



On the Comparison of Waiting Times in Tandem Queues

Shun-Chen Niu

Journal of Applied Probability, Vol. 18, No. 3 (Sep., 1981), 707-714.

Stable URL:

<http://links.jstor.org/sici?sici=0021-9002%28198109%2918%3A3%3C707%3AOTCOWT%3E2.0.CO%3B2-V>

Journal of Applied Probability is currently published by Applied Probability Trust.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/apt.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

ON THE COMPARISON OF WAITING TIMES IN TANDEM QUEUES

SHUN-CHEN NIU,* *The Cleveland State University*

Abstract

Using a definition of partial ordering of distribution functions, it is proven that for a tandem queueing 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 and service distributions becomes smaller with respect to that ordering. Some stronger conclusions are also given under stronger order relations. Using these results, bounds for the expected total waiting time in system are then readily obtained for wide classes of tandem queues.

TANDEM QUEUES; TOTAL WAITING TIMES; PARTIAL ORDERINGS; BOUNDS

1. Introduction

Bounds and approximations are important areas of research in queueing theory because many practical queueing models are still analytically intractable. One method of finding bounds for queueing systems is to look for monotonicity properties among various systems such that the performance measures of difficult systems may be bounded by that of simpler ones. A considerable amount of literature has been devoted to this direction of research. For a survey of recent work in this area, see Stoyan (1977b) and the references there. The main purpose of this paper is to extend an important comparison theorem for $GI/G/1$ queues (see Rolski and Stoyan (1976) and Stoyan and Stoyan (1969)) to some tandem queues.

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 are other customers present at the time of his arrival, he joins the end of the queue and

Received 15 April 1980; revision received 14 July 1980.

* Present address: School of Management and Administration. The University of Texas at Dallas, Box 688, Richardson, TX 75080, U.S.A.

Research 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.

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 types of service in general. Every customer has to go through all stations according to prefixed order and leaves the system after finishing service at the last station. We shall denote such a system with n stations in tandem by $GI/G_1/1 \rightarrow G_2/1 \rightarrow \cdots \rightarrow G_n/1$, where $G_i, i = 1, 2, \dots, n$, is the service-time distribution for station i , GI means the arrival process of customers to the first station is a renewal process, and there is one server at each station. Similar notations will be used throughout.

Following Stoyan (1977a) consider the following partial orderings of distribution functions.

Definition. For two distribution functions F and G , we say $F \stackrel{(1)}{\cong} G$ iff $\int f(x)dF(x) \leq \int f(x)dG(x)$ for all non-decreasing functions f , and $F \stackrel{(2)}{\cong} G$ iff $\int f(x)dF(x) \leq \int f(x)dG(x)$ for all non-decreasing convex functions f . For two random variables X and Y , we say $X \stackrel{(i)}{\cong} Y, i = 1, 2$, if their distribution functions satisfy the above orderings.

The ordering $\stackrel{(1)}{\cong}$ is usually called stochastic ordering. The ordering $\stackrel{(2)}{\cong}$ is called ‘stochastically smaller in mean’ by Bessler and Veinott (1966). It orders the relative variabilities of two distribution functions (or random variables) when they have the same mean. The following useful comparison theorem was proven by Stoyan and Stoyan (1969).

Theorem 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 $G_i, i = 1, 2$. Assume the systems start operation at time 0 with no customers present. If $\int x dF_1(x) = \int x dF_2(x)$, $F_1 \stackrel{(2)}{\cong} F_2$, and $G_1 \stackrel{(2)}{\cong} G_2$, then $D_k^1 \stackrel{(2)}{\cong} D_k^2$, where $D_k^i, i = 1, 2, k = 1, 2, \dots$, is the delay of the k th customer in system $F_i/G_i/1$.

The main results in this paper are extensions of Stoyan’s theorem to some tandem queueing systems. It is shown in Section 2 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 smaller with respect to partial ordering $\stackrel{(2)}{\cong}$. 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 in Section 3.

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 service distribution for both systems, we prove in Section 4 that the system with stochastically larger interarrival times and smaller service times with respect to partial ordering $\stackrel{(i)}{\cong}$, $i = 1, 2$, at the second station has smaller delay with respect to partial ordering $\stackrel{(i)}{\cong}$ for every customer in front of the second station.

2. Comparison of waiting times in tandem queues

We now proceed to generalize Theorem 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 $G_j, j = 1, 2, \dots, n$. Let

T_k = interarrival time between the $(k - 1)$ th and k th customer,
 $k = 1, 2, \dots$.

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

Z_k^j = epoch at which the k th customer leaves station j ,

$$j = 1, 2, \dots, n, k = 1, 2, \dots. \quad Z_k^0 \equiv \sum_{i=1}^k T_i.$$

We make the assumption that the system starts operation at time 0 with no customer in the system, and customers are served in the order of their arrival at each station. The following lemmas are essential in the proofs below.

Lemma 1 (Bessler and Veinott (1966), Theorem 14). Let $X_i, Y_i, i = 1, 2, \dots, n$, be independent random variables. Then, $X_i \stackrel{(2)}{\cong} Y_i$ for all $i = 1, 2, \dots, n$, if and only if $f(X_1, X_2, \dots, X_n) \stackrel{(2)}{\cong} f(Y_1, Y_2, \dots, Y_n)$ for all increasing convex functions f .

Proof. The proof is straightforward by induction, noting that increasing convex functions of increasing convex functions are still increasing convex.

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

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

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

Since $Z_k^0 = \sum_{i=1}^k T_i$, $k = 1, 2, \dots$, 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 the tilde (e.g. \tilde{Z}_k) denote the corresponding quantities for system Δ_2 . It follows from Lemmas 1 and 2 that

$$(1) \quad Z_k^i \stackrel{(2)}{\cong} \tilde{Z}_k^i \quad \text{for all } j = 1, 2, \dots, n, \quad k = 1, 2, \dots,$$

if $F \stackrel{(2)}{\cong} \tilde{F}$ and $G_j \stackrel{(2)}{\cong} \tilde{G}_j$ for all $j = 1, 2, \dots, n$.

Letting $W_k (\tilde{W}_k)$ be the total waiting time of the k th customer in system $\Delta_1 (\Delta_2)$, we have the following results.

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

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 subtracting, we have

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

where the last inequality follows from (1).

Theorem 3. If $F(x) = \tilde{F}(x)$ for all $x \geq 0$ and $G_j \stackrel{(2)}{\cong} \tilde{G}_j$ for all $j = 1, 2, \dots, n$, then $W_k \stackrel{(2)}{\cong} \tilde{W}_k$ for all $k = 1, 2, \dots$.

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

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

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

Removing the conditioning, we have $E[f(W_k)] \leq E[f(\tilde{W}_k)]$.

Remarks. (i) In the proof of Theorem 3, we only needed the assumption that the sequence $\{t_i\}_{i=1}^\infty$ has the same joint distribution for both systems. Therefore, the interarrival times may be arbitrary.

(ii) Clearly, Theorem 3 is true when $\stackrel{(2)}{\cong}$ is replaced by $\stackrel{(1)}{\cong}$ and $F(x) = \tilde{F}(x)$ is replaced by $F(x) \leq \tilde{F}(x)$.

Another interesting question one might ask is: Can Theorem 1 be generalized to parallel-server queues? Ross (1978) gave a counterexample showing that similar comparisons do not hold in general. However, it may be shown that Theorem 1 can be generalized to the case of m constant servers in parallel (see

Stoyan (1977a), p. 854, or Niu (1977), p. 36). It should be noted that in case of constant service times the ordering $\stackrel{(2)}{\cong}$ is simply ordering by size. Combining Theorems 2, 3 and the above, we have the following.

Theorem 4. 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 Theorems 2 and 3, 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 as in the proofs of Theorems 2 and 3.

3. 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.

Example 1. Consider the following four systems of tandem queues:

- (A) $M/A_1/1 \rightarrow A_2/1 \rightarrow \dots \rightarrow A_n/1$
- (B) $M/B_1/1 \rightarrow B_2/1 \rightarrow \dots \rightarrow B_n/1$
- (C) $M/C_1/1 \rightarrow C_2/1 \rightarrow \dots \rightarrow C_n/1$
- (D) $M/D_1/1 \rightarrow D_2/1 \rightarrow \dots \rightarrow D_n/1$

where

- $A_i, i = 1, 2, \dots, n$ have NWUE distributions,
- $B_i, i = 1, 2, \dots, n$ have exponential distributions,
- $C_i, i = 1, 2, \dots, n$ have NBUE distributions,
- $D_i, i = 1, 2, \dots, n$ are deterministic.

Assume that A_i, B_i, C_i, D_i have the same mean for all $i = 1, 2, \dots, n$ and the arrival processes are Poisson with the same rate for all four systems. It is well known (see Stoyan (1977a), p. 853) that $D_i \stackrel{(2)}{\cong} C_i \stackrel{(2)}{\cong} B_i \stackrel{(2)}{\cong} A_i$ for all $i = 1, 2, \dots, n$.

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, we have $W_D \stackrel{(2)}{\cong} W_C \stackrel{(2)}{\cong} W_B \stackrel{(2)}{\cong} W_A$. 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 $E(W'_C)$ and a lower bound for $E(W'_A)$ for all real numbers $r \geq 1$.

Example 2. Tembe and Wolff (1974) showed that the expected delay in front of the second server in an $M/D/1 \rightarrow M/1$ (the first server has deterministic service times) system is bounded above by $\lambda/[\mu(\mu - \lambda)]$ where λ is the arrival rate and μ is the service rate at the second station.

Theorem 3 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 4.5 in Niu (1980), we have

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

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

Remark. A lower bound for the expected delay at station j , denoted by d_j , in $GI/G_1/1 \rightarrow G_2/1 \rightarrow \dots \rightarrow G_n/1$ system is

$$\frac{\lambda\sigma_j^2}{2(1 - \rho_j)} - \frac{m_j}{2},$$

where σ_j^2 is the variance of G_j , m_j is the mean of G_j , and $\rho_j = \lambda m_j$. See Formula (5.7.1) on p. 90 of Stoyan (1977b).

4. 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 tandem queues $F/G/1 \rightarrow H/1$ and $\bar{F}/G/1 \rightarrow \bar{H}/1$ 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 $F/G/1 \rightarrow H/1$, let D_n (D_n^*) 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 between the n th and $(n + 1)$ th customer. Let all barred notations (e.g., \bar{D}_n^*) denote the corresponding quantities in system $\bar{F}/G/1 \rightarrow \bar{H}/1$.

It is not difficult to show (see Niu (1980)) that

$$(2) \quad D_{n+1}^* = \max[0, \min[D_n + S_n + D_n^* + R_n - T_n - S_{n+1}, D_n^* + R_n - S_{n+1}]]$$

and

$$(3) \quad \bar{D}_{n+1}^* = \max[0, \min[\bar{D}_n + \bar{S}_n + \bar{D}_n^* + \bar{R}_n - \bar{T}_n - \bar{S}_{n+1}, \bar{D}_n^* + \bar{R}_n - \bar{S}_{n+1}]].$$

We need the following well-known lemma.

Lemma 3. Let $X_i, Y_i, i = 1, 2, \dots, n$, be independent random variables. Then $X_i \stackrel{(1)}{\leq} Y_i$ for all $i = 1, 2, \dots, n$ if and only if

$$f(X_1, X_2, \dots, X_n) \stackrel{(1)}{\leq} f(Y_1, Y_2, \dots, Y_n)$$

for all non-decreasing functions f .

Theorem 5. If $F \stackrel{(1)}{\geq} \bar{F}$ and $H \stackrel{(1)}{\leq} \bar{H}$, then $D_k^* \stackrel{(1)}{\leq} \bar{D}_k^*$ for all $k = 1, 2, \dots$.

Proof. Conditioning on the sequence of service times, $\{S_j\}_{j=1}^k$, at the first station, it follows from (2) and (3) that $D_k^*(\bar{D}_k^*)$ are increasing functions of $(-T_1, \dots, -T_{k-1}, R_1, \dots, R_{k-1})((-\bar{T}_1, \dots, -\bar{T}_{k-1}, \bar{R}_1, \dots, \bar{R}_{k-1}))$ for all $k = 1, 2, \dots$. Hence by Lemma 3,

$$[D_k^*|\{S_j\}_{j=1}^k] \stackrel{(1)}{\leq} [\bar{D}_k^*|\{S_j\}_{j=1}^k] \text{ for all } k = 1, 2, \dots.$$

To complete the proof we remove the conditioning.

Theorem 6. If $F \stackrel{(1)}{\geq} \bar{F}$ and $H \stackrel{(2)}{\leq} \bar{H}$, then $D_k^* \stackrel{(2)}{\leq} \bar{D}_k^*$ for all $k = 1, 2, \dots$.

Proof. Let f be an arbitrary increasing convex function. Denote $F_{T_1, \dots, T_{k-1}}, F_{\bar{T}_1, \dots, \bar{T}_{k-1}}$, and G_{S_1, \dots, 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

$$\begin{aligned} E\{f(D_k^*)\} &= \int \int E\{f(D_k^*)|\{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^k\} \\ & dF_{T_1, \dots, T_{k-1}}(t_1, \dots, t_{k-1}) dG_{S_1, \dots, S_k}(s_1, \dots, s_k) \\ & \leq \int \int E\{f(\bar{D}_k^*)|\{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^k\} \\ & dF_{T_1, \dots, T_{k-1}}(t_1, \dots, t_{k-1}) dG_{S_1, \dots, S_k}(s_1, \dots, s_k) \\ & \leq \int \int E\{f(\bar{D}_k^*)|\{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^k\} \\ & dF_{\bar{T}_1, \dots, \bar{T}_{k-1}}(t_1, \dots, t_{k-1}) dG_{S_1, \dots, S_k}(s_1, \dots, s_k) \\ & = E\{f(\bar{D}_k^*)\}, \end{aligned}$$

where the first inequality follows since $[f(D_k^*)|\{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^k]$ and $[f(\bar{D}_k^*)|\{t_i\}_{i=1}^{k-1}, \{s_i\}_{i=1}^k]$ are increasing convex functions of (R_1, \dots, R_{k-1}) and $(\bar{R}_1, \dots, \bar{R}_{k-1})$ respectively (see (2) and (3)), and the second inequality follows from Lemma 3.

Remarks.

(i) In the proofs of Theorems 5 and 6, we needed only $(T_1, \dots, T_{k-1}) \stackrel{(1)}{\cong} (\bar{T}_1, \dots, \bar{T}_{k-1})$ for all $k = 1, 2, \dots$ where $(T_1, \dots, T_{k-1}) \stackrel{(1)}{\cong} (\bar{T}_1, \dots, \bar{T}_{k-1})$ if $f(T_1, \dots, T_{k-1}) \stackrel{(1)}{\cong} f(\bar{T}_1, \dots, \bar{T}_{k-1})$ for all non-decreasing functions f . Hence the arrival processes may be arbitrary. Also, only $(R_1, \dots, R_{k-1}) \stackrel{(1)}{\cong} (\bar{R}_1, \dots, \bar{R}_{k-1})$ for all $k = 1, 2, \dots$ was used in Theorem 5.

(ii) Analogues of Theorems 5 and 6 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.

(iii) For some related discussion of monotonic properties of the output of queues, see Stoyan and Stoyan (1976).

References

- BESSLER, S. A. AND VEINOTT, A. F., JR (1966) Optimal policy for a dynamic multi-echelon inventory model. *Naval Res. Logist. Quart.* **13**, 355–389.
- NIU, S. C. (1977) Bounds and comparisons for some queueing systems. ORC 77–32, Operations Research Center, University of California, Berkeley.
- NIU, S. C. (1980) Bounds for the expected delays in some tandem queues. *J. Appl. Prob.* **17**, 831–838.
- ROLSKI, T. AND STOYAN, D. (1976) On the comparison of waiting times in $GI/G/1$ queues. *Operat. Res.* **24**, 197–200.
- ROSS, S. M. (1978) Average delay in queues with nonstationary Poisson arrivals. *J. Appl. Prob.* **15**, 602–609.
- STOYAN, D. (1977a) Bounds and approximations in queueing through monotonicity and continuity. *Operat. Res.* **24**, 851–863.
- STOYAN, D. (1977b) *Qualitative Eigenschaften und Abschätzungen Stochastischer Modelle*. Akademie-Verlag, Berlin.
- STOYAN, D. AND STOYAN, H. (1969) Monotonieigenschaften der Kundenwartezeiten im Modell $GI/G/1$. *Z. Angew. Math.* **49**, 729–734.
- STOYAN, D. AND STOYAN, H. (1976) Some qualitative properties of single server queues. *Math. Nachr.* **70**, 29–34.
- TEMBE, S. V. AND WOLFF, R. W. (1974) Optimal order of service in tandem queues. *Operat. Res.* **22**, 824–832.