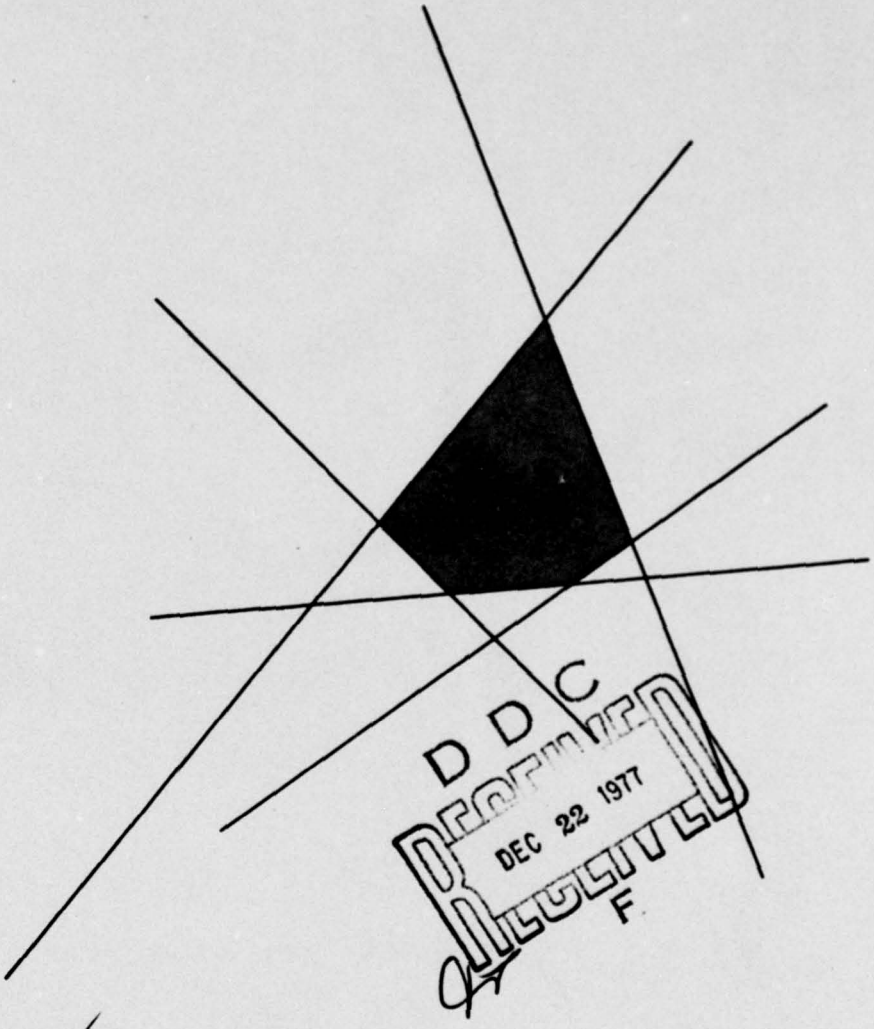
AD-A047 951 CALIFORNIA UNIV BERKELEY OPERATIONS RESEARCH CENTER F/G 12/1
BOUNDS AND COMPARISONS FOR SOME QUEUEING SYSTEMS.(U)
NOV 77 S NIU N00014-77-C-0299
UNCLASSIFIED ORC-77-32 NL

END
DATE FILMED
1-78
DDC

ORC 77-32
NOVEMBER 1977

BOUNDS AND COMPARISONS FOR SOME QUEUEING SYSTEMS

by
SHUN-CHEN NIU



OPERATIONS RESEARCH CENTER

UNIVERSITY OF CALIFORNIA · BERKELEY

BOUNDS AND COMPARISONS FOR SOME QUEUEING SYSTEMS

by

Shun-Chen Niu
Operations Research Center
University of California, Berkeley

NOVEMBER 1977 ORC 77-32

This research has been partially supported by the Office of Naval Research under Contract N00014-77-C-0299 and the Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-77-3213 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 77-32	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: BOUNDS AND COMPARISONS FOR SOME QUEUEING SYSTEMS.	5. TYPE OF REPORT & PERIOD COVERED: Research Report,
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Shun-Chen Niu	8. CONTRACT OR GRANT NUMBER(s): NØØØ14-77-C-ØØ29, AFOSR-77-3213
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Operations Research Center, University of California, Berkeley, California 94720	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2304/A5 61102F
11. CONTROLLING OFFICE NAME AND ADDRESS: United States Air Force, Air Force Office of Scientific Research, Bolling AFB, D.C. 20332	12. REPORT DATE: November 1977
	13. NUMBER OF PAGES: 68
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: 69 p.	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)*

Tandem Queues
Total Waiting Times
Delays
Bounds
Comparisons
Loss Models

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

(SEE ABSTRACT)

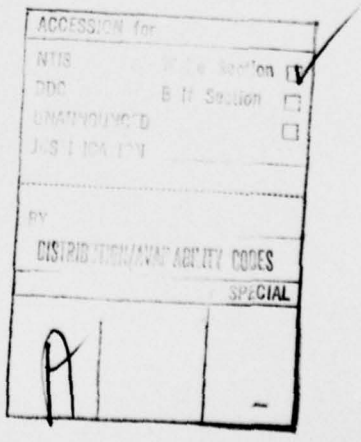
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

270 750 AB

DEDICATION

To my mother and my wife.



ACKNOWLEDGEMENTS

I would like to sincerely thank my thesis advisor, Professor Sheldon M. Ross, for his supervision and encouragement throughout this work. I would also like to thank Professor Ronald W. Wolff for many helpful discussions and Professor Gordon F. Newell for serving on my committee.

ABSTRACT

Bounds and comparisons for various performance measures in tandem queueing systems and loss systems are studied.

An upper bound for the expected delay in front of the second server is found for a sequence of two queues in tandem where the first server has deterministic service times, the second server has general service distribution, and the arrival process is an arbitrary renewal process. The result is extended to the case of n queues in tandem where all the servers except the last one have constant service times.

Using a definition of variability of random variables, it is proven that for a tandem queueing system with n stations in series, where each station can have either one server with an arbitrary service distribution or a number of constant servers in parallel, the expected total waiting time in system of every customer decreases as the variability of interarrival and service distributions decreases. A new sufficient condition for customer average delay to be smaller (larger) than time average delay in single server queues is also given.

A heterogeneous arrival single server queueing loss model is also analyzed, where the arrival process is a nonstationary Poisson process with an intensity function whose evolution is governed by a two-state continuous time Markov chain. The explicit loss formula for this model is obtained. In a special case, it is shown that as the arrival process becomes more stationary the loss decreases. For single server loss systems with renewal arrivals, counterexamples are given to show that regularity of arrival and service distributions do not work to good effect in general. Two sufficient conditions for it to be true are given.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This thesis is concerned with two types of problems: bounds for expected delays in some tandem queueing systems which are very difficult to analyze exactly in general and the effect of regularity of arrival process and service times on the performance of some queueing systems.

A tandem queue is a number of service facilities in series. Customers arrive according to a renewal process. Upon arrival, each customer goes to the first station and requires a random amount of service time. If there were other customers present at the time of his arrival, he joins the end of the queue and waits for service. The order of service of customers is first come first served. After being served at the first station, he goes to the second station. Each station operates in a similar fashion but may provide different type of service in general. Every customer has to go through all stations according to a prefixed order and leaves the system after finishing service at the last station.

Most known results in queueing theory are for queues with renewal arrivals. The fact that, in general, the output process of a queueing station is not a renewal process makes tandem queues very difficult to analyze.

A number of authors have tried to analyze the output process of a queueing system with the hope that it might be helpful in analyzing the behavior of a subsequent station.

Burke [5], [6] showed that the output of a stationary M/M/s queue is a Poisson process with the same rate as the arrival process. Hence in a tandem queue with Poisson arrival and exponential service times, the waiting times at each station can be computed by solving a standard M/M/1 queue. Furthermore, it has been shown by Reich [16], [17] that the waiting times of a customer at different stations are independent of each other.

For a tandem queueing system where all the servers have constant service times except maybe one whose service times are random and larger than that of others with probability one, Friedman [8] showed that, for any arbitrary arrival process, the epoch at which each customer departs the system is independent of the order of servers.

If we are allowed to choose the order of servers in tandem queues where the service times for different servers are nonoverlapping, Tembe and Wolff [23], [24] showed that the total waiting time in system is stochastically minimal when the servers are in decreasing order of their service times. They also proved that, for two queues in tandem with one constant server and one arbitrary server, it is better to put the constant server in the first station for any arbitrary arrival process.

Except in these instances, few useful results are known concerning the waiting times in tandem queueing systems.

In view of the difficulties with analyzing tandem queues, it is natural to look for some approximations and bounds for various quantities of interest in these models. In Chapter 2 and 3, some results in this direction are presented.

Tembe and Wolff [24] obtained an upper bound for the expected waiting time for a system of two queues in tandem where the first server has constant service times, the second server has exponential service times, and the arrival process to the system is Poisson.

In Chapter 2, we study a more general system with a deterministic first server but the second server has an arbitrary service time distribution and the arrival process is a general renewal process. Motivated by Tembe and Wolff's bound, a conjecture is made to the effect that, for a fixed arrival process, the expected delay in front of the second server is a decreasing function of the magnitude of the constant service time at the first station.

If the conjecture is true, then, letting the service time at the first station go to zero, we have that the expected delay in front of the second server is bounded above by the corresponding one for a system without the first server or just a conventional GI/G/1 queue.

Well-known upper bounds for the expected delay in GI/G/1 queue are readily available. Hence the conjecture will give upper bound for the expected delay in front of the second server in the above system.

We have been unable to verify the conjecture at this time. However the upper bound which can be obtained in this fashion is shown to be valid.

The above result can be generalized to find bounds for expected delays in tandem queues with n stations $(n \geq 2)$ in series where all servers except the last server whose service distribution may be arbitrary have constant service times and the arrival process is general renewal.

As a by-product of the analysis, we also give some simple bounds for expected idle periods, variance of interoutput times, and some correlation relations among various quantities in some tandem queueing systems.

In Chapter 3, we investigate the effect of regularity of the arrival process and service times on the waiting time of a customer in some tandem queueing systems.

The following "conventional wisdom" is generally believed to be true for many queueing systems: The more regular the arrival process and the service times are, the better the system performance will be.

Wolff [25] gave a summary of known results in the literature which are in support of the conventional wisdom and also presented a number of systems which exhibit contrary behaviors.

Stoyan [18], [21] proved a theorem on the comparison of delays in GI/G/1 queues. He gave a definition of variability which is very useful in this context, and proved that, in a GI/G/1 queue, as the interarrival times and service times become more regular, the delay decreases. In this thesis, when we say "more regular," it will mean more regular in Stoyan's sense in most cases.

His result can readily be applied to give a new sufficient condition for customer average delay to be smaller (or larger) than time average delay in a GI/G/1 queue. Improvements of some known bounds by Marshall [13] for the expected delay in certain classes of GI/G/1 queues also follow easily.

The main results in this chapter are extensions of Stoyan's theorem to some tandem queueing systems. It is shown that for a tandem system

with n stations in series, where each station can have either one server with an arbitrary service distribution or a number of constant servers in parallel, the expected total waiting time of every customer decreases as the interarrival and service times become more regular.

The results are quite useful for bounding purposes since we can bound the expected total waiting time for systems which are difficult to analyze by the corresponding quantity for easier ones. Two examples of this sort are given.

Under stronger assumptions on the arrival process and service times, stronger comparison results can also be obtained.

Specifically, consider a sequence of two queues in tandem. For two such systems where the servers at the first station have the same distribution for both systems, it is proven that the system with stochastically larger interarrival times and stochastically smaller (more regular) service times at the second station has stochastically smaller (more regular) delay for every customer in front of the second station.

In Chapter 4, the effects of regularity on some single server queueing loss models are examined.

Fond and Ross [22] considered a single server exponential queueing loss model where the arrival and service rates alternate between two phases (λ_1,μ_1) and (λ_2,μ_2). The amounts of time the system spends in each phase have exponential distributions with rate $c\alpha_1$ and $c\alpha_2$ respectively. All arrivals who find the system busy are lost. By solving balance equations, they showed that the proportion of customers lost is a decreasing function of c, i.e., the more stationary the arrival process is, the smaller the proportion of customers lost will be.

CHAPTER 2

BOUNDS FOR SOME TANDEM QUEUEING SYSTEMS

2.1 Definitions and Notations

The first system we will consider is a system of two queues in tandem. We will denote such a system by $GI/G_1/1 \rightarrow G_2/1$ where G_i, $i = 1,2$ is the distribution of the service times at station i and GI means that the arrival process is a general renewal process. Similar notations will be used throughout, e.g., $GI/D/1 \rightarrow G/1$, $M/D/1 \rightarrow M/1 \rightarrow G/1$ where D, M represent constant and exponential service or interarrival times respectively.

By "delay" we shall always mean the amount of time a customer spends waiting in queue in front of a service station. The waiting time of a customer is equal to his delay plus service time.

Unless otherwise specified, it is assumed that the queueing process starts at time zero with all stations empty. Hence the delay of the first customer will be zero.

Let the subscript n (e.g., D_n) refer to the n^{th} customer. The following notations will be used:

T_n = the interarrival time between the n^{th} and $(n+1)^{th}$ arrival. $E(T_n) = \frac{1}{\lambda}$, $\text{Var } T_n = \sigma_T^2 < \infty$. T_n, $n = 1,2, \ldots$, are i.i.d. random variables.

S_n = service time of the n^{th} customer at the first station. $E(S_n) = \frac{1}{\mu_1}$, $\text{Var } S_n = \sigma_S^2 < \infty$. S_n, $n = 1,2, \ldots$, are i.i.d. random variables.

R_n = service time of the n^{th} customer at the second station.

$E(R_n) = \frac{1}{\mu_2}$, Var $R_n = \sigma_R^2 < \infty$. S_n , $n = 1,2, \ldots$, are i.i.d. random variables.

D_n = delay in queue of the n^{th} customer in front of the first station.

D_n^* = delay in queue of the n^{th} customer in front of the second station.

$\rho_1 = \lambda/\mu_1$, $\lambda < \mu_1$.

$\rho_2 = \lambda/\mu_2$, $\lambda < \mu_2$.

$W_n = D_n + S_n$.

$W_n^* = D_n^* + R_n$.

2.2 A Conjecture

Tembe and Wolff [24] proved, among other things, a useful result concerning the optimal order of servers in tandem queues: for a tandem queue with two servers in series, if one of the server has constant service time, the total waiting time in system for every customer is stochastically smaller when the constant server is in the first station. The conclusion is true for any arbitrary arrival process.

By reversing the order of servers, the above was used to find an upper bound for the expected stationary waiting time, $E(W^*)$, in front of the second server in a $M/D/1 \rightarrow M/1$ system. Specifically, they obtained

$$E(W^*) \leq \frac{1}{\mu_2 - \lambda} . \tag{2.2.1}$$

Notice that the right-hand side is just the expected waiting time of a M/M/1 queue. The intuitive explanation of this is that the constant server at the first station tends to make the arrival process to the second station more regular than it would be without the first server.

It is felt that the validity of the above explanation should not depend on the distributions of the arrival process and service time at the second station. Hence, we make the following

Conjecture:

In a GI/D/1 → G/1 system, the expected delay in front of the second server is smaller than it would be if there were no first server at all.

In fact we believe a stronger conjecture should also be true, namely, the larger the constant service time at the first station, the smaller the expected delay in front of the second server will be. Thus, the first conjecture follows by letting the service time at the first station go to zero.

If the conjecture were true, then it follows from well-known upper bound for the expected delay in GI/G/1 queues by Kingman [9], [10] and Marshall [13] that

$$E\left(D_\infty^*\right) \leq \frac{\lambda\left(\sigma_T^2 + \sigma_R^2\right)}{2(1 - \rho_2)} \tag{2.2.2}$$

where D_∞^* denotes the stationary delay of a customer in front of the second station.

We have been unable to verify the conjecture at this time. However, it is proven in the next section that (2.2.2) is indeed valid.

2.3 Bounds for the Expected Delay in Front of the Second Server

It is well-known that

$$D_{n+1} = \max [0, D_n + S_n - T_n] \tag{2.3.1}$$

or equivalently,

$$D_{n+1} - X_n = D_n + S_n - T_n \tag{2.3.2}$$

where $X_n = -\min [0, D_n + S_n - T_n]$.

We will first give relations similar to (2.3.1) and (2.3.2) and then use them to prove (2.2.2) by an approach parallel to Kingman and Marshall's.

To start with, we do not assume that the service times of the first server are deterministic. Let e_n be the epoch at which the n^{th} customer arrives. Observe that $e_n + D_n + S_n + D_n^* + R_n$ is the epoch he leaves the system and $e_n + T_n + D_{n+1} + S_{n+1}$ is the epoch the $(n+1)^{th}$ customer departs the first station. Thus

$$D_{n+1}^* = \begin{cases} D_n + S_n + D_n^* + R_n - T_n - D_{n+1} - S_{n+1}, & \text{if } D_n + S_n + D_n^* + R_n > T_n + D_{n+1} + S_{n+1} \\ 0 \quad \text{otherwise.} \end{cases}$$

Equivalently,

$$D_{n+1}^* = \max \left[0, D_n + S_n + D_n^* + R_n - T_n - D_{n+1} - S_{n+1}\right] . \tag{2.3.3}$$

Let $Y_n = -\min \left[0, D_n + S_n + D_n^* + R_n - T_n - D_{n+1} - S_{n+1}\right]$, we have

$$(2.3.4) \quad D^*_{n+1} - Y_n = D_n + S_n + D^*_n + R_n - T_n - D_{n+1} - S_{n+1} .$$

In the derivations below, we will assume stationarity of the processes $\{D_n, n \geq 1\}$ and $\left\{D^*_n, n \geq 1\right\}$.

Taking expectations in (2.3.4), we have

$$E(Y_n) = E(T_n) - E(R_n)$$

$$(2.3.5) \qquad = \frac{1}{\lambda}(1 - \rho_2) .$$

Note that (2.3.5) can be used to find a lower bound for the expected idle period, $E(I_2)$, at the second station by using $E(Y_n) = a_0 E(I_2) \leq E(I_2)$ where a_0 = the probability that a stationary customer departing from the first station finds the second station empty.

Let $B_n = R_n - T_n - S_{n+1}$, (2.3.4) can be written as

$$\left(D_{n+1} + D^*_{n+1}\right) - Y_n = \left(D_n + D^*_n\right) + S_n + B_n .$$

Squaring both sides and noting that $D^*_{n+1} Y_n \equiv 0$ gives

$$\left(D_{n+1} + D^*_{n+1}\right)^2 - 2Y_n D_{n+1} + Y_n^2 = \left(D_n + D^*_n\right)^2 + S_n^2 + B_n^2$$
$$+ 2\left(D_n + D^*_n\right)S_n + 2\left(D_n + D^*_n\right)B_n + 2S_n B_n .$$

Note that the pairs D_n and S_n, D_n and B_n, D^*_n and B_n, S_n and B_n are independent. After taking expectations and cancelling, we get

$$E\left(Y_n^2\right) - 2E(Y_n D_{n+1}) = E\left(S_n^2\right) + E\left(B_n^2\right) + 2E(D_n)E(S_n)$$

$$+ 2E\left(D_n^* S_n\right) + 2E(D_n)E(B_n) + 2E(S_n)E(B_n) \quad . \tag{2.3.6}$$

Substituting $E(Y_n D_{n+1}) = E(Y_n)E(D_{n+1}) + \text{Cov}\,[Y_n, D_{n+1}]$, $E\left(S_n D_n^*\right) = E(S_n)E\left(D_n^*\right) + \text{Cov}\left[S_n, D_n^*\right]$ and noting that $E(S_n) + E(B_n) = -E(Y_n)$, (2.3.6) becomes

$$E\left(Y_n^2\right) - 2\ \text{Cov}\,[Y_n, D_{n+1}] = E\left(S_n^2\right) + E\left(B_n^2\right) - 2E\left(D_n^*\right)E(Y_n)$$

$$+ 2\ \text{Cov}\left[S_n, D_n^*\right] + 2E(S_n)E(B_n) \quad .$$

After some rearrangements, we have proven the following

Theorem 2.3.1:

Under stationary conditions,

$$E\left(D_n^*\right) = \frac{\lambda}{2(1 - \rho_2)} \left\{ 2\sigma_S^2 + \sigma_R^2 + \sigma_T^2 - \text{Var}\ Y_n \right.$$

$$\left. + 2\ \text{Cov}\,[Y_n, D_{n+1}] + 2\ \text{Cov}\left[S_n, D_n^*\right] \right\} . \tag{2.3.7}$$

Remark:

In the derivations of (2.3.5) and (2.3.6), we needed the conditions that $E\left(D_n^*\right) < \infty$ and $E\left(D_n^{*2}\right) < \infty$ in order to make cancellations. It is assumed that these cancellations are valid. In practice, this assumption is usually justified.

The unknowns in (2.3.7) are $\text{Var}\ Y_n$, $\text{Cov}\,[Y_n, D_{n+1}]$, and $\text{Cov}\left[S_n, D_n^*\right]$. However, we always have $\text{Var}\ Y_n \geq 0$ and, in the case

of constant first server, $\text{Cov}\left[S_n, D_n^*\right] \equiv 0$. Hence, we only have to find an upper bound for $\text{Cov}\,[Y_n, D_{n+1}]$. In order to do that, the concept of association of random variables is used. It's definition and basic properties are given in Esary, Proschan and Walkup [7]. For convenience, a summary is given here:

Definition:

Random variables $V_1, \ldots, V_n$ are "associated" if $\text{Cov}\,[\Gamma(\underline{V}), \Delta(\underline{V})] \geq 0$ for all pairs of increasing binary functions Γ , Δ . We shall also say the vector $\underline{V} = (V_1, \ldots, V_n)$ is associated.

Associated random variables have the following properties:

(P1): Any subset of associated random variables are associated.

(P2): The set consisting of a single random variable is associated.

(P3): Increasing functions of associated random variables are associated.

(P4): If two sets of associated random variables are independent of each other, then their union is a set of associated random variables.

(P5): Independent random variables are associated.
(Note: this follows immediately from P2 and P4).

(P6): If random variables $V_1, \ldots, V_n$ are associated, then $\text{Cov}\,[f(\underline{V}), g(\underline{V})] \geq 0$ for all increasing functions f and g .

Before proving the next theorem, we need the following

Lemma 2.3.2:

$$(2.3.8)\quad D^*_{n+1} = \max\left[0,\ \min\left[D_n + D^*_n + S_n + R_n - T_n - S_{n+1}, D^*_n + R_n - S_{n+1}\right]\right],$$

$$(2.3.9)\quad Y_n = -\min\left[0, D_n + S_n + D^*_n + R_n - T_n - S_{n+1}, D^*_n + R_n - S_{n+1}\right],$$

and

$$(2.3.10)\qquad D^*_{n+1} - Y_n = -X_n + D^*_n + R_n - S_{n+1}\ .$$

Proof:

Fix a point ω in the sample space of $T_1, \ldots, T_n, S_1, \ldots, S_{n+1}$, and $R_1, \ldots, R_n$. Then, from (2.3.3),

$$\begin{aligned}
D^*_{n+1}(\omega) &= \max\left[0, D_n(\omega) + S_n(\omega) + D^*_n(\omega) + R_n(\omega) - T_n(\omega) - D_{n+1}(\omega) - S_{n+1}(\omega)\right] \\
&= \max\Big[0, D_n(\omega) + S_n(\omega) + D^*_n(\omega) + R_n(\omega) - T_n(\omega) \\
&\qquad - \max\left[0, D_n(\omega) + S_n(\omega) - T_n(\omega)\right] - S_{n+1}(\omega)\Big] \\
&= \max\Big[0, D_n(\omega) + S_n(\omega) + D^*_n(\omega) + R_n(\omega) - T_n(\omega) \\
&\qquad + \min\left[0, -D_n(\omega) - S_n(\omega) + T_n(\omega)\right] - S_{n+1}(\omega)\Big] \\
&= \max\Big[0,\ \min\Big[D_n(\omega) + S_n(\omega) + D^*_n(\omega) + R_n(\omega) - T_n(\omega) - S_{n+1}(\omega)\ , \\
&\qquad D^*_n(\omega) + R_n(\omega) - S_{n+1}(\omega)\Big]\Big]\ .
\end{aligned}$$

Since ω was arbitrary, we have

$$D^*_{n+1} = \max\left[0, \min\left[D_n + S_n + D^*_n + R_n - T_n - S_{n+1}, D^*_n + R_n - S_{n+1}\right]\right]$$

everywhere.

The proofs of (2.3.9) and (2.3.10) are similar.■

Unless otherwise stated, we will assume from here on in this section that $S_n \equiv s$, $\forall n \geq 1$, where s is a constant, i.e., the first server has constant service time s . In this case, (2.3.8) and (2.3.9) reduce to

$$(2.3.11)\quad D^*_{n+1} = \max\left[0, \min\left[D_n + D^*_n + R_n - T_n, D^*_n + R_n - s\right]\right] .$$

$$(2.3.12)\quad Y_n = -\min\left[0, D_n + D^*_n + R_n - T_n, D^*_n + R_n - s\right] .$$

We proceed to the following important

Theorem 2.3.3:

For all $n \geq 1$, D_n and D^*_n are associated random variables.

Proof:

The proof is by induction.

Obviously, D_1 and D^*_1 are associated since $D_1 = D^*_1 = 0$.

Assume D_n and D^*_n are associated. It is clear from (2.3.1) and (2.3.11) that both D_{n+1} and D^*_{n+1} are increasing functions of the vector $\left(D_n, D^*_n, -T_n, R_n\right)$. Now, $-T_n$ and R_n are independent of each other and $\left(D_n, D^*_n\right)$. Therefore, it follows by induction hypothesis and P4 that $\left(D_n, D^*_n, -T_n, R_n\right)$ is associated. Hence, by P3, D_{n+1} and D^*_{n+1} are associated.■

Lemma 2.3.4:

For all $n \geq 1$, $-Y_n$ and D_{n+1} are associated random variables.

Proof:

It follows from (2.3.1) and (2.3.12) that both D_{n+1} and $-Y_n$ are increasing functions of the vector $\left(D_n, D_n^*, -T_n, R_n\right)$. But, by Theorem 2.3.3, P4, and P5, $\left(D_n, D_n^*, -T_n, R_n\right)$ is associated. Hence, the conclusion follows from P3.■

Remark:

As n goes to infinity, $-Y_n$ and D_{n+1} converge in distribution to their stationary distributions respectively. Since $(-Y_n, D_n)$ is associated for all $n \geq 1$, it follows that the stationary distributions of $-Y_n$ and D_n are also associated.

We are now ready for the following main

Theorem 2.3.5:

For a stationary GI/D/1 → G/1 system,

$$E\left(D_n^*\right) \leq \frac{\lambda\left(\sigma_T^2 + \sigma_R^2\right)}{2(1 - \rho_2)}$$

where equality holds for D/D/1 → D/1 systems.

Proof:

If $S_n \equiv s$, $\forall n \geq 1$, then $\sigma_S^2 = 0$ and $\mathrm{Cov}\left[S_n, D_n^*\right] = 0$. Hence, (2.3.7) specializes to

$$E\left(D_n^*\right) = \frac{\lambda}{2(1-\rho_2)}\left\{\sigma_T^2 + \sigma_R^2 - \text{Var } Y_n + 2 \text{ Cov } [Y_n, D_{n+1}]\right\}$$

$$\leq \frac{\lambda}{2(1-\rho_2)}\left\{\sigma_T^2 + \sigma_R^2 + 2 \text{ Cov } [Y_n, D_{n+1}]\right\} \quad \text{since} \quad \text{Var } Y_n \geq 0$$

$$\leq \frac{\lambda\left(\sigma_T^2 + \sigma_R^2\right)}{2(1-\rho_2)} \quad \text{since} \quad \text{Cov } [Y_n, D_{n+1}] \leq 0 \quad \text{by Lemma 2.3.4.} \blacksquare$$

Remark:

Bounds for the variance of the interdeparture time from the second station can also be obtained. Let τ_n = time between the n^{th} and $(n+1)^{th}$ departure from the second station. It follows from the fact that $\tau_n = Y_n + R_{n+1}$ that $E(\tau_n) = 1/\lambda$. Since Y_n and R_{n+1} are independent, we have

$$\text{Var } \tau_n = \text{Var } Y_n + \text{Var } R_{n+1} \geq \sigma_R^2 . \tag{2.3.13}$$

In the case $S_n \equiv s$, $\forall n \geq 1$, (2.3.7) becomes:

$$\text{Var } Y_n = \sigma_R^2 + \sigma_T^2 + 2 \text{ Cov } [Y_n, D_{n+1}] - 2E\left(D_n^*\right)E(Y_n)$$

$$\leq \sigma_R^2 + \sigma_T^2 + 2 \text{ Cov } [Y_n, D_{n+1}] \quad \text{since} \quad E\left(D_n^*\right)E(Y_n) \geq 0 \tag{2.3.14}$$

$$\leq \sigma_R^2 + \sigma_T^2 \quad \text{since} \quad \text{Cov } [Y_n, D_{n+1}] \leq 0 .$$

Combining (2.3.13) and (2.3.14) gives

$$\sigma_R^2 \leq \text{Var } \tau_n \leq 2\sigma_T^2 + \sigma_T^2 ,$$

where equality holds both sides for $D/D/1 \rightarrow D/1$ systems.

2.4 A Generalization

Consider the following system of $n(n \geq 2)$ queues in tandem. Customers arrive according to a general renewal process where the inter-arrival times T_n , $n = 1,2, \ldots$ are i.i.d. random variables and have an arbitrary distribution with $E(T_n) = \frac{1}{\lambda}$, Var $T_n = \sigma_T^2$. The first through $(n - 1)^{th}$ stations have constant servers with deterministic service times H_j , $j = 1,2, \ldots, n-1$, and the n^{th} server has an arbitrary service distribution H_n with $E(H_n) = 1/\mu_n$, Var $H_n = \sigma_H^2$. Customers are served according to first in first out rule at each station. Under stationary conditions, an upper bound for the expected total waiting time in system of a customer is found below.

Friedman [8] has shown that if all the servers in a tandem queueing system have constant service times, then, for any arbitrary arrival process, the epoch at which every customer leaves the entire system does not change with changes in order of servers. Now, if we consider the first through $(n - 1)^{th}$ server as the first subsystem and the n^{th} server as the second subsystem, then our system becomes two subsystems in tandem.

In the first subsystem, since the total waiting time is independent of the order of servers, we can rearrange them such that their service times are in decreasing order. Let $d_{[j]}$, $j = 1, \ldots, n-1$, be the delay in front of the j^{th} station after the rearrangement, then $d_{[j]} = 0$ for all $j = 2, \ldots, n-1$, and, by Kingman and Marshall's bound, $E(d_{[1]}) \leq \frac{\lambda\sigma_T^2}{2(1 - H_k)}$ where $k \in \{i \mid i \in \{1, \ldots, n-1\}$, $H_i \geq H_j \ \forall j = 1, \ldots, n-1\}$.

Let $\{t_i , i = 1,2, \ldots\}$ be any realization of the departure epochs from the first station after rearrangement, then it is easy to see that the departure epochs from the first subsystem are

$$\left\{ t_i + \sum_{j=1}^{n-1} H_j - H_k , i \geq 1 \right\} .$$

Now, if we compare our system with a $GI/H_k/1 \rightarrow G/1$ system (i.e., delete the second through $(n-1)^{th}$ stations), the above observation says that the arrival processes to the last station in the two systems differ by a constant $\sum_{j=1}^{n-1} H_j - H_k$. Hence every customer will experience the same amount of delay in front of the last station for any realization of the processes in both systems. Therefore, by Theorem 2.3.5, the expected delay in front of the last station is bounded above by $\frac{\lambda\left(\sigma_T^2 + \sigma_H^2\right)}{2(1 - \lambda/\mu_n)}$.

We have proven the following

Theorem 2.4.1:

$$E(W) \leq \frac{\lambda\sigma_T^2}{2(1 - \lambda H_k)} + \sum_{j=1}^{n-1} H_j + \frac{\lambda\left(\sigma_T^2 + \sigma_H^2\right)}{2(1 - \lambda/\mu_n)} + \frac{1}{\mu_n}$$

where W = total waiting time in system of a stationary customer. Again, equality holds if both T_n and H_n are deterministic.

Remark:

(i) Friedman's result holds even if the deterministic servers have multiple parallel channels. Theorem 2.4.1 can be easily modified to this case.

(ii) Tembe and Wolff's bound, (2.2.1), can also be generalized in exactly the same way.

2.5 Some Correlation Relations

In this section, we investigate some properties of stationary covariances and their relationships. We begin with the following intuitively obvious theorem:

Theorem 2.5.1:

In GI/G/1 → G/1 systems, $\text{Cov}\left[S_n, D_n^*\right] \leq 0$ and $\text{Cov}\left[S_{n+1}, Y_n\right] \geq 0$ for all $n \geq 1$.

Proof:

Conditioning on D_{n-1} , D_{n-1}^* , R_{n-1} , S_{n-1} , T_{n-1} , we have

$$\text{Cov}\left[S_n, D_n^*\right] = E\left\{\text{Cov}\left[S_n, D_n^* \mid D_{n-1}, D_{n-1}^*, R_{n-1}, S_{n-1}, T_{n-1}\right]\right\}$$

$$+ \text{Cov}\left\{E\left(S_n \mid D_{n-1}, D_{n-1}^*, R_{n-1}, S_{n-1}, T_{n-1}\right), E\left(D_n^* \mid D_{n-1}, D_{n-1}^*, R_{n-1}, S_{n-1}, T_{n-1}\right)\right\}.$$

Since S_n is independent of D_{n-1} , D_{n-1}^* , R_{n-1} , S_{n-1} , T_{n-1} , it follows $E\left(S_n \mid D_{n-1}, D_{n-1}^*, R_{n-1}, S_{n-1}, T_{n-1}\right) = E(S_n)$, a constant. Thus the second term above is zero.

$\text{Cov}\left[S_n, D_n^* \mid D_{n-1}, D_{n-1}^*, R_{n-1}, S_{n-1}, T_{n-1}\right] \leq 0$ follows from the fact that D_n^* is a decreasing function of S_n given D_{n-1} , D_{n-1}^* , R_{n-1} , S_{n-1} , T_{n-1} (see (2.3.8)). Therefore, $\text{Cov}\left[S_n, D_n^*\right] \leq 0$.

Using a similar argument, $\text{Cov}\left[S_{n+1}, Y_n\right] \geq 0$ follows by conditioning on D_n , D_n^* , S_n , R_n , T_n . ■

In the case of deterministic first server, D_n and D_n^* are associated for all $n \geq 1$ (Theorem 2.3.3). We have

Theorem 2.5.2:

Each of the following vector is associated for all $n \geq 1$:
$(-X_n, D_n)$, $\left(-X_n, D_n^*\right)$, (D_n, D_{n+1}) , $\left(-Y_n, D_n^*\right)$, (X_n, Y_n) , $\left(-X_n, D_{n+1}^*\right)$,
$\left(-Y_n, D_{n+1}^*\right)$, $\left(D_n^*, D_{n+1}^*\right)$, $\left(D_n, D_{n+1}^*\right)$, and $\left(D_{n+1}, D_n^*\right)$.

Proof:

Similar to the proof of Lemma 2.3.4.■

Next, we will derive some relationships and bounds for various stationary covariances.

Observe that (2.3.4) can also be written as:

$$(2.5.1)\quad \left(D_{n+1} + S_{n+1} + D_{n+1}^*\right) - Y_n = \left(D_n + S_n + D_n^*\right) + R_n - T_n \ .$$

Now, if we apply the same procedures to (2.3.10) and (2.5.1) as in the derivation of (2.3.7), we can get (omitting the details):

$$(2.5.2)\quad E\left(D_n^*\right) = \frac{1}{2E(Y_n)} \left\{ \text{Var } X_n + \sigma_R^2 + \sigma_S^2 - \text{Var } Y_n - 2 \text{ Cov } \left[X_n, D_n^*\right] \right\}$$

and

$$(2.5.3)\quad E\left(D_n^*\right) = \frac{1}{2E(Y_n)} \left\{ \sigma_R^2 + \sigma_T^2 - \text{Var } Y_n + 2 \text{ Cov } [Y_n, D_{n+1}] + 2 \text{ Cov } [Y_n, S_{n+1}] \right\} .$$

Therefore, by comparing (2.3.7), (2.5.2), and (2.5.3), we obtain

Theorem 2.5.3:

For a stationary GI/G/1 → G/1 system,

$$\text{(2.5.4)}\quad \text{Var } X_n - 2 \text{ Cov } \left[X_n, D_n^*\right] = \sigma_S^2 + \sigma_T^2 + 2 \text{ Cov } [Y_n, D_{n+1}] + 2 \text{ Cov } \left[S_n, D_n^*\right] ,$$

$$\text{(2.5.5)}\quad \sigma_S^2 + \text{Cov } \left[S_n, D_n^*\right] = \text{Cov } [Y_n, S_{n+1}] ,$$

$$\text{(2.5.6)}\quad \text{Var } X_n + \sigma_S^2 - 2 \text{ Cov } \left[X_n, D_n^*\right] = \sigma_T^2 + 2 \text{ Cov } [Y_n, D_{n+1}] + 2 \text{ Cov } [Y_n, S_{n+1}] .$$

Corollary 2.5.4:

Under stationary conditions,

$$-\sigma_S^2 \leq \text{Cov } \left[S_n, D_n^*\right] \leq 0$$

$$0 \leq \text{Cov } [Y_n, S_{n+1}] \leq \sigma_S^2$$

for GI/G/1 → G/1 systems. Equality holds both sides if S_n's are deterministic.

Proof:

By Theorem 2.5.1 and (2.5.5).■

Corollary 2.5.5:

Under stationary conditions,

$$\text{Var } X_n - 2 \text{ Cov } \left[X_n, D_n^*\right] = \sigma_T^2 + 2 \text{ Cov } [Y_n, D_{n+1}]$$

if S_n's are deterministic.

Proof:

By (2.5.4) and (2.5.6).■

Corollary 2.5.6:

If the arrival process is Poisson and $S_n = s \quad \forall n \geq 1$, then, under stationary conditions

$$-\frac{s^2}{2} \leq \text{Cov}\left[X_n, D_n^*\right] \leq 0$$

$$-\frac{s^2}{2} \leq \text{Cov}\left[Y_n, D_{n+1}\right] \leq 0 \ .$$

Proof:

For Poisson arrivals, $\sigma_T^2 = \frac{1}{\lambda^2}$ and

$$\begin{aligned} \text{Var } X_n &= E\left(X_n^2\right) - \{E(X_n)\}^2 \\ &= (1 - \lambda s)\,\frac{2}{\lambda^2} - \left(\frac{1}{\lambda} - s\right)^2 \\ &= \frac{1}{\lambda^2} - s^2 \ . \end{aligned}$$

Hence, the inequalities follow from Corollary 2.5.5, Lemma 2.3.4, and Theorem 2.5.2. ■

Corollary 2.5.7:

For a stationary $M/M/1 \rightarrow M/1$ system,

$$\text{Cov}\left[Y_n, D_{n+1}\right] + \text{Cov}\left[S_n, D_n^*\right] + \frac{1}{\mu_1^2} = 0 \ , \tag{2.5.7}$$

and

$$\text{Cov}\left[Y_n, D_{n+1}\right] + \text{Cov}\left[Y_n, S_{n+1}\right] = 0 \ . \tag{2.5.8}$$

Proof:

It is well-known that W_n and W_n^* are independent (see Reich [17]). Since $X_n = -\min [0, W_n - T_n]$ and T_n is independent of W_n^*, it follows from the above fact that X_n and W_n^* are independent. This, in turn, implies X_n and D_n^* are independent. Therefore, $\text{Cov}\left[X_n, D_n^*\right] = 0$.

Noting that $\text{Var } X_n = \frac{1}{\lambda^2} - \frac{1}{\mu_1^2}$, (2.5.4), (2.5.6) simplify to (2.5.7) and (2.5.8) respectively. ■

The exact value of $\text{Cov}[Y_n, D_{n+1}]$ can be found by using the fact that W_n and W_n^* are independent for the $M/M/1 \rightarrow M/1$ system. Explicitly, we have

$$E(Y_n D_{n+1}) = E(Y_n D_{n+1} \mid D_{n+1} > 0) P\{D_{n+1} > 0\}$$

$$= E(Y_n \mid D_{n+1} > 0) E(D_{n+1} \mid D_{n+1} > 0) P\{D_{n+1} > 0\},$$

where we have used the assertion that $(Y_n \mid D_{n+1} > 0)$ and $(D_{n+1} \mid D_{n+1} > 0)$ are independent. The assertion follows since $(Y_n \mid D_{n+1} > 0) = \max\left[0, S_{n+1} - W_n^*\right]$, $(D_{n+1} \mid D_{n+1} > 0) = (W_n - T_n \mid W_n > T_n)$, and the vectors $\left(S_{n+1}, W_n^*\right)$ and (W_n, T_n) are independent of each other. Now,

$$E(D_{n+1} \mid D_{n+1} > 0) P\{D_{n+1} > 0\} = E(D_{n+1})$$

$$= \frac{\lambda}{\mu_1(\mu_1 - \lambda)}.$$

$$E(Y_n \mid D_{n+1} > 0) = E\left\{\max\left[0, S_{n+1} - W_n^*\right]\right\}$$

$$= E\left(S_{n+1} - W_n^* \mid S_{n+1} > W_n^*\right) P\left\{S_{n+1} > W_n^*\right\}$$

$$= \frac{1}{\mu_1}\left(\frac{\mu_2 - \lambda}{\mu_1 + \mu_2 - \lambda}\right).$$

Hence, $E(Y_n D_{n+1}) = \frac{1}{\mu_1}\left(\frac{\mu_2 - \lambda}{\mu_1 + \mu_2 - \lambda}\right)\left(\frac{\lambda}{\mu_1(\mu_1 - \lambda)}\right)$, and

$$\text{Cov}\,[Y_n, D_{n+1}] = E(Y_n D_{n+1}) - E(Y_n)E(D_{n+1})$$

$$(2.5.9) \qquad = \frac{\lambda(\mu_2 - \lambda)}{\mu_1^2(\mu_1 - \lambda)(\mu_1 + \mu_2 - \lambda)} - \frac{(\mu_2 - \lambda)}{\lambda\mu_2} \cdot \frac{\lambda}{\mu_1(\mu_1 - \lambda)}$$

$$= -\frac{(\mu_2 - \lambda)(\mu_1 + \mu_2)}{\mu_1^2\mu_2(\mu_1 + \mu_2 - \lambda)}.$$

Therefore, $\text{Cov}\left[S_n, D_n^*\right]$ and $\text{Cov}\,[Y_n, S_{n+1}]$ can be obtained via (2.5.9) and Corollary 2.5.7.

CHAPTER 3

COMPARISON OF WAITING TIMES FOR TANDEM QUEUEING SYSTEMS

3.1 Introduction

It is generally believed that, in a queueing system, the more regular the arrival process and the service times are, the better the system performance will be.

Stoyan [18], [21] gave a definition of regularity which is very useful in this context, and proved that the above principle holds for GI/G/1 queues under his definition of regularity.

The main purpose of this chapter is to extend his theory to tandem queueing systems. It is proven that for a tandem system with many stations in series, where each station can have either one server with an arbitrary service distribution or a number of constant servers in parallel, the expected total waiting time in system of every customer decreases as the interarrival times and service times become more egular.

We will start with some preliminaries.

3.2 Some Preliminaries

Define the following partial ordering on the set of all distribution functions:

Definition:

For two distribution functions F and G, we say that $F \overset{v}{\leq} G$ if

$$\int f(x)dF(x) \leq \int f(x)dG(x)$$

for all increasing convex function f .

Definition:

For two arbitrary random variables X and Y , we say that $X \overset{v}{\leq} Y$ if their corresponding distribution functions satisfy the above ordering.

A list of some useful properties of this ordering follows:

(a) If $X \overset{v}{\leq} Y$, then $E(x) \leq E(Y)$.

(b) $F \overset{v}{\leq} G$ if and only if

$$\int_t^\infty [1 - F(x)]dx \leq \int_t^\infty [1 - G(x)]dx \quad \forall t .$$

(See Theorem 14 of Bessler and Veinott [2].)

(c) If F and G have the same mean, then $F \overset{v}{\leq} G$ if and only if

$$\int f(x)dF(x) \leq \int f(x)dG(x)$$

for all convex function f (see Ross [19]).

(d) Let F and G be two distribution functions with the same mean and $F(0) = G(0) = 0$. Suppose $\bar{F}$ ($\bar{F}(x) = 1 - F(x)$) crosses $\bar{G}$ ($\bar{G}(x) = 1 - G(x)$) at most once, and that if such a crossing occurs, $\bar{F}$ crosses $\bar{G}$ from above. Then $F \overset{v}{\leq} G$ (see Lemma 7.5 of Barlow and Proschan [1]).

Before giving the next property, we need the following definition:

Definition:

A nonnegative random variable X with finite mean is said to have NBUE (NWUE) distribution if

$$E(X - t \mid X > t) \underset{(\geq)}{\leq} E(X) \quad \forall t \geq 0 .$$

(e) Let X be a nonnegative random variable, and Y be an exponential random variable with the same mean as X. Then $X \overset{v}{\underset{(\geq)}{\leq}} Y$ if and only if X has NBUE (NWUE) distribution. (See Theorem 4.6 of Marshall and Proschan [15]).

It should be noted that, for two random variables X and Y with the same mean, property (c) implies that $\text{Var } X \leq \text{Var } Y$ if $X \overset{v}{\leq} Y$. Hence, this ordering indeed orders the relative variabilities of random variables with the same mean.

We now state the following theorem of Stoyan:

Theorem 3.2.1:

Let $F_1/G_1/1$ and $F_2/G_2/1$ be two single server queues with interarrival time distributions F_i, $i = 1,2$, and service time distributions G_i, $i = 1,2$. If $\int_0^\infty x dF_1(x) = \int_0^\infty x dF_2(x)$, $F_1 \overset{v}{\leq} F_2$, and $G_1 \overset{v}{\leq} G_2$, then $D_\infty^1 \overset{v}{\leq} D_\infty^2$, where D_∞^i, $i = 1,2$, are the stationary delays of systems $F_i/G_i/1$, $i = 1,2$, respectively.

3.3 Customer Average and Time Average Delays

In this section, some relations between customer average and time average delays for some classes of single server queues are investigated. By "customer average delay," we mean the average amount of work in system as seen by an arriving customer, and time average delay is defined as the average over time of the amount of work in system at an arbitrary time point t. It should be noted that the amount of work in system at time t is equal to the amount of time a customer has to wait if he arrives at time t.

Marshall and Wolff [14] showed that the time average delay is greater than the customer average delay in a single server queue if $\mu\sigma_a^2 \leq 1$ where μ is the service rate and σ_a is the coefficient of variation of the interarrival time.

In the next theorem, we use Stoyan's theorem to obtain a new sufficient condition for time average delay to be greater (smaller) than customer average delay in GI/G/1 queues.

Let $E(V)$ and $E(D)$ be the time average delay and customer average delays in a GI/G/1 queue respectively. Brumelle [4] showed that the following relation holds:

$$E(V) = \rho E(D) + \frac{\lambda E(S^2)}{2},$$

where λ is the arrival rate, $E(S)$ and $E(S^2)$ are the first and second moments of the service time distribution, and $\rho = \lambda E(S)$. It follows that $E(V) \underset{(\leq)}{\geq} E(D)$ if and only if

$$E(D) \underset{(\geq)}{\leq} \frac{\lambda E(S^2)}{2(1-\rho)} . \tag{3.3.1}$$

Now, compare the original GI/G/1 system with a M/G/1 system where the service distribution stays the same but the arrival process is Poisson with rate λ. Notice that the right-hand side of (3.3.1) is just the expected delay for the M/G/1 system. Therefore, the next theorem follows from property (e) of variability ordering and Theorem 3.2.1.

Theorem 3.3.1:

In GI/G/1 queues, customer average delay is smaller (greater) than time average delay if the interarrival times have i.i.d. NBUE (NWUE) distributions.

Denote a single server queueing system with i.i.d. NBUE (NWUE) interarrival time distributions by a NBUE/G/1 (NWUE/G/1) system. As a consequence of above, Marshall's bounds [13] for the expected delays in NBUE/G/1 and NWUE/G/1 queues can be improved. Let σ_a^2 and σ_g^2 be the variances of the interarrival time and service distributions respectively. We have

(i) For NBUE/G/1 queues with $\rho < 1$,

$$\max\left[\ell, J - \frac{(1+\rho)}{2\lambda}\right] \leq E(D) \leq \min\left[J, \frac{\lambda E(S^2)}{2(1-\rho)}\right] . \tag{3.3.2}$$

(ii) For NWUE/G/1 queues with $\rho < 1$,

$$\max\left[\ell, \frac{\lambda E(S^2)}{2(1-\rho)}\right] \leq E(D) \leq J - \frac{(1+\rho)}{2\lambda} , \tag{3.3.3}$$

where $J = \frac{\lambda\left(\sigma_a^2 + \sigma_g^2\right)}{2(1 - \rho)}$ and ℓ is the solution of $\int_{-x}^{\infty} [1 - K(u)]du$, $x \geq 0$, K is the distribution of the difference of a service time and an interarrival time.

Letting $C_a = \lambda\sigma_a$, we have

$$\frac{\lambda E(S^2)}{2(1 - \rho)} = J - \frac{C_a^2 - \rho^2}{2\lambda(1 - \rho)} .$$

Thus,

$$\frac{\lambda E(S^2)}{2(1 - \rho)} \to J - \frac{1 + \rho}{2\lambda} \quad \text{as} \quad C_a \to 1 .$$

Therefore, the upper and lower bounds converge to each other as $C_a \to 1$ in both (3.3.2) and (3.3.3). This implies

$$E(D) \to \frac{\lambda E(S^2)}{2(1 - \rho)} \quad \text{as} \quad C_a \to 1 .$$

Hence $|E(D) - E(V)| \to 0$ as $C_a \to 1$ for both NBUE/G/1 and NWUE/G/1 queus.

Remark:

Marshall's bound for IFR/G/1 and DFR/G/1 queues can also be improved in the same way.

3.4 Comparison of Waiting Times

We now proceed to generalize Theorem 3.2.1 to tandem queueing systems.

Consider a queueing system, Δ_1, with n stations in series. There is only one server at each station. Customers arrive according to a renewal process with interarrival time distribution F. The service times at station j have distribution G_j, $j = 1,2, \ldots, n$. Let

T_k = interarrival time between the $(k-1)^{th}$ and k^{th} customer, $k = 1,2, \ldots$.

S_k^j = service time of the k^{th} customer at station j, $j = 1,2, \ldots, n$, $k = 1,2, \ldots$.

Z_k^j = epoch at which the k^{th} customer leaves station j, $j = 1,2, \ldots, n$, $k = 1,2, \ldots$. $Z_k^o \equiv \sum_{i=1}^{k} T_i$.

We make the assumption that the system starts operation at time zero with no customer in system, and customers are served in the order of their arrival at each station.

The following lemmas are essential in the proofs below:

Lemma 3.4.1:

Let X_i, Y_i, $i = 1,2, \ldots, n$ be independent random variables. Then, $X_i \overset{v}{\leq} Y_i$ for all $i = 1,2, \ldots, n$ if and only if

$$f(X_1, X_2, \ldots, X_n) \overset{v}{\leq} f(Y_1, Y_2, \ldots, Y_n)$$

for all increasing convex function f.

Proof:

The if part is trivial, so we only have to prove the other direction. The proof is by induction on n.

The lemma obviously holds for $n = 1$ by definition of variability. Suppose it is true for k. Denote $F_{X_{k+1}}\left(F_{Y_{k+1}}\right)$ the distribution of $X_{k+1}(Y_{k+1})$. Let f and g be arbitrary increasing convex functions. Noting that increasing convex functions of increasing convex functions are still increasing convex, we have

$$E\{g[f(X_1,X_2, \ldots, X_k,X_{k+1})]\}$$

$$= E\{g[f(X_1,X_2, \ldots, X_k,x_{k+1})] \mid X_{k+1} = x_{k+1}\}dF_{X_{k+1}}(x_{k+1})$$

$$\leq E\{g[f(Y_1,Y_2, \ldots, Y_k,x_{k+1})] \mid X_{k+1} = x_{k+1}\}dF_{X_{k+1}}(x_{k+1})$$

by induction hypothesis

$$\leq E\{g[f(Y_1,Y_2, \ldots, Y_k,x_{k+1})] \mid Y_{k+1} = x_{k+1}\}dF_{Y_{k+1}}(x_{k+1})$$

since $E\{g[f(Y_1,Y_2, \ldots, Y_k,x_{k+1})] \mid X_{k+1} = x_{k+1}\}$ is an increasing convex function of x_{k+1} and $X_{k+1} \overset{v}{\leq} Y_{k+1}$

$$= E\{g[f(Y_1,Y_2, \ldots, Y_k,Y_{k+1})]\} . \blacksquare$$

Lemma 3.4.2:

Z_k^j is an increasing convex function of $T_1, \ldots, T_k, S_1^1, \ldots, S_k^1, \ldots, S_1^j, \ldots, S_k^j$ for all $j = 1,2, \ldots, n$, $k = 1,2, \ldots$.

Proof:

Observe that $\max\left[Z_k^{j-1}, Z_{k-1}^j\right]$ is the epoch at which the k^{th} customer enters service at station j. Hence

$$Z_k^j = \max\left[Z_k^{j-1}, Z_{k-1}^j\right] + S_k^j \quad \text{for all} \quad j = 1,2, \ldots, n \ , \ k = 1,2, \ \ldots \ .$$

Since $Z_k^o = \sum_{i=1}^{k} T_i$, $k = 1,2, \ \ldots$, the conclusion follows by induction on j and k .■

Now, consider a second such system, Δ_2 , with interarrival distribution $\tilde{F}$ and service distribution $\tilde{G}_j$ at station j . Let all wiggled notations (e.g. $\tilde{Z}_k$) denote the corresponding entities for system Δ_2 . It follows from Lemma 3.4.1 and 3.4.2 that

(3.4.1) $\quad Z_k^j \overset{v}{\leq} \tilde{Z}_k^j$ for all $j = 1,2, \ldots, n$, $k = 1,2, \ \ldots$,

if $F \overset{v}{\leq} \tilde{F}$ and $G_j \overset{v}{\leq} \tilde{G}_j$ for all $j = 1,2, \ldots, n$.

Let $W_k(\tilde{W}_k)$ be the total waiting time in system $\Delta_1(\Delta_2)$, we have

Theorem 3.4.3:

If $\int x dF(x) = \int x d\tilde{F}(x) < \infty$, $F \overset{v}{\leq} \tilde{F}$, and $G_j \overset{v}{\leq} \tilde{G}_j$ for all $j = 1,2, \ldots, n$, then $E(W_k) \leq E(\tilde{W}_k)$ for all $k = 1,2, \ \ldots$.

Proof:

$$W_k = Z_k^n - \sum_{i=1}^{k} T_i$$

$$\tilde{W}_k = \tilde{Z}_k^n - \sum_{i=1}^{k} \tilde{T}_i \ .$$

Taking expectations and substracting, we have:

$$E(W_k) - E(\tilde{W}_k) = E\left(Z_k^n\right) - E\left(\tilde{Z}_k^n\right) \leq 0 ,$$

where the last inequality follows from (3.4.1).■

If the arrival processes of Δ_1 and Δ_2 are the same, a stronger relation between W_k and $\tilde{W}_k$ holds.

Theorem 3.4.4:

If $F(x) = \tilde{F}(x)$ for all $x \geq 0$, then

$$W_k \overset{v}{\leq} \tilde{W}_k \quad \text{for all} \quad k = 1,2, \ldots .$$

Proof:

Let f be an arbitrary increasing convex function. Conditioning on the sequence of arrival epochs $\{t_i\}_{i=1}^k$, we have

$$E\left\{f(W_k) \mid \{t_i\}_{i=1}^k\right\} \leq E\left\{f(\tilde{W}_k) \mid \{t_i\}_{i=1}^k\right\}$$

since $\left[W_k \mid \{t_i\}_{i=1}^k\right]\left(\left[\tilde{W}_k \mid \{t_i\}_{i=1}^k\right]\right)$ is an increasing function of $S_i^j \left(\tilde{S}_i^j\right)$, $j = 1,2, \ldots, n$, $i = 1,2, \ldots, k$.

Unconditioning, we have $E[f(W_k)] \leq E[f(\tilde{W}_k)]$.■

Remark:

In the proof of Theorem 3.4.4, we only needed the assumption that the sequence $\{t_i\}_{i=1}^k$ has the same joint distribution for both systems. Therefore, the interarrival times may be "dependent."

Another interesting question one might ask is: Can Theorem 3.2.1 be generalized to parallel server queues? Ross [19] gave a counter-example showing that similar comparisons do not hold in general. However, we have the following theorem:

Theorem 3.4.5:

Let $A/S_1/m$ and $B/S_2/m$ be two parallel server queues with interarrival time distribution A and B, deterministic service distributions S_1 and S_2, and m servers in parallel respectively. If $\int x dA(x) = \int x dB(x)$, $A \overset{v}{\leq} B$, and $S_1 \overset{v}{\leq} S_2$, then

$$D_k \overset{v}{\leq} \bar{D}_k \quad \text{for all} \quad k = 1,2, \ldots ,$$

where $D_k(\bar{D}_k)$ is the delay in queue of the k^{th} customer in system $A/S_1/m$ $(B/S_2/m)$.

Proof:

Let $T_i(\bar{T}_i)$, $i = 1,2, \ldots,$ be the interarrival time of the i^{th} and $(i+1)^{th}$ customer in system $A/S_1/m$ $(B/S_2/m)$. Since $T_i \overset{v}{\leq} \bar{T}_i$ and $E(T_i) = E(\bar{T}_i)$ for all $i = 1,2, \ldots$, we have by property (c) of variability that $-T_i \overset{v}{\leq} -\bar{T}_i$ for all $i = 1,2, \ldots$. Let $S_i = s_i$, $i = 1,2$, with probability one. Noting that customers depart in the order of their arrivals in a parallel server queue with constant service times, we have the following relations:

$$(3.4.2) \qquad D_k = \max \left[0, D_{k-m} + s_1 - \sum_{j=k-m}^{k-1} T_j \right] \quad \text{for all} \quad k > n ,$$

$$\bar{D}_k = \max\left[0, \bar{D}_{k-m} + s_2 - \sum_{j=k-m}^{k-1} \bar{T}_j\right] \quad \text{for all} \quad k > n \,, \tag{3.4.3}$$

$$\text{and} \quad D_k \equiv 0 \,, \; \bar{D}_k \equiv 0 \quad \text{for all} \quad 1 \leq k \leq n \,.$$

Clearly, $D_k \overset{v}{\leq} \bar{D}_k$ for all $1 \leq k \leq n$. It follows from (3.4.2) and (3.4.3) that $D_k(\bar{D}_k)$ is an increasing convex function of $\left(D_{k-m}, -\sum_{j=k-m}^{k-1} T_j\right)\left(\left(\bar{D}_{k-m}, -\sum_{j=k-m}^{k-1} \bar{T}_j\right)\right)$. Therefore, the conclusion follows by induction on k. ■

Remark:

The same proof will go through if we allow the service times to be random but satisfying the following conditions:

(i) $S_1 \overset{v}{\leq} S_2$

(ii) $P\{a_i \leq S_i \leq b_i\} = 1 \,, \; i = 1,2 \,,$

(iii) $A(b_1 - a_1) = 0$ and $B(b_2 - a_2) = 0$.

In other words, we want the customers to depart in the order of their arrivals. Under this condition, relation (3.4.2) and (3.4.3) remain valid and the inductive proof still holds.

Finally, combining Theorem 3.4.3, 3.4.4, and 3.4.5, we have the following

Theorem 3.4.6:

Consider a tandem queueing system with n stations in series where each station can have either one server with general service distribution or any number of servers in parallel with constant

service times. For two such systems, if the interarrival and service distributions are ordered in the same way as in Theorem 3.4.3, 3.4.4, and 3.4.5, then similar order relations hold for the total waiting time in system for every customer.

Proof:

It can be shown by induction that the time until any customer leaves the whole system is an increasing convex function of all the interarrival and service times of all customers up to him. Hence, the conclusion follows in a similar fashion as in the proofs of Theorem 3.4.3, 3.4.4, and 3.4.5.■

3.5 Examples

The results obtained in the last section are quite useful for finding bounds for total waiting times in tandem queueing systems. Suppose we know the ordering between two systems and one of them can be analyzed exactly. Then, bounds for the corresponding quantities in the other system can be readily obtained. We give two specific examples below.

I. Consider the following four systems of tandem queues:

(A) $M/A_1/1 \to A_2/1 \to \cdots \to A_n/1$

(B) $M/B_1/1 \to B_2/1 \to \cdots \to B_n/1$

(C) $M/C_1/1 \to C_2/1 \to \cdots \to C_n/1$

(D) $M/D_1/1 \to D_2/1 \to \cdots \to D_n/1$.

where

A_i , $i = 1,2, \ldots, n$ have NWUE distributions,

B_i , $i = 1,2, \ldots, n$ have exponential distributions,

C_i , $i = 1,2, \ldots, n$ have NBUE distributions,

D_i , $i = 1,2, \ldots, n$ are deterministic.

Assume that A_i , B_i , C_i , D_i have the same mean for all $i = 1,2, \ldots, n$ and the arrival processes are Poisson with the same rate for all four systems. It follows from property (e) of variability ordering that $D_i \overset{v}{\leq} C_i \overset{v}{\leq} B_i \overset{v}{\leq} A_i$ for all $i = 1,2, \ldots, n$.

Let $W_A(W_B,W_C,W_D)$ be the stationary total waiting time of a customer in system $A(B,C,D)$. By Theorem 3.4.4, we have

$$W_D \overset{v}{\leq} W_C \overset{v}{\leq} W_B \overset{v}{\leq} W_A .$$

The point of this example is that the distributions of W_D and W_B are known exactly. Hence, we have both upper and lower bounds for $E\left(W_C^r\right)$ and a lower bound for $E\left(W_A^r\right)$ for all real number $r \geq 1$.

II. Tembe and Wolff showed that the expected delay in front of the second server in a $M/D/1 \rightarrow M/1$ system is bounded above by $\frac{\lambda}{\mu(\mu - \lambda)}$ (see (2.2.1)) where λ is the arrival rate and μ is the service rate at the second station.

Theorem 3.4.4 implies that the corresponding expected delay, d^* , is smaller for a $M/D/1 \rightarrow G/1$ system where G has NBUE distribution and the same arrival and service rates are assumed. Therefore, combining with Theorem 2.3.5, we have

$$d^* \leq \min\left[\frac{\lambda}{\mu(\mu-\lambda)}, \frac{\lambda\left(\sigma_a^2+\sigma_g^2\right)}{2(1-\lambda/\mu)}\right]$$

where $\sigma_a^2 = 1/\lambda^2$ and σ_g^2 is the variance of G.

3.6 Some Stronger Comparisons

Sometimes it is of interest to know not only comparisons of total waiting times in system but also the delays in individual stations in tandem queues. Under stronger assumptions (e.g., stochastic ordering) between interarrival and service times of two systems, some results in this direction can be obtained.

Consider two systems of tandem queues $F/G/1 \to H/1$ and $\bar{F}/G/1 \to \bar{H}/1$ where $F(\bar{F})$ is the interarrival distribution, $H(\bar{H})$ is the service distribution of the second station, and G is the service distribution of the first station for both systems. For system $F/G/1 \to H/1$, let $D_n\left(D_n^*\right)$ be the delay of the n^{th} customer in front of the first (second) station, S_n and R_n be the service times of the n^{th} customer at the first and second station respectively, and T_n be the interarrival time of the n^{th} and $(n+1)^{th}$ customer. Let all barred notations $\left(\text{e.g., } \bar{D}_n^*\right)$ denote the corresponding quantities in system $\bar{F}/G/1 \to \bar{H}/1$.

From (2.3.8), we have that

$$(3.6.1)\quad D_{n+1}^* = \max\left[0, \min\left[D_n + S_n + D_n^* + R_n - T_n - S_{n+1}, D_n^* + R_n - S_{n+1}\right]\right],$$

and

$$(3.6.2)\quad \bar{D}_{n+1}^* = \max\left[0, \min\left[\bar{D}_n + \bar{S}_n + \bar{D}_n^* + \bar{R}_n - \bar{T}_n - \bar{S}_{n+1}, \bar{D}_n^* + \bar{R}_n - \bar{S}_{n+1}\right]\right].$$

Denote stochastic ordering by $\overset{st}{\leq}$. Before proving the next theorem, we need the following

Lemma 3.6.1:

Let X_i , Y_i , $i = 1,2, \ldots, n$, be independent random variables. Then, $X_i \overset{st}{\leq} Y_i$ for all $i = 1,2, \ldots, n$ if and only if

$$f(X_1,X_2, \ldots, X_n) \overset{st}{\leq} f(Y_1,Y_2, \ldots, Y_n)$$

for all increasing function f .

Proof:

Straightforward by induction on n . ■

Theorem 3.6.2:

If $F \overset{st}{\geq} \bar{F}$ and $H \overset{st}{\leq} \bar{H}$, then

$$D_k^* \overset{st}{\leq} \bar{D}_k^* \quad \text{for all} \quad k = 1,2, \ldots .$$

Proof:

Conditioning on the sequence of service times, $\{S_j\}_{j=1}^k$, at first station, it follows from (3.6.1) and (3.6.2) that $D_k^*\left(\bar{D}_k^*\right)$ are increasing functions of $(-T_1, \ldots, -T_{k-1},R_1, \ldots, R_{k-1})$ $((-\bar{T}_1, \ldots, -\bar{T}_{k-1},\bar{R}_1, \ldots, \bar{R}_{k-1}))$ for all $k = 1,2, \ldots$. Hence by Lemma 3.6.1,

$$\left[D_k^* \mid \{S_j\}_{j=1}^k\right] \overset{st}{\leq} \left[\bar{D}_k^* \mid \{S_j\}_{j=1}^k\right] \quad \text{for all} \quad k = 1,2, \ldots .$$

Unconditioning the above finishes the proof. ■

Theorem 3.6.3:

If $F \overset{st}{\geq} \bar{F}$ and $H \overset{v}{\leq} \bar{H}$, then

$$D_k^* \overset{v}{\leq} \bar{D}_k^* \quad \text{for all} \quad k = 1,2, \ldots .$$

Proof:

Let f be an arbitrary increasing convex function. Denote $F_{T_1, \ldots, T_{k-1}}$, $F_{\bar{T}_1, \ldots, \bar{T}_{k-1}}$, and $G_{S_1, \ldots, S_k}$ the joint distribution of $\{T_i\}_{i=1}^{k-1}$, $\{\bar{T}_i\}_{i=1}^{k-1}$ and $\{S_i\}_{i=1}^{k}$ respectively. Conditioning on $\{T_i\}_{i=1}^{k-1}$ and $\{S_i\}_{i=1}^{k}$, we have

$$E\left\{f\left(D_k^*\right)\right\} = \iint E\left\{f\left(D_k^*\right) \mid \{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^{k}\right\} \cdot$$

$$dF_{T_1, \ldots, T_{k-1}}(t_1, \ldots, t_{k-1}) dG_{S_1, \ldots, S_k}(s_1, \ldots, s_k)$$

$$\leq \iint E\left\{f\left(\bar{D}_k^*\right) \mid \{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^{k}\right\} \cdot$$

$$dF_{T_1, \ldots, T_{k-1}}(t_1, \ldots, t_{k-1}) dG_{S_1, \ldots, S_k}(s_1, \ldots, s_k)$$

since $\left[f\left(D_k^*\right) \mid \{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^{k}\right]$ and $\left[f\left(\bar{D}_k^*\right) \mid \{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^{k}\right]$ are increasing convex functions of $(R_1, \ldots, R_{k-1})$ and $(\bar{R}_1, \ldots, \bar{R}_{k-1})$ respectively (see (3.6.1) and (3.6.2))

$$\leq \iint E\left\{f\left(\bar{D}_k^*\right) \mid \{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^{k}\right\} \cdot$$

$$dF_{\bar{T}_1, \ldots, \bar{T}_{k-1}}(t_1, \ldots, t_{k-1}) dG_{S_1, \ldots, S_k}(s_1, \ldots, s_k)$$

by Lemma 3.6.1

$$= E\{f(\bar{D}_k)\} . \blacksquare$$

Remark:

(i) In the proofs of Theorem 3.6.2 and 3.6.3, we only needed $(T_1, \ldots, T_{k-1}) \overset{st}{\geq} (T_1, \ldots, T_{k-1})$ for all $k = 1,2, \ldots$. Hence the arrival processes may be "dependent." Also, only $(R_1, \ldots, R_{k-1}) \overset{st}{\leq} (\bar{R}_1, \ldots, \bar{R}_{k-1})$ for all $k = 1,2, \ldots$ was used in Theorem 3.6.2.

(ii) Analogues of Theorem 3.6.2 and 3.6.3 can easily be proven for two systems with n $(n \geq 2)$ queues in tandem where the first through $(n - 1)^{th}$ servers have the same service distribution for both systems.

CHAPTER 4

A HETEROGENEOUS ARRIVAL QUEUEING LOSS MODEL

4.1 Descriptions of the Model

Consider a single server queueing loss system. The system alternates between two phases (1 and 2). During phase i, $i = 1,2$, customers arrive according to a Poisson process with rate λ_i, and the service times required by customers have distribution G_i with finite mean. The amounts of time the system spends in phase i, $i = 1,2$, are exponential random variables with rate $c\alpha_i$, $\alpha_i \neq 0$, $i = 1,2$. An arriving customer who finds the system busy departs immediately and is called a lost customer.

Notice that the average arrival rate, $\bar{\lambda}$, of customers to the system is

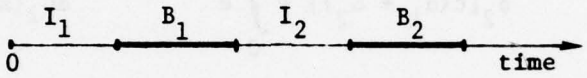
$$\bar{\lambda} = \frac{\lambda_1\alpha_2 + \lambda_2\alpha_1}{\alpha_1 + \alpha_2}$$

which is independent of c. However, as c increases, the rate at which the system switches between phases increases, and the arrival process converges to a stationary Poisson process with rate $\bar{\lambda}$ as c goes to infinity.

Let $L(c)$ be the long run proportion of customers lost for a given c in this model. In the next section, we will find an explicit formula for $L(c)$, and examine the properties of it in some special cases.

4.2 The Loss Formula

The system operates similar to a counter model. Each time an arriving customer finds the system empty, he enters service immediately and block the system for a random period of time whose distribution depends on which phase the system is in at the time of his arrival. During the blocking period, all arriving customers are lost. Hence the system alternates between busy and idle periods.



Let $Z_n = i$, $i = 1,2$, if the system is in phase i at the beginning of the n^{th} busy period. We will start by proving the following

Theorem 4.2.1:

The embedded Markov chain $\{Z_n , n = 1,2, \ldots\}$ has a stationary distribution and is given by

$$\pi_1 = \lim_{n\to\infty} P\{Z_n = 1\} = \frac{P_{21}}{P_{12} + P_{21}} , \tag{4.2.1}$$

and

$$\pi_2 = \lim_{n\to\infty} P\{Z_n = 2\} = \frac{P_{12}}{P_{12} + P_{21}} , \tag{4.2.2}$$

where

$$P_{12} = \frac{\alpha_1\lambda_1\lambda_2 + c\alpha_1^2\lambda_2 + c\alpha_1\alpha_2\lambda_2 - \alpha_1\lambda_1\lambda_2\phi_1[c(\alpha_1 + \alpha_2)]}{(\alpha_1 + \alpha_2)(\lambda_1\lambda_2 + c\alpha_2\lambda_1 + c\alpha_1\lambda_2)} , \tag{4.2.3}$$

$$(4.2.4)\quad P_{21} = \frac{\alpha_2\lambda_1\lambda_2 + c\alpha_2^2\lambda_1 + c\alpha_1\alpha_2\lambda_1 - \alpha_2\lambda_1\lambda_2\phi_2[c(\alpha_1 + \alpha_2)]}{(\alpha_1 + \alpha_2)(\lambda_1\lambda_2 + c\alpha_2\lambda_1 + c\alpha_1\lambda_2)} ,$$

and

$$\phi_1[c(\alpha_1 + \alpha_2)] = \int_0^\infty e^{-c(\alpha_1+\alpha_2)x} dG_1(x) ,$$

$$\phi_2[c(\alpha_1 + \alpha_2)] = \int_0^\infty e^{-c(\alpha_1+\alpha_2)x} dG_2(x) .$$

Proof:

Let the transition probability matrix of the Markov chain $\{Z_n , n = 1,2, \ldots\}$ be

$$\underline{P} = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$$

where $P_{ij} = P\{Z_{n+1} = j \mid Z_n = i\}$, $i = 1,2$, $j = 1,2$.

If P_{ij}'s were given, then $\underline{\pi} = (\pi_1,\pi_2)$ can be obtained by solving $\underline{\pi} = \underline{\pi}\,\underline{P}$ and $\pi_1 + \pi_2 = 1$. It is easy to verify that (4.2.1) and (4.2.2) indeed satisfies the above. Hence, we only have to find P_{ij} , $i = 1,2$, $j = 1,2$.

Let

P_{B_i,E_ℓ} = the probability that the system will be in phase ℓ at the end of a busy period, given that it is in phase i at the beginning of that busy period, $i = 1,2$, $\ell = 1,2$;

and

P_{E_ℓ,B_j} = the probability that the system will be in phase j at the beginning of next busy period, given that it is in phase ℓ at the end of a busy period, $\ell = 1,2$, $j = 1,2$.

Conditioning on the phase of the system at the end of a busy period, we have

$$(4.2.5) \qquad P_{ij} = \sum_{\ell=1}^{2} P_{B_i,E_\ell} \cdot P_{E_\ell,B_j} \ , \ i = 1,2 \ , \ j = 1,2 \ .$$

We have to find P_{B_i,E_ℓ} and P_{E_ℓ,B_j} , $i = 1,2$, $j = 1,2$.

Let $Z(t) = i$, $i = 1,2$, if the system is in phase i at time t . Then, $\{Z(t), t \geq 0\}$ is a continuous time Markov chain. Denote $P_{ij}(t)$ the probability that the phase of the system is j at time t , given that it is in phase i at time 0 . It may be readily verified that

$$P_{11}(t) = e^{-c(\alpha_1+\alpha_2)t} + \left[1 - e^{-c(\alpha_1+\alpha_2)t}\right] \cdot \left[\frac{\alpha_2}{\alpha_1 + \alpha_2}\right] ,$$

$$P_{12}(t) = \left[1 - e^{-c(\alpha_1+\alpha_2)t}\right] \cdot \left[\frac{\alpha_1}{\alpha_1 + \alpha_2}\right] ,$$

$$P_{22}(t) = e^{-c(\alpha_1+\alpha_2)t} + \left[1 - e^{-c(\alpha_1+\alpha_2)t}\right] \cdot \left[\frac{\alpha_1}{\alpha_1 + \alpha_2}\right] ,$$

and

$$P_{21}(t) = \left[1 - e^{-c(\alpha_1+\alpha_2)t}\right] \cdot \left[\frac{\alpha_2}{\alpha_1 + \alpha_2}\right] ,$$

satisfies the set of Kolmogorov backward differential equations (see Ross [20, p. 111) for the process $\{Z(t), t \geq 0\}$. Therefore, by conditioning on the length of a busy period, we have

$$P_{B_i,E_\ell} = \int_0^\infty P_{i\ell}(t)dG_i(t) \ , \ i = 1,2 \ , \ \ell = 1,2 \ ,$$

or explicitly,

$$P_{B_1,E_1} = \frac{\alpha_2}{\alpha_1 + \alpha_2} + \frac{\alpha_1}{\alpha_1 + \alpha_2} \cdot \phi_1[c(\alpha_1 + \alpha_2)] \ ,$$

$$P_{B_1,E_2} = \frac{\alpha_1}{\alpha_1 + \alpha_2} \cdot [1 - \phi_1[c(\alpha_1 + \alpha_2)]] \ ,$$

$$P_{B_2,E_1} = \frac{\alpha_2}{\alpha_1 + \alpha_2} \cdot [1 - \phi_2[c(\alpha_1 + \alpha_2)]] \ ,$$

and

$$P_{B_2,E_2} = \frac{\alpha_1}{\alpha_1 + \alpha_2} + \frac{\alpha_2}{\alpha_1 + \alpha_2} \cdot \phi_2[c(\alpha_1 + \alpha_2)] \ .$$

Next, we will compute P_{E_ℓ,B_j} , $\ell = 1,2$, $j = 1,2$. Since

P_{E_1,B_1} = P{an arrival occurs before the system switches to phase 2} · 1

+ P{no arrival occurs before the system switches to phase 2} · P_{E_1,B_1}

$$= \frac{\lambda_1}{\lambda_1 + c\alpha_1} + \left(\frac{c\alpha_1}{\lambda_1 + c\alpha_1}\right) \cdot \left(\frac{c\alpha_2}{\lambda_2 + c\alpha_2}\right) \cdot P_{E_1,B_1} \ .$$

Solving for P_{E_1,B_1}, we get

$$P_{E_1,B_1} = \frac{\lambda_1(\lambda_2 + c\alpha_2)}{\lambda_1\lambda_2 + c\alpha_2\lambda_1 + c\alpha_1\lambda_2} .$$

By symmetry, we have

$$P_{E_2,B_2} = \frac{\lambda_2(\lambda_1 + c\alpha_1)}{\lambda_1\lambda_2 + c\alpha_2\lambda_1 + c\alpha_1\lambda_2} .$$

Also,

$$P_{E_1,B_2} = 1 - P_{E_1,B_1}$$

$$= \frac{c\alpha_1\lambda_2}{\lambda_1\lambda_2 + c\alpha_2\lambda_1 + c\alpha_1\lambda_2} ,$$

and

$$P_{E_2,B_1} = 1 - P_{E_2,B_2}$$

$$= \frac{c\alpha_2\lambda_1}{\lambda_1\lambda_2 + c\alpha_2\lambda_1 + c\alpha_1\lambda_2} .$$

Substituting P_{B_i,E_ℓ} and P_{E_ℓ,B_j}, $i = 1,2$, $j = 1,2$, $\ell = 1,2$, into (4.2.5) gives (4.2.3) and (4.2.4). ■

Let K_i be the number of arrivals in the i^{th} busy period, $i = 1,2, \ldots$. The distribution of K_i depends on the phase of the system at the beginning of the i^{th} busy period. The explicit form of it will be difficult to obtain. However, we have the following lemmas which are needed in the proof below.

Lemma 4.2.2:

The Markov chain $\{(Z_n,K_n)\ ,\ n = 1,2,\ ...\}$ has a stationary distribution.

Proof:

Let r_{ik} , $i = 1,2$, $k = 1,2,\ ...$, be the probability that the total number of arrivals in a busy period, which starts when the phase of the system is in i , is k . Then, it is easy to check that

$$P\{(Z_n,K_n) = (i,k)\} = \pi_i r_{ik}\ ,\ i = 1,2\ ,\ k = 1,2,\ ...$$

is a stationary distribution of $\{(Z_n,K_n)\ ,\ n = 1,2,\ ...\}$. ■

Lemma 4.2.3:

If K_1 has the distribution

$$P\{K_1 = k\} = \pi_1 r_{1k} + \pi_2 r_{2k}\ ,\ k = 1,2,\ ...\ ,$$

then $\{K_n\ ,\ n = 1,2,\ ...\}$ is a stationary ergodic process.

Proof:

Since a stationary distribution for the Markov chain $\{(Z_n,K_n)\ ,\ n = 1,2,\ ...\}$ exists and the chain is indecomposable, it follows from Theorem 7.16 of Breiman [3] that the process $\{(Z_n,K_n)\ ,\ n = 1,2,\ ...\}$ with the stationary initial distribution given in Lemma 4.2.2 is ergodic. Therefore, $\{K_n\ ,\ n = 1,2,\ ...\}$ is also a stationary ergodic process since K_n is just a coordinate mapping of (Z_n,K_n) (see Theorem 6.31 of Breiman). ■

Now we are ready for the main

Theorem 4.2.4:

$$L(c) = 1 - \frac{1}{E(K_1)}$$

where

$$E(K_1) = 1 + \pi_1 \int_0^\infty \int_0^t [\lambda_1 P_{11}(y) + \lambda_2 P_{12}(y)]dydG_1(t)$$

(4.2.6)

$$+ \pi_2 \int_0^\infty \int_0^t [\lambda_1 P_{21}(y) + \lambda_2 P_{22}(y)]dydG_2(t) .$$

Proof:

Observe that in each busy period, only the first customer can enter the system and receive service, the rest of them are lost. Hence the proportion of customers lost in n busy period, $L_n(c)$, is given by

$$L_n(c) = \frac{\sum_{i=1}^{n} K_i - n}{\sum_{i=1}^{n} K_i}$$

$$= 1 - \frac{n}{\sum_{i=1}^{n} K_i}$$

$$= 1 - \frac{1}{\sum_{i=1}^{n} K_i/n} .$$

Letting n goes to infinity, we get the long run proportion of customers lost which is also the stationary probability that an arriving customer will be lost

$$L(c) = \lim_{n\to\infty} L_n(c)$$

$$= 1 - \frac{1}{\lim\limits_{n\to\infty} \sum\limits_{i=1}^{n} K_i/n}$$

$$= 1 - \frac{1}{E(K_1)} .$$

The last equality follows because, without loss of generality, we can assume that $\{K_n , n = 1,2, \ldots\}$ is a stationary ergodic process by letting K_1 have the stationary distribution given in Lemma 4.2.3, and hence, by ergodic theorem, (e.g., Theorem 6.28 of [3])

$$\lim_{n\to\infty} \frac{\sum\limits_{i=1}^{n} K_i}{n} \to E(K_1) \quad \text{W.P.1.}$$

So, we only have to compute $E(K_1)$.

Let N_i , $i = 1,2$, be the total number of customers who arrive in a busy period with length T and starting with the system in phase i . For $i = 1,2$, $j = 1,2$, define

$$I_{ij}(y) = \begin{cases} 1 & \text{if the phase of the system at time } y \text{ is } j \text{ , given that at time } 0 \text{ it is } i \\ 0 & \text{otherwise,} \end{cases}$$

then $T_{ij} = \int_0^T I_{ij}(y)dy$ is the amount of time the system spends in phase j in $[0,T]$ if it starts from phase i. Therefore,

$$E(N_i) = \lambda_1 E(T_{i1}) + \lambda_2 E(T_{i2})$$

$$= \lambda_1 E\left[\int_0^T I_{i1}(y)dy\right] + \lambda_2 E\left[\int_0^T I_{i2}(y)dy\right]$$

$$= \lambda_1 \int_0^T P_{i1}(y)dy + \lambda_2 \int_0^T P_{i2}(y)dy \ .$$

Hence (4.2.6) follows by conditioning on the length and the phase of the system at the beginning of a busy period.■

4.3 Special Cases

Case 1: $c \to \infty$

Let G_1 and G_2 have the same mean $\frac{1}{\mu}$. As we mentioned earlier in Section 4.1, the parameter c in this model controls the rate at which the system switches between phases. And, as c goes to infinity, the arrival process converge to a stationary Poisson process with rate $\bar{\lambda}$. Indeed, from (4.2.6), we have

$$\lim_{c\to\infty} E(K_1) \to 1 + \frac{\bar{\lambda}}{\mu}$$

or equivalently,

$$\lim_{c\to\infty} L(c) = \frac{\bar{\lambda}}{\bar{\lambda} + \mu}$$

which is exactly the loss formula for a no queue allowed M/G/1 loss system with arrival rate $\bar{\lambda}$ and service rate μ .

Case 2: $\lambda_2 = 0$

The expression we obtained for $L(c)$ is, unfortunately, too complicated to determine whether it is a decreasing function of c . Some numerical experiences indicate that $L(c)$ is decreasing for Erlang, exponential and hyperexponential service distributions.

One interesting special case is to let $\lambda_2 = 0$. Note that, in this case, the arrival process becomes a renewal process, and is sometimes called an interrupted Poisson process in teletraffic terminology. It arises naturally as an over flow process from some queueing loss systems. Letting $\lambda_2 = 0$ in (4.2.6), we have

$$(4.3.1) \quad E(K_1) = 1 + \left(\frac{\alpha_1\lambda_1}{c(\alpha_1 + \alpha_2)^2}\right) \cdot [1 - \phi_1[c(\alpha_1 + \alpha_2)]] + \frac{\bar{\lambda}}{\mu}$$

which may be shown by differentiating to be a decreasing function of c . Thus, $L(c)$ is also a decreasing function of c . Hence, in this system, with a given service distribution G_1 , the more stationary the arrival process is the smaller $L(c)$ will be.

For a fixed c , we can also compare $L(c)$ for two such systems with different service distributions H_1 and H_2 . Suppose H_1 and H_2 have the same mean and

$$(4.3.2) \quad \int_0^\infty e^{-st} dH_1(t) \leq \int_0^\infty e^{-st} dH_2(t) ,$$

i.e., the Laplace transforms of H_1 and H_2 are ordered. Then, (4.3.1) and (4.3.2) imply that

$$L^1(c) \geq L^2(c) ,$$

where $L^i(c)$, $i = 1,2$, is the loss formula for the system with service distribution H_i .

One sufficient condition for (4.3.2) to hold is $H_1 \overset{v}{\leq} H_2$ and $\int x dH_1(x) = \int x dH_2(x)$ since e^{-st} is a convex function of t . Therefore, for interrupted Poisson arrival process, the more regular the service distribution is the "worse" the system performance will be for a single server loss system. This is perhaps a surprising phenomenon and we will look into this further in the next section.

4.4 Some Related Results

Regularity of arrival process and service times seems to work to good effect with respect to the usual performance measures in a number of queueing systems. Therefore, it is natural to ask: Does this general principle holds for loss systems?

We consider two types of comparisons. One can either compare two systems with different service distributions for fixed arrival process, or, for fixed service distribution, compare two systems with different interarrival distributions. Counterexamples of both types will be presented below. Hence, the answer to the above question is: no.

Consider a single server loss system. All arriving customers who find the system empty are lost. Let the service times be

deterministic, say 10 minutes for each customer. Compare two such systems: one has i.i.d. interarrival times T_i , $i = 1,2, \ldots$, such that

$$T_i = \begin{cases} 7 \text{ minutes} & \text{w.p. } 1/2 \\ 11 \text{ minutes} & \text{w.p. } 1/2 \end{cases},$$

and one with deterministic interarrival times with $T_i \equiv 9$ minutes for all $i = 1,2, \ldots$. Then, it is easy to see that

Proportion of customers lost in first system = 1/3

Proportion of customers lost in second system = 1/2 .

Thus, the second system which has more regular interarrival times behaves "worse" than the first system.

Another example in the same vein is obtained by reversing the roles of interarrival and service time distributions in the above. Consider two loss systems with deterministic interarrival times, say 10 minutes for each customer to arrive. Let the service times S_i , $i = 1,2, \ldots$ for customers in the first system be such that

$$S_i = \begin{cases} 9 \text{ minutes} & \text{w.p. } 1/2 \\ 13 \text{ minutes} & \text{w.p. } 1/2 \end{cases},$$

and $S_i \equiv 11$ for the second system. Again, we have

Proportion of customers lost in first system = 1/3

Proportion of customers lost in second system = 1/2 .

Hence, the system with more regular service times has a "larger" loss. Note that the system discussed at the end of the last section also exhibits the same contrary behavior, but the example here is simpler.

The next question is: Is there any condition under which a loss system with more regular interarrival and service distributions does perform better? Some partial answers to this question for single server loss systems are given below.

Consider two single server loss systems $F_1/G/1$ and $F_2/G/1$ where F_1 , F_2 are interarrival distributions and G is the service distribution for both systems. Let $L_1(L_2)$ be the proportion of customers lost in system $F_1/G/1$ $(F_2/G/1)$.

Theorem 4.4.1:

Suppose $\int_0^\infty x dF_1(x) = \int_0^\infty x dF_2(x) = 1/\lambda$, $F_1(F_2)$ has NBUE (NWUE) distribution, and $\int_0^\infty x dG(x) = 1/\mu$. Then,

$$L_1 \leq \frac{\lambda}{\lambda + \mu} \leq L_2$$

where $\frac{\lambda}{\lambda + \mu}$ is the corresponding loss formula for a M/G/1 loss system with arrival rate λ and service rate μ .

Proof:

Let $m_i(t)$ be the expected number of arrivals in $[0,t]$ for the i^{th} system, $i = 1,2$. From Theorem 4.2.4, we have

$$L_i = 1 - \frac{1}{E(C_1)} \ , \ i = 1,2$$

where $E(C_i) = 1 + \int_0^\infty m_i(t)dG(t)$ is the expected number of arrivals in a busy period.

Hence, we only have to show

$$(4.4.1) \qquad \int_0^\infty m_1(t)dG(t) \leq \lambda t \leq \int_0^\infty m_2(t)dG(t) \ .$$

But, (4.4.1) follows directly from a well-known fact from renewal theory that $m_1(t) \leq \lambda t \leq m_2(t)$. ■

Theorem 4.4.2:

Suppose $\int_0^\infty e^{-sx}dF_1(x) \leq \int_0^\infty e^{-sx}dF_2(x)$, $\int_0^\infty xdF_1(x) = \int_0^\infty xdF_2(x)$, and G has hyperexponential distribution. Then,

$$L_1 \leq L_2 \ .$$

Proof:

Again, we only have to prove

$$\int_0^\infty m_1(t)dG(t) \leq \int_0^\infty m_2(t)dG(t) \ .$$

Let $\phi_i(x) = \int_0^\infty e^{-sx}dF_i(x)$, $i = 1,2$. Now, since $m_i(t) = \sum_{n=1}^\infty F_i^{[n]}(t)$, we have

$$\tilde{m}_i(s) = \frac{\phi_i(s)}{1 - \phi_i(s)} , \quad i = 1,2 ,$$

where $\tilde{m}_i(s)$ is the Laplace transform of $m_i(t)$. Therefore, $\phi_1(s) \leq \phi_2(s) \Leftrightarrow \tilde{m}_1(s) \leq \tilde{m}_2(s)$. Integrating by part, we get

$$\tilde{m}_i(s) = \int_0^\infty e^{-st} dm(t)$$

$$= \int_0^\infty m_i(t) s e^{-st} dt .$$

Thus, $\int_0^\infty m_1(t) s e^{-st} dt \leq \int_0^\infty m_2(t) s e^{-st} dt$ for all s.

Notice that se^{-st} is exactly the density of an exponential distribution with rate s. It follows that

$$\int_0^\infty m_1(t) dG(t) \leq \int_0^\infty m_2(t) dG(t)$$

if G is a mixture of exponential distributions. ■

REFERENCES

[1] Barlow, R. E. and F. Proschan, STATISTICAL THEORY OF RELIABILITY AND LIFE TESTING: PROBABILITY MODELS, Holt, Rinehart, and Winston, New York, 1975.

[2] Bessler, S. A. and A. F. Veinott, Jr., "Optimal Policy for a Dynamic Multi-Echelon Inventory Model," Naval Research Logistics Quarterly, Vol. 13, pp. 355-389, 1966.

[3] Breiman, L., PROBABILITY, Addison-Wesley, Reading, Mass., 1968.

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

[5] Burke, P. J., "The Output of a Queueing System," Operations Research, Vol. 4, pp. 699-704, 1956.

[6] Burke, P. J., "The Output Process of a Stationary M/M/s Queueing System," Annals of Mathematical Statistics, Vol. 39, pp. 1144-1152, 1968.

[7] Esary, J. D., F. Proschan and D. W. Walkup, "Association of Random Variables, with Applications," Annals of Mathematical Statistics, Vol. 38, pp. 1466-1474, 1967.

[8] Friedman, H. D., "Reduction Methods for Tandem Queueing Systems," Operations Research, Vol. 13, pp. 121-131, 1965.

[9] Kingman, J. F. C., "On Queues in Heavy Traffic," Journal of Royal Statistical Society, B 24, pp. 383-392, 1962.

[10] Kingman, J. F. C., "Some Inequalities for the GI/G/1 Queue," Biometrika, Vol. 49, pp. 315-324, 1962.

[11] Kuzura, A., "Loss Systems with Mixed Renewal and Poisson Inputs," Operations Research, Vol. 21, No. 3, pp. 787-795, 1973.

[12] Kuzura, A., "The Interrupted Poisson Process as an Overflow Process," Bell System Technical Journal, Vol. 51, pp. 437-448, 1973.

[13] Marshall, K. T., "Some Inequalities in Queueing," Operations Research, Vol. 16, pp. 651-665, 1968.

[14] Marshall, K. T. and R. W. Wolff, "Customer Average and Time Average Queue Lengths and Waiting Times," Journal of Applied Probability, Vol. 8, No. 3, pp. 535-542, 1971.

[15] Marshall, A. W. and F. Proschan, "Classes of Distributions Applicable in Replacement, with Renewal Theory Implications," Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, pp. 395-415, University of California Press, 1972.

[16] Reich, E., "Waiting Times when Queues are in Tandem," Annals of Mathematical Statistics, Vol. 28, pp. 768-773, 1957.

[17] Reich, E., "Note on Queues in Tandem," Annals of Mathematical Statistics, Vol. 34, pp. 338-341, 1963.

[18] Rolski, T. and D. Stoyan, "On the Comparison of Waiting Times in GI/G/1 Queues," Operations Research, Vol. 24, pp. 197-200, 1976.

[19] Ross, S. M., "Average Delay in Queues with Nonstationary Poisson Arrivals," ORC 77-13, Operations Research Center, University of California, Berkeley, California, 1977.

[20] Ross, S. M., APPLIED PROBABILITY MODELS WITH OPTIMIZATION APPLICATION, Holden Dav, San Francisco, California, 1970.

[21] Stoyan, H. and D. Stoyan, "Monotonieeigenschaften der Kundenwartezeiten im Modell GI/G/1," Z. Angew. Math. Vol. 49, pp. 729-734, 1969.

[22] Fond, S. and S. M. Ross, "A Heterogeneous Arrival and Service Queueing Loss Model," ORC 77-12, Operations Research Center, University of California, Berkeley, California, 1977.

[23] Tembe, S. V., "Effect of Order of Servers on Tandem Queues," ORC 71-32, Operations Research Center, University of California, Berkeley, California, 1971.

[24] Tembe, S. V. and R. W. Wolff, "Optimal Order of Service in Tandem Queues," Operations Research, Vol. 22, No. 4, pp. 824-832, 1974.

[25] Wolff, R. W., "The Effect of Service Time Regularity on System Performance," ORC 77-7, Operations Research Center, University of California, Berkeley, California, 1977.