R(0)}^2 - \tfrac{1}{8}p_1 ET_{R(1)}^2) \leqq \tfrac{1}{4}(p_0 + p_1)ES^2.$$

Unconditioning (15) gives: $\frac{1}{4}p_0 ES_{R(0)}^2 \leqq \frac{1}{4}p_0 ES^2$. Adding these two inequalities and using (10) gives (13).

3. The $M/G/C$ system

In this section we use the work equalization construction to equate the work in the standard $M/G/C$ queue with that in a single server queue. The following definitions are required for the C server system:

$\{T_n\}$ = i.i.d. sequence with T_1 distributed exponential with parameter λ.

CS_n = service time for the nth customer.

$\{CS_n\}$ = i.i.d. sequence with S_1 having mean $1/\mu$.

$\lambda/\mu = \rho$ = system load factor (assumed to be less than one).

B_n = arrival epoch of the nth customer.

$C_n(C)$ = nth departure epoch. By convention quantities subscripted by zero or negative numbers are set to zero.

$D_n(C)$ = delay in queue of the nth customer, when FCFS scheduling discipline is used.

$d(C)$ = mean delay in queue.

$W(C, i, t)$ = ith largest component of the work vector at time t (under FCFS), $i = 1, 2, \cdots, C$.

$W(C, C, t)$ = virtual delay.

$V(C, t)$ = total work in the system at time t.

$W(C, i)$ = time average of $W(C, i, t)$, $i = 1, 2, \cdots, C$.

$V(C) = \Sigma_{i=1}^{C} W(C, i)$ = time average work in the system.

$W^P(1, t)$ = modified work at time t in the single server system.

We shall be using a construction similar to the one used for the two server case to equate the time average work in the C server system with that in a modified single server system. However, instead of the detailed construction and justification (as in Lemmas 1 to 3) we shall only indicate the main steps below. Let Z be an instant such that at time Z^- either $(k+1)$ or $(k-1)$ servers were busy, and at time Z exactly k servers are busy. This

event can happen either due to a service completion or the arrival of a customer. Here k can take values between 1 and $(C-1)$. Each such value will constitute a case. Assume that at time Z, $W^P(1, Z)=V(C, Z)/C$. Let T_r be the residual inter-arrival time at time Z. For easier notation we shall let $X=\min[T_r, W(C, k, Z)]$. At time $(Z+X)$ the next event, either a customer arrival or departure occurs. For equating the work in the single and C server systems (and for continuing this construction) we shall be setting $W^P(1, (Z+X)^-)=\Sigma_{i=1}^{k}(W(C, i, Z)-X)/C$. This (as in the two server system) will add an extra area under the work curve of the single server system, which will be denoted by $A(k)$. To compute $A(k)$ we need to consider two cases.

Case (i). $V(C, Z)/C \geqq X$:

$$A(k)=\frac{1}{C}\left(\sum_{i=1}^{k}(W(C, i, Z)-\tfrac{1}{2}X)X\right)-(V(C, Z)/C-\tfrac{1}{2}X)X$$

$$+\frac{1}{2}\left[\frac{1}{C}\left(\sum_{i=1}^{k}W(C, i, Z)-X\right)\right]^2-\tfrac{1}{2}(V(C, Z)/C-X)^2 \tag{16}$$

$$=\frac{C-k}{C^2}(V(C, Z)X-\tfrac{1}{2}kX^2).$$

Case (ii). $V(C, Z)/C<X$:

$$A(k)=\frac{1}{C}\left(\sum_{i=1}^{k}(W(C, i, Z)-\tfrac{1}{2}X)X\right)-\tfrac{1}{2}V(C, Z)^2/C^2$$

$$+\frac{1}{2}\left[\frac{1}{C}\left(\sum_{i=1}^{k}W(C, i, Z)-X\right)\right]^2 \tag{17}$$

$$=\frac{C-k}{C^2}(V(C, Z)X-\tfrac{1}{2}kX^2).$$

We shall use $EA(k)$ to denote the conditional expectation of the area $A(k)$ given that case k has occurred. Note that, as found in the two server case, (16) equals (17). Moreover, the expression in parentheses in (16) is the area under the work curve during the time $[Z, Z+X]$ in the C server system. This interpretation will be used in the following. Let,

$I(k, i, n)=$ the indicator of the event that case (k) happened, and the next customer to arrive was customer n who saw i customers in the system. Here $i=0, 1, 2, \cdots, k$.

Then the expected value of the padded area EA_n attached to customer n can be written as

$$EA_n=\sum_{i=1}^{C-1}\sum_{k=i}^{C-1}E(A(k)I(k, i, n))+\sum_{k=1}^{C-1}E(A(k)I(k, 0, n)). \tag{18}$$

Also define the time averaged (over the whole time line) area under the work curve during the time when there are 1 *through* $C-1$ customers in the C server system, as

$$A_I = \lambda \sum_{j=1}^{C-1} \beta_j EA(j) \left(\frac{C^2}{C-j} \right) \tag{19}$$

where, $\beta_j = \lim_{N \to \infty} \Sigma_{n=1}^{N} \Sigma_{i=0}^{j} I(j, i, n)/N$ and the subscript I is used to denote that this is the area when some server is idle. Note that the equality in (19) is from an application of $H = \lambda G$, because the expression on the right-hand side of (19) is λ times the customer averaged area under the work curve as seen by an arriving customer when there are 1 *through* $C-1$ customers in the system. It follows from (18) and (19) that

$$\left(\frac{C-1}{C^2} \right) A_I \geqq \lambda EA_n. \tag{20}$$

By the work equalization argument (or by a constructive proof) we have that the mean delay in the modified single server system is $V(C)/C$, and the average contribution to the area under the work curve by a customer is $(\frac{1}{2}ES^2 + EA_n)$. Therefore, using $H = \lambda G$,

$$\begin{aligned} V(C)/C &= \rho V(C)/C + \lambda(\tfrac{1}{2}ES^2 + EA_n) \\ &\Leftrightarrow V(C)/C = d_1 + \lambda EA_n/(1-\rho). \end{aligned} \tag{21}$$

Let the time averaged (over the whole time line) area under the work curve in the C server system when all C servers are busy be A_B, where the subscript B stands for busy. We have

$$A_I + A_B = V(C). \tag{22}$$

We also know that

$$V(C)/C = \rho d_C + \tfrac{1}{2} C \lambda ES^2. \tag{23}$$

If A_I is less than $\frac{1}{2}C^2\lambda ES^2(1-\rho)$ then from (23), i.e., the expression for $V(C)/C$, (20) and (21):

$$\begin{aligned} V(C)/C &= \rho d_C + \tfrac{1}{2} C \lambda ES^2 < d_1 + \tfrac{1}{2}(C-1)\lambda ES^2 \\ &\Leftrightarrow d_C < d_1. \end{aligned}$$

On the other hand, if $A_I \geqq C^2\lambda ES^2(1-\rho)/2$, then from (22) and (23):

$$Cd_C < A_B \leqq C\rho d_C + \tfrac{1}{2}\rho C^2 \lambda ES^2 \Rightarrow d_C < (C\rho) d_1.$$

The first inequality follows from the fact that the total work in the system in the C server system when all C servers are busy, when time averaged over the entire time line, is surely larger than Cd_C. Thus we have:

Proposition 2. For the $M/G/C$ system when the traffic intensity is less than or equal to $1/C$, the mean delay in queue d_C is less than that in the $M/G/1$ system operating under the same load.

4. Conclusions

We have shown that the mean delay for the $M/G/2$ queue is smaller than that for the fast server $M/G/1$ queue under the FCFS scheduling discipline and either DMRL or NBU service time distribution. For the general $M/G/C$ system we have a limited result for traffic intensities smaller than $1/C$, which strengthens the bound of Suzuki and Yoshida (1970). We intend to extend this result using the work equalization argument to show that, under the same load, the mean delay is decreasing in the number of servers.

Acknowledgments

I am grateful to Professor Wolff for introducing me to the problem, for several helpful discussions and for pointing out that the proof in Proposition 1 goes through for the NBU case. Professor Shanthikumar helped me in writing out the precise formulation of the work equalization argument, and I am grateful to him for several discussions on earlier versions of this paper. The comments by an anonymous referee have helped a lot in improving the exposition. This work was partly supported by grants to the University of California, Berkeley, from the California State MICRO program and the Semiconductor Research Corporation.

References

Cox, D. R. and Smith, W. L. (1961) *Queues.* Wiley, New York.

Daley, D. J. and Rolski, T. (1984) Some comparability results for waiting times in single- and many-server queues. *J. Appl. Prob.* **21**, 887–900.

Mori, M. (1975) Some bounds for queues. *J. Operat. Res. Soc. Japan* **18**, 151–181.

Stoyan, D. (1983) *Comparison Methods for Queues and Other Stochastic Models.* Wiley, New York.

Suzuki, T. and Yoshida, Y. (1970) Inequalities for the many-server queue and other queues. *J. Operat. Res. Soc. Japan* **13**, 59–77.

Wolff, R. W. (1989) *Stochastic Modeling and the Theory of Queues.* Prentice Hall, New York.