Journal of the Operations Research
Society of Japan
Vol. 31, No. 2, June 1988

MEAN SOJOURN TIMES IN A MULTI-STAGE TANDEM QUEUE SERVED BY A SINGLE SERVER

Tsuyoshi Katayama
NTT Communication Switching Laboratories

(Received July 24, 1987; Revised November 24, 1987)

Abstract This paper presents explicit expressions for the mean sojourn times in a multi-stage tandem queue, with Poisson arrival and general service times, served by a single server with a cyclic switching rule, $M/G_1-G_2-\cdots-G_N/1$. The expressions are derived using the results of the two-stage tandem queue, $M/G_1-G_2/1$, already considered by some authors and the well-known Pollaczek-Khintchine formula. With the first-in, first-out service discipline, the mean waiting times in the first stage are derived for some switching rules and upper and lower bounds for the mean sojourn times and the mean waiting times for workload conserving switching rules are also obtained.

1. Introduction

There are several practical examples of tandem queueing systems served by a single server: a labor and machine limited production system [1], a repairable system with a repair man [2], and operating systems in computer and telephone switching systems [6]-[9]. T. Katayama [6]-[8] dealt with a tandem queueing model served by a moving server for call processing in a single processor switching system. A moving server, which corresponds to a single processor control point, visits each processing program, i.e. input processing, internal processing, output processing and so on, in accordance with a switching rule given by a scheduling table. Call (Customer) services in some tandem queues are performed by a moving server. (The queues may be called hoppers and buffers.)

There have been some analitical studies of these models of tandem queues. M. T. Netto [2] considered the steady-state probabilities in a two-stage tandem queue with Poisson arrival and general service times, denoted by $M/G_1-G_2/1$ for simplicity in this paper, which is served according to a zero switching rule, i.e. the server stays in a stage until its queue becomes empty and

then it switches to the other stage. S. S. Nair analyzed the transient behavior of the tandem queue (M/G_1-$G_2/1$) with a zero switching rule [3][4] and with a non-zero switching rule [5] (See Section 2). T. Katayama considered the steady-state probabilities in the tandem queue (M/G_1-$G_2/1$) with gated service [6][7] (see Section 2), and he also investigated the influence of the server's walking time (sometimes known as switchover time or overhead time) on the mean sojourn time in a tandem queue with a non-zero switching rule [8].

These authors were concerned only with the two-stage tandem queue (M/G_1-$G_2/1$) served by a single server [1]-[8]. To author's best knowledge, there are only a few studies treating the multi-stage tandem queue with a zero switching rule: the three-stage tandem queue (M/G_1-G_2-$G_3/1$) by M. Murakami et al. [9], the N-stage tandem queue (M/G_1-G_2-$\cdots$-$G_N/1$) by T. Nishida et al. [10] and D. König et al. [11] (E. G. Enns considered the multiple feedback queue with priorities [12] similar to the N-stage tandem queue.)

The main object of this paper is to provide expressions for the mean sojourn time, that is, the mean total time spent by a customer in the N-stage tandem queue, and to obtain upper and lower bounds for this expression in the tandem queue (M/G_1-G_2-$\cdots$-$G_N/1$) with some switching rules, including all the above switching rules, and also the expressions for the mean waiting times for receiving the service in the first stage. (In practical situations, the mean waiting time is also important since this means the waiting time for input processing in the call processing in some switching systems [6]-[8].)

Section 2 of this paper describes in detail an N-stage tandem queue served by a single server, and certain switching rules. Section 3 is devoted to the derivation of expressions for the mean sojourn times and the mean waiting times in the N-stage tandem queue using the results of the two-stage tandem queue analysis and the well-known Pollaczek-Khintchine formula. Section 4 gives upper and lower bounds for the mean sojourn times and the mean waiting times for a general switching rule. Section 5 summarizes the paper.

2. Multi-Stage Tandem Queueing Model

This section presents the tandem queueing model served by a single server in detail, together with definitions of notations.

The queueing system has N ($<\infty$) service stages connected in series. The i-th stage has a service counter S_i and a queue with infinite capacity, Q_i, $i=1,2,\ldots,N$. Customers arrive at the first stage according to a Poisson process with rate λ. Each customer requires exactly N services before leaving the queueing system. That is, after completion of service in S_i, the customer

goes to Q_{i+1} to receive the service in S_{i+1}, $i=1,2,\ldots,N-1$, and after service completion in S_N, the customer leaves the system.

Customers in each queue are served in the order of their arrivals (FIFO). Service times τ_i at each counter S_i, $i=1,2,\ldots,N$ are independent and identically distributed random variables with a general distribution function $H_i(t)$, with finite first and second moments h_i and $h_i^{(2)}$. The Laplace-Stieltjes transform (LST) of $H_i(t)$ is denoted by $H^*_i(s)$, $i=1,2,\ldots,N$.

All the queues are served by a single server that moves among the counters according to a cyclic switching rule. That is, the server advances to the next counter in cyclic order, $S_1 \to S_2 \to \cdots \to S_N \to S_1 \to$, and so on. The following switching rules [13] will be considered for each counter.

(a) Exhaustive service, also called a zero switching rule (ZS): when the server visits a queue, its customers are served until that queue is empty. (This is denoted by ZS in this paper.)

(b) K-limited service, also called a non-zero switching rule (NZ): when the server visits a queue, it is served until either the queue becomes empty, or at most, a fixed number of customers, say K, are served, whichever occurs first.

(c) Gated service (GA): when the server visits a queue, the customers found upon arrival at the queue are served.

Remark 2.1: An example of the non-cyclic switching rule is as follows [14]:

(d) Deterministic switching rule (DT): For example, the order of switching is given by $S_1 \to S_2 \to \cdots \to S_N \to S_{N-1} \to S_{N-2} \to \cdots \to S_1 \to S_2 \to \cdots$.

In the above switching rules, the server's walking time needed to switch service from one counter to another is assumed to be zero. When no customers are present in the system, the server waits for a new arrival at S_1.

For simplicity, the utilization at Q_i and the mean total service time are denoted by the expressions

$$\rho_i := \lambda h_i, \quad i=1,2,\ldots,N, \quad h := \sum_{i=1}^{N} h_i. \tag{2.1}$$

Server utilization $\rho := \lambda h < 1$ is assumed in order to ensure that there are stationary distributions of all relevant queueing quantities.

The probability that m customers arrive at Q_1 during service times τ_i, $i=1,2,\ldots,N$ is denoted by $q^{(i)}_m$, and the generating function is denoted by $Q_i(x)$, i.e.,

$$(2.2)\qquad q^{(i)}_m := \int_0^\infty e^{-\lambda t}\frac{(\lambda t)^m}{m!}\,dH_i(t),\quad m=0,1,2,\ldots,$$

$$(2.3)\qquad Q_i(x) := \sum_{m=0}^{\infty} q^{(i)}_m x^m,\qquad i=1,2,\ldots,N.$$

Then, from (2.2) and (2.3)

$$(2.4)\qquad Q_i(x) = H^*_i\{\lambda(1-x)\}.$$

The following notations are used for a differentiable generating function $G_i(x,y)$,

$$(2.5)\qquad G'_{i,x}(a,b) := \left[\frac{\partial}{\partial x}G_i(x,y)\right]_{x=a,\ y=b},$$

$$G''_{i,y}(a,b) := \left[\frac{\partial^2}{\partial y^2}G_i(x,y)\right]_{x=a,\ y=b}.$$

3. Analysis of Mean Sojourn Times for Cyclic Switching Rules

This section considers an N-stage tandem queue model with cyclic switching rules. The mean sojourn time in the N-stage, $E(\Theta_N)$ and the mean waiting time in the first stage, $E(W_N)$, will be analysed using the results of two-stage tandem queue analysis.

3.1 Some results for queue length generating function

First, we will discuss the generating function of the joint queue length distribution. Denote by $\pi_n(i_1,i_2,\ldots,i_N)$ the steady-state probability that i_k customers are waiting in Q_k, $k=1,2,\ldots,N$, just after a customer has completed service at counter S_n, $n=1,2,\ldots,N$ and define the generating function by the expression

$$(3.1)\qquad G_n(X_1,X_2,\ldots,X_N) := \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty}\cdots\sum_{i_N=0}^{\infty}\pi_n(i_1,i_2,\ldots,i_N)X_1^{i_1}X_2^{i_2}\cdots X_N^{i_N},$$

$$n=1,2,\ldots,N,\quad |X_1|,|X_2|,\ldots,|X_N| \le 1.$$

To obtain simple expressions for $E(\Theta_N)$ and $E(W_N)$, the following lemma will be necessary.

Lemma 1. For the cyclic switching rule with a ZS (NZ,GA) scheme at counter S_n, $n=1,2,\ldots,N$, the following relationships hold:

$$(3.2)\qquad G_1(1,1,0,\ldots,0) = G_2(1,1,1,0,\ldots,0) = \cdots = G_N(1,0,\ldots,0,1) = \frac{1}{N},$$

$$(3.3)\qquad N\,G_1(X_1,X_2,0,\ldots,0) = 2\measuredangle\cdot G_1(X_1,X_2),$$

and

$$N\,G_N(1,0,\ldots,0,X) = 2\,\measuredangle\cdot G_2(1,X), \tag{3.3a}$$

where the capital letter $\measuredangle$ denotes an operator resulting from substituting $H_2(t)$ for $H_2*H_3*\ldots*H_N(t)$. (The symbol * denotes convolution.)

Proof: For the cyclic switching rule with a ZS (NZ,GA) scheme at counter S_n, $n=1,2,3,\ldots,N$, $G_n(X_1,X_2,\ldots,X_N)$ is equal to $G_n(X_1,0,\ldots,0,X_n,X_{n+1},0,\ldots,0)$ since no customers are present in $Q_2,\ldots,Q_{n-1}$ and $Q_{n+2},\ldots,Q_N$ when the server serves at S_n. Since exactly N services are required for each customer, the number of customers served at each counter S_n, $n=1,2,\ldots,N$ is also equal. Thus, the following relationship holds:

$$G_1(1,1,0,\ldots,0) = G_2(1,1,1,0,\ldots,0) = \cdots = G_N(1,0,\ldots,0,1). \tag{3.4}$$

Hence, using the normalization condition

$$\sum_{n=1}^{N} G_n(1,1,\ldots,1)=1, \tag{3.5}$$

(3.2) is obtained.

Next, a modified tandem queue with two stages, defined as follows, is introduced: Service time T_1 at the first counter has the same distribution as the original tandem queue. However, service time T_2 at the second counter is equal to the sum of $\tau_2,\tau_3,\ldots$ and τ_N. That is, $T_1:=\tau_1$, $T_2:=\tau_2+\tau_3+\cdots+\tau_N$. Everything else, e.g. arrival process, server switching rule and so on is exactly the same as in the original tandem queue. Hence, it is clear that in the original N-stage tandem queue and the modified tandem queue with two stages, both the first queue length distributions at the instant when the server has arrived at counter S_1 are equal. Since both the services at S_1 are performed in the same manner, both of the queue length distributions at service completion of a customer at S_1 are also equal. That is, the following equation holds:

$$\frac{G_1(X_1,X_2,0,\ldots,0)}{G_1(1,1,0,\ldots,0)} = \measuredangle\cdot\{\frac{G_1(X_1,X_2)}{G_1(1,1)}\}. \tag{3.6}$$

Thus, using (3.2), this gives (3.3).

Since no customer arrives at Q_n when the server serves at S_n, $n=2,3,\ldots,N$, the second queue length distribution just after switching to S_2 is equal to the $(i+1)$-th queue length distribution just after service completion of all customers in Q_i, $i=2,3,\ldots,N-1$. Hence, the following equations are derived:

$$\frac{G_2(1,x,1,0,\ldots,0)}{G_2(1,1,1,0,\ldots,0)} = \measuredangle\cdot\{\frac{G_2(1,x)}{G_2(1,1)}\}, \tag{3.7}$$

(3.8) $G_2(1,x,1,0,\ldots,0) = G_3(1,0,x,1,0,\ldots,0) = \cdots = G_N(1,0,\ldots,0,x)$.

Thus, using (3.2), (3.7) and (3.8) give (3.3a). □

The following lemma will also be necessary.

Lemma 2. For the cyclic switching rule with a ZS (NZ,GA) scheme at counter S_n, $n=1,2,\ldots,N$, the mean sojourn time, $E(\theta_N)$, is expressed by

$$E(\theta_N) = E(W)_{M/G/1} + h + \sum_{i=1}^{N-1} h_i \cdot E(Q_N^*), \tag{3.8}$$

where

$$E(W)_{M/G/1} := \frac{\lambda}{2(1-\rho)} \left[h^2 + \sum_{i=1}^{N} (h_i^{(2)} - h_i^2)\right], \tag{3.9a}$$

$E(Q_N^*)$: the mean queue length in Q_N at service completion at S_N.

Proof: Since services in each queue are performed by the first-in, first-out service discipline (FIFO), all customers in front of an arbitrary customer, C^*, must first be served before C^* leaves the system. Thus, the mean sojourn time, $E(\theta_N)$, of an arbitrary customer, C^*, is given by the summation of the following three terms [8][12]:

$$E(\theta_N) = E(W)+E(\tau)+E(U). \tag{3.10}$$

The first term, $E(W)$, denotes the mean total service time remaining for all of the customers in the queueing system on arrival of C^*. It is noticed that $E(W)$ is invariant for any switching rule included in the workload conserving switching rule [16] described in Section 4. Thus, using the Pollaczek-Khintchine formula [15] on the condition that the service time distribution function is $H_1 * H_2 * \cdots * H_N(t)$ gives (3.9a).

The second term, $E(\tau)$, denotes the mean total service time of C^* at S_1, $S_2,\ldots,$ and S_N. That is, $E(\tau) = h$.

The third term, $E(U)$, denotes the mean total service time at $S_1, S_2, \ldots,$ and S_{N-1}, for customers who have arrived at the system behind C^*, to receive their services at $S_1, S_2, \ldots,$ and S_{N-1} during the sojourn time of C^*. (Note that for a ZS (NZ,GA) scheme, the number of waiting customers in $Q_2, Q_3, \ldots,$ and Q_{N-1} is zero at the departure time of C^*.) Denoting the average number of customers in Q_N when C^* leaves S_N by $E(Q_N^*)$, gives the equation

$$E(U) = \sum_{i=1}^{N-1} h_i \cdot E(Q_N^*), \tag{3.11}$$

Since all customers in Q_N have already received services at $S_1, S_2, \ldots, S_{N-2}$ and S_{N-1}. Thus, (3.9) is obtained. □

Remark 3.1: The following switching rule is an example of the cyclic switching rule, for which Lemmas 1, 2 do not hold.
Probabilistic switching rule (PB): An example of this is the Bernoulli switching rule [14] which is parameterized by a vector $(p_1,p_2,\ldots,p_N)$. Here, at the completion of service at S_i, if Q_i is empty, then Q_{i+1} is served. Otherwise, Q_i is served again with probability p_i, and Q_{i+1} is served with probability $1-p_i$, $i=1,2,\ldots,N$.

Remark 3.2: For a tandem queue with server walking time, (3.10) seems to hold, giving a different definition of the third term $E(U)$. However, $E(W)$, i.e. the mean workload in the system at an arbitrary instant [15], is not given by such simple expression as $E(W)_{M/G/1}$. It also depends on the server switching rule.

3.2 Analysis for exhaustive service

This subsection considers an N-stage tandem queueing system with a ZS scheme, in which all the queues are served according to the exhaustive service. With the aid of Lemmas 1, 2, and the results of two-stage tandem queue analysis [2], the mean sojourn time, $E(\theta_N)_{ZS}$, and the mean waiting time, $E(W_N)_{ZS}$, can be obtained.

Theorem 1. For the exhaustive service, $E(\theta_N)_{ZS}$ and $E(W_N)_{ZS}$ are given by

$$E(\theta_N)_{ZS} = E(W)_{M/G/1} + h + \sum_{i=1}^{N-1} h_i \cdot E(Q_N^*)_{ZS}, \tag{3.12}$$

and

$$E(W_N)_{ZS} = \frac{\rho_1 \sum_{i=2}^{N} \rho_i}{(1-\rho)(1-\rho_1+\sum_{i=2}^{N}\rho_i)} \cdot (h-h_1) + \frac{\lambda(1-\rho_1)}{2(1-\rho)(1-\rho_1+\sum_{i=2}^{N}\rho_i)} \cdot \{h_1^{(2)} + (h-h_1)^2 + \sum_{i=2}^{N}(h_i^{(2)}-h_i^2)\}, \tag{3.13}$$

where

$$E(Q_N^*)_{ZS} := \frac{1}{(1-\rho)(1-\rho_1+\sum_{i=2}^{N}\rho_i)} [\rho_1(1-\rho_1) + \frac{1}{2}\lambda^2\{h_1^{(2)}+(h-h_1)^2 + \sum_{i=2}^{N}(h_i^{(2)}-h_i^2)\}]. \tag{3.12a}$$

Proof: Using Lemmas 1, 2, we need to calculate $E(Q_N^*)_{ZS}$ that is given by

$$(3.15)\qquad E(Q_N^*)_{ZS} := G'_{N,x_N}(1,0,\ldots,0,1) \;/\; G_N(1,0,\ldots,0,1)$$
$$= 2\mathcal{L}\cdot\{G'_{2,x_2}(1,1)\}.$$

From the results of two-stage tandem queue analysis, the first derivative in (3.15) is given by (A1.3) in Appendix 1. By operating $\mathcal{L}$ for (A1.3), i.e. the following replacement expressions

$$(3.16)\qquad \{h_2 \to (h-h_1),\ h_2^{(2)} \to \sum_{i=2}^{N}(h_i^{(2)}-h_i^2)+(h-h_1)^2,\ \rho_2 \to \sum_{i=2}^{N}\rho_i\},$$

(3.12a) is obtained.

Lemma 1 leads to the following relationship:

$$(3.17)\qquad G_1(1-S/\lambda,1,0,\ldots,0) \;/\; (\tfrac{1}{N}) = \mathcal{L}\cdot\{G_1(1-S/\lambda,1) \;/\; (\tfrac{1}{2})\}.$$

The lefthand side of (3.17) represents the LST of the waiting time in the N-stage tandem queue [2] (denoted by $W^*_N(S)_{ZS}$). Hence (3.17) can be also rewritten as

$$(3.18)\qquad W^*_N(S)_{ZS} = \mathcal{L}\cdot\{W^*_2(S)_{ZS}\}.$$

That is,

$$(3.18a)\qquad E(W_N)_{ZS} = \mathcal{L}\cdot E(W_2)_{ZS}$$

is obtained.

The mean waiting time in the two-stage tandem queue, $E(W_2)_{ZS}$, is given by (A1.5) in Appendix 1. Thus, operating $\mathcal{L}$ for (A1.5) gives (3.13). □

Letting $N=2$ in Theorem 1, $E(\theta_2)_{ZS}$ and $E(W_2)_{ZS}$ are derived. This agrees with the results of M. T. Netto [2] obtained by a different analysis. (cf. (A1.4) and (A1.5) in Appendix 1.)

Remark 3.3: The LST of the sojourn time distribution in an N-stage tandem queue with a ZS scheme, denoted by $\theta^*_N(S)_{ZS}$, is given by [2][13]

$$(3.19)\qquad \theta^*_N(S)_{ZS} = G_N(1-S/\lambda,0,\ldots,0,1-S/\lambda)/G_N(1,0,\ldots,0,1).$$

The above $W^*_N(S)_{ZS}$ is also given by

$$(3.19a)\qquad W^*_N(S)_{ZS} = G_1(1-S/\lambda,1,0,\ldots,0)/G_1(1,1,0,\ldots,0).$$

Hence, without using Lemmas 1 and 2, $E(\theta_N)_{ZS}$ and $E(W_N)_{ZS}$ can be obtained directly by letting $S \to 0$ after differentiating (3.19) and (3.19a) with respect to S. However, it seems to need a lengthy and complicated calculation to obtain the values of $G'_{N,x_1}(1,0,\ldots,1)$, $G'_{N,x_N}(1,0,\ldots,0,1)$ and $G'_{1,x_1}(1,1,0,\ldots,0)$.

3.3 Analysis for K-limited service

Consider an N-stage tandem queueing system with an NZ scheme, in which all the queues are served according to the K-limited service. The mean sojourn time, $E(\theta_N)_{NZ}$ can be derived.

Theorem 2. For the K-limited service, $E(\theta_N)_{NZ}$ is given by

$$E(\theta_N)_{NZ} = E(W)_{M/G/1} + h + \sum_{i=1}^{N-1} h_i \cdot E(Q_N^*)_{NZ}. \tag{3.20}$$

where

$$E(Q_N^*)_{NZ} := \frac{1}{2}\{K-1-(1-\rho)\sum_{r=1}^{K-1}(K-r)\frac{\sigma_r}{(r-1)!}\}. \tag{3.21}$$

Here, σ_r, $r=1,2,\ldots,K-1$ are solutions of the following simultaneous linear equations. (Using Cramer's formula, $\sigma_r (=|D_r|/|D|)$ can be represented by the determinants $|D_r|$ and $|D|$ formed by the coefficients of (3.21a).)

$$\sum_{r=1}^{K-1}\frac{1}{r!}\{\omega_j^K \prod_{i=2}^{N} Q_i^r(\omega_j) - \omega_j^r Q_1^{K-r}(\omega_j)\prod_{i=2}^{N} Q_i^K(\omega_j)\}\cdot\sigma_r$$

$$= (1-\omega_j)\prod_{i=1}^{N} Q_i^K(\omega_j), \qquad \text{for } j=1,2,\ldots,K-1, \tag{3.21a}$$

$$\omega_j := \sum_{n=1}^{\infty}\frac{(-\lambda)^{n-1}}{n!}\varepsilon_j^n \frac{d^{n-1}}{d\lambda^{n-1}}\{\prod_{i=1}^{N} H^*_i(\lambda)\}^n, \tag{3.21b}$$

and

$$\varepsilon_j := \exp\{\frac{2\pi j}{K} i\}, \qquad \text{where } i^2 = -1 \qquad \text{for } j=1,2,\ldots,K-1. \tag{3.21c}$$

Proof: Using Lemma 1 for the K-limited service, $E(Q_N^*)_{NZ}$ is given by

$$E(Q_N^*)_{NZ} = \measuredangle \cdot \{G'_{2,x_2}(1,1)\}. \tag{3.22}$$

The first derivative in (3.22) is given by (A2.3). By operating $\measuredangle$ for (A2.3), i.e. the replacement in (3.16), gives (3.21).
Thus, the following replacement in (A2.4) and (A2.5) leads (3.21a) and (3.21b),

$$\{Q_2(x) = H^*_2(\lambda(1-x)) \to \prod_{i=2}^{N} H^*_i(\lambda(1-x)) = \prod_{i=2}^{N} Q_i(x)\}. \tag{3.23}$$

□

Remark 3.4: The relationship, $E(W_N)_{NZ} = \measuredangle \cdot E(W_2)_{NZ}$, also holds in this tandem queueing system with the NZ scheme. However, the expression of $E(W_2)_{NZ}$ seems to be extremely complicated. This is the reason that Theorem 2 does not contain any results for $E(W_N)_{NZ}$.

3.4 Analysis for gated service

Consider an N-stage tandem queueing system with a GA scheme, in which all the queues are served according to the gated service. A virtual queue with infinite capacity, Q_0, is introduced in front of the first queue, Q_1, in order to analyze the mean sojourn time, $E(\theta_N)_{GA}$, and the mean waiting time, $E(W_N)_{GA}$. First, customers join Q_0. Then, all customers waiting in Q_0 move instantaneously to Q_1 in the order of their arrival when a single server arrives at S_1. After that, all customers in Q_1 are served at S_1 and they move to stage 2 for the second service and so on. (When there are no customers in Q_0 on a server-arrival at S_1, the server stays at S_1. After a new customer arrives, the server begins to serve the customer at S_1 and further continues to serve the customer at $S_2,\ldots,S_{N-1}$ and S_N sequentially.)

The steady-state probability that i_k customers are waiting in Q_k, $k=0,1,2,\ldots,N$, just after a customer has completed service at S_n, $n=1,2,\ldots,N$ is denoted by $\pi_n(i_0,i_1,i_2,\ldots,i_N)$, and the generating function is defined by the expression

$$G_n(x_0,x_1,\ldots,x_N) := \sum_{i_0=0}^{\infty}\sum_{i_1=0}^{\infty}\cdots\sum_{i_N=0}^{\infty}\pi_n(i_0,i_1,\ldots,i_N)x_0^{i_0}x_1^{i_1}\cdots x_N^{i_N}, \tag{3.24}$$

$$n=1,2,\ldots,N, \quad |x_0|,|x_1|,\ldots,|x_N| \leq 1.$$

With the aid of the results of two-stage tandem queue analysis [6][7], $E(\theta_N)_{GA}$ and $E(W_N)_{GA}$ can be derived.

Theorem 3. For the gated service, $E(\theta_N)_{GA}$ and $E(W_N)_{GA}$ are given by

$$E(\theta_N)_{GA} = E(W)_{M/G/1} + h + \sum_{i=1}^{N-1} h_i \cdot E(Q_N^*)_{GA}, \tag{3.25}$$

and

$$E(W_N)_{GA} = \frac{\rho_1(1+\rho_1)}{1-\rho^2}\cdot(h-h_1) + \frac{\lambda(1+\rho_1)}{2(1-\rho^2)}\{h_1^{(2)}+(h-h_1)^2 + \sum_{i=2}^{N}(h_i^{(2)}-h_i^2)\} \tag{3.26}$$

where

$$E(Q_N^*)_{GA} := \frac{2}{1-\rho^2}\left[\rho_1\sum_{i=2}^{N}\rho_i + \frac{\lambda^2}{2}\{h_1^{(2)}+(h-h_1)^2 + \sum_{i=2}^{N}(h_i^{(2)}-h_i^2)\}\right], \tag{3.25a}$$

Proof: Using Lemma 1 for the GA scheme, $E(Q_N^*)_{GA}$ is given by

(3.27) $E(Q_N^*)_{GA} = 2 \measuredangle \cdot G'_{2,x_2}(1,0,1).$

The first derivative in (3.27) is given by (A3.3) in Appendix 3. Hence, operating $\measuredangle$ for (A3.3) gives (3.25a). The expression of $E(W_2)_{GA}$ is also given by (A3.5). Since the same relation as (3.18a) holds, (3.26) is also obtained by operating $\measuredangle$ for (A3.5). □

4. Minimum and Maximum of Mean Sojourn Times

This section considers an N-stage tandem queueing system with a more general switching rule than the cyclic switching rule defined in Section 2. It is considered under the FIFO service discipline in each queue. By this switching rule, however, a single server is not idle as long as there are customers in the queueing system and all interrupted services are resumed, i.e. the workload in the system is conserved perfectly [15][16]. Hence, it is called workload conserving switching rule, denoted by *WL* service scheme in this paper. Everything else, e.g. arrival process, service time distribution and so on is the same as the N-stage tandem queueing model described in Section 2.

Three switching rules will also be considered.

(a) Sequential service, also called a single thread service (*ST*) [7]: customers in Q_{i+1} have priority over customers in Q_i, $i=1,2,\ldots,N-1$. This is equivalent to a basic queueing model $M/G/1$ with service times being the total sum of τ_i, $i=1,2,\ldots,N$.

(b) Preemptive priority service (*PR*): customers in Q_i have priority over customers in Q_{i+1}, $i=1,2,\ldots,N-1$. If a customer arrives at Q_i, $i=1,2,\ldots N$, when the server stays in S_j, $j>i$, the server interrupts the current service and immediately starts serving the customer in Q_i. The serving of the customer in S_j is resumed at an interrupted point when there are no customers in Q_k, $k=1,2,\ldots j-1$, $j>i$ in the system [12].

(c) Decrementing service (*DC*), also called a semi-exhaustive service [13]: when the server visits a queue, it serves until the number of customers in the queue decreases to one less than that found upon its arrival at the queue.

Note that the workload conserving switching rule includes all cyclic switching rules, in which each counter is served by *ZS*, *NZ*, *GA*, *PB*, *ST* and *DC* schemes and non-cyclic switching rules, such as *DT*, *PR* and so on.

For the *WL* service scheme, upper and lower bounds for the mean sojourn time, $E(\theta_N)_{WL}$, and upper and lower bounds for the mean waiting time, $E(W_N)_{WL}$ will be investigated.

Theorem 4. In the workload conserving switching rule, the minimum and maximum of $E(\theta_N)_{WL}$ arise from the *ST* scheme and the *PR* scheme, respectively. That is, the following inequality holds:

$$E(\theta_N)_{ST} \leqq E(\theta_N)_{WL} \leqq E(\theta_N)_{PR}, \tag{4.1}$$

where

$$E(\theta_N)_{ST} := h + \frac{\lambda}{2(1-\rho)}\left[h^2 + \sum_{i=1}^{N}(h_i^{(2)} - h_i^2)\right], \tag{4.2}$$

and

$$E(\theta_N)_{PR} := \frac{h}{(1-\sum_{i=1}^{N-1}\rho_i)} + \frac{\lambda}{2(1-\rho)(1-\sum_{i=1}^{N-1}\rho_i)} \cdot \left\{\sum_{i=1}^{N-1}(h_i^{(2)} - h_i^2) + h_N^{(2)} + 2(h-h_N) + (h-h_N)^2\right\}. \tag{4.3}$$

The inverse relationship for $E(W_N)_{WL}$ also holds, i.e.

$$E(W_N)_{ST} \geqq E(W_N)_{WL} \geqq E(W_N)_{PR}, \tag{4.4}$$

where

$$E(W_N)_{ST} := \frac{\lambda}{2(1-\rho)}\left[h^2 + \sum_{i=1}^{N}(h_i^{(2)} - h_i^2)\right], \tag{4.5}$$

and

$$E(W_N)_{PR} := \frac{\lambda}{2(1-\rho_1)} h_1^{(2)}. \tag{4.6}$$

Proof: Even for the workload conserving switching rule, the mean sojourn time $E(\theta_N)_{WL}$ of an arbitrary customer C^* can be represented by

$$E(\theta_N)_{WL} = E(W) + E(\tau) + E(U), \tag{4.7}$$

since customers in each queue are served in the order of their arrival (i.e. (3.10) also holds for the *WL* service scheme.)
Note that the first term, $E(W)$, is invariant for the *WL* service scheme (cf. Theorem 6.1 in Ref. (16)). Thus, the expression $\{E(W) + E(\tau)\}$ equals $E(W)_{M/G/1}+h$, where $E(W)_{M/G/1}$ is given by (3.9a).

Only the third term, $E(U)$, depends on the server switching rule. (See the proof of Lemma 2.) In the *ST* scheme, customers behind C^* receive no service at $S_1, S_2, \ldots$, and S_N during the sojourn time of C^*. That is, $E(U)$ is zero. In the *PR* scheme, there are never any customers in $Q_1, Q_2, \ldots$, and Q_{N-1} when C^* leaves the last counter S_N. Hence, $E(U)$ becomes maximum. Thus, the inequality (4.1) holds.

By the FIFO service discipline in Q_1, the mean waiting time, $E(W_N)_{WL}$, of

an arbitrary customer C^*, is given by the summation of the following two terms:

$$(4.8) \qquad E(W_N)_{WL} = E(W^*) + E(V).$$

The first term, $E(W^*)$, denotes the mean waiting time of C^* in Q_1 assuming that the server always stays at S_1 to serve customers in Q_1. That is, $E(W^*)$ is the minimum of the mean time spent by C^* in Q_1. Hence, it follows from the Pollaczek-Khintchine formula that

$$(4.9) \qquad E(W^*) = \frac{\lambda}{2(1-\rho_1)} h_1^{(2)}.$$

The second term, $E(V)$, denotes the mean total time spent by the server except in S_1 (i.e. $S_2, S_3, \ldots, S_N$) during the waiting time of C^* in Q_1. Only $E(V)$ depends on the server switching rule.

In the *PR* scheme, the server is in S_1 as long as there are customers in Q_1. Hence, $E(V)$ is zero.

In the *ST* scheme, there are no customers in $Q_2, Q_3, \ldots$, and Q_N when the customer C^* begins to receive the service in S_1. Hence, $E(V)$ becomes maximum. Thus, the inequality (4.4) is obtained.

Therefore, (4.2), (4.5) and (4.6) are derived by the Pollaczek-Khintchine formula.

Next, to derive (4.3) from the following result for a two-stage tandem queue with the *PR* scheme [12][15],

$$(4.10) \qquad E(\theta_2)_{PR} = \frac{h}{1-\rho_1} + \frac{\lambda}{2(1-\rho_1)(1-\rho)} \cdot \{h_1^{(2)} + 2h_1 h_2 + h_2^{(2)}\},$$

a modified two-stage tandem queue with the *PR* scheme described as follows is first introduced.

The service time T_1 at the first counter is equal to the sum of $\tau_1, \tau_2, \ldots$, and τ_{N-1}, i.e. $T_1 := \tau_1 + \tau_2 + \cdots + \tau_{N-1}$ and service time T_2 at the second counter is equal to τ_N. Everything else, e.g. arrival process and so on, are the same as with the original queueing model with the *PR* scheme. In both the original N-stage tandem queue and the modified two-stage tandem queue, (4.7) also holds and the mean workload in the system just before the arrivals of customers at Q_1 is given by $E(W)_{M/G/1}$ in (3.9a). It is noticed that in the N-stage and the 2-stage tandem queue models above, both of the last queue length distributions just after the instants at which a customer leaves the system are equal. Thus, the above arguments imply that both mean sojourn times are also equal. Hence,

$$(4.11) \qquad E(\theta_N)_{PR} = \mathcal{L}^* \cdot E(\theta_2)_{PR}$$

is obtained, where $\measuredangle$ * is an operator for substituting $H_1(t)$ for $H_1 * H_2 * \cdots * H_{N-1}(t)$ and $H_2(t)$ for $H_N(t)$, i.e. $\measuredangle$ * means the following replacement,

$$(4.12) \quad \{h_1 \to (h-h_N),\ h_1^{(2)} \to \sum_{i=1}^{N-1} (h_i^{(2)} - h_i^2) + (h-h_N)^2,\ \rho_1 \to \sum_{i=1}^{N-1} \rho_i,$$

$$h_2 \to h_N,\ h_2^{(2)} \to h_N^{(2)},\ \rho_2 \to \rho_N\}.$$

Therefore, operating $\measuredangle$ * for (4.10) gives (4.3). □

5. Conclusion

In this paper, mean sojourn times in the system and mean waiting times in the first stage have been derived for a multi-stage tandem queue served by a single server with a cyclic switching rule. These were derived by a simple method using the results of the two-stage tandem queue analysis and the Pollaczek-Khintchine formula. In a general switching rule (called the work-load conserving switching rule), upper and lower bounds for the mean sojourn times and the mean waiting times were obtained. Further study is needed for mean sojourn time analysis in the multi-stage tandem queue with server walking time.

References

[1] Nelson, R.T.: "Dual-Resource Constrained Series Service Systems", Oper. Res., 16, 2, pp.324-341 (1968).

[2] Taube-Netto, M.: "Two Queues in Tandem Attended by a Single Server", Oper. Res., 25, 1, pp.140-147 (1977).

[3] Nair, S.S.: "Semi-Markov Analysis of Two Queues in Series Attended by a Single Server", Bull. Soc. Math. Belgique, 22, pp.355-367 (1970).

[4] Nair, S.S.: "Two Queues in Series Attended by a Single Server", Ibid., 25, 160-176 (1973).

[5] Nair, S.S.: "A Single Server Tandem Queue", J. Appl. Prob., 8, 1, pp.95-109 (1971).

[6] Katayama, T.: "Analysis of a Tandem Queueing System with Gate Attended by a Moving Server", Review of the Electrical Communi. Lab., 29, 3-4, pp.254-267 (1981).

[7] Katayama, T.: "Analysis of a Tandem Queueing System with Gate Attended by a Moving Server with Walking Time", Trans. of IECE of Japan, J64-B, 9, pp.931-938 (1981).

[8] Katayama, T.: "Analysis of a Finite Intermediate Waiting Room Tandem Queue Attended by a Moving Server with Walking Time", Trans. of IECE of Japan, E-64, 9, pp.571-578 (1981).

[9] Murakami, K. and Nakamura, G.: "A Model for Event Handling in a Functionally Dedicated Processor and Analysis", Trans. of IECE of Japan, J61-D, 7, pp.465-472 (1978).

[10] Nishida, T., Serikawa, Y. and Yoneyama, K.: "n Queues in Tandem Attended by a Single Server", Rep. Stat. Appl. Res., JUSE, 26, 1, pp.14-19 (1979).

[11] König, D. and Schmit, V.: "Relationships between Time/Customer Stationary Characteristics of Tandem Queues Attended by a Single Server", J. Oper. Res. Soc. Japan, 27, 3, pp.191-204 (1984).

[12] Enns, E.G.: "Some Waiting Time Distributions for Queues with Multiple Feedback and Priorities", Oper. Res., 17, 3, pp.519-525 (1969).

[13] Takagi, H.: "Analysis of Polling System", The MIT Press (1985).

[14] Takagi, H.: "A Survey of Queueing Analysis of Polling Models", Proc. of the Seminar on Queueing Theory and Its Application, pp.97-116, Dep. of Appl. Math. and Phys., Faculty of Eng. Kyoto Univ. (1987).

[15] Cooper, R.B.: "Introduction to Queueing Theory", Second Edition, Edward-Arnold (1981).

[16] Gelenbe, E. and Mitrani, I.: "Analysis and Synthesis of Computer Systems", Academic Press (1980).

Tsuyoshi KATAYAMA: NTT Communication Switching Laboratories, 9-11, Midori-Cho, 3-Chome, Musashino-Shi, Tokyo 180, Japan

Appendix 1. Generating function for the exhaustive service

In the two-stage tandem queue model with a *ZS* scheme, the following functional relationships [9][10] for $G_n(x_1,x_2)$, n=1,2 defined by (3.1) are obtained:

$$\text{(A1.1)}\quad G_1(X_1,X_2) = \{G_1(X_1,X_2) - G_1(0,X_2)\}\cdot Q_1(X_1)\cdot X_2/X_1 + \{G_2(X_1,0) - \pi_2(0,0)\}\cdot Q_1(X_1)\cdot X_2/X_1 + \pi_2(0,0)\cdot X_1\cdot Q_1(X_1)\cdot X_2/X_1,$$

and

$$\text{(A1.2)}\quad G_2(X_1,X_2) = \{G_2(X_1,X_2) - G_2(X_1,0)\}\cdot Q_2(X_1)/X_2 + G_1(0,X_2)\cdot Q_2(X_1)/X_2.$$

From analysis of the two-stage tandem queue model with a *ZS* scheme, the following expressions have been obtained [2][6][7]:

$$\text{(A1.3)}\quad G'_{2,x_2}(1,1) = \frac{1}{2(1-\rho)(1-\rho_1+\rho_2)}\left[\rho_1(1-\rho_1) + \frac{\lambda^2}{2}\left(h_1^{(2)}+h_2^{(2)}\right)\right],$$

$$\text{(A1.4)}\quad E(\theta_2)_{ZS} = \frac{(1-\rho_1)(1+\rho_2)}{(1-\rho)(1-\rho_1+\rho_2)}\cdot h_1+h_2+\frac{\lambda(1+\rho_2)}{2(1-\rho)(1-\rho_1+\rho_2)}\left(h_1^{(2)}+h_2^{(2)}\right),$$

and

$$\text{(A1.5)}\quad E(W_2)_{ZS} = \frac{\rho_1\,\rho_2}{(1-\rho)(1-\rho_1+\rho_2)}\cdot h_2 + \frac{\lambda(1-\rho_1)}{2(1-\rho)(1-\rho_1+\rho_2)}\left(h_1^{(2)}+h_2^{(2)}\right).$$

Appendix 2. Generating function for the *K*-limited service

In the two-stage tandem queue model with a *NZ* service scheme, the following functional relationships [8] for $G_N(X_1,X_2)$ are obtained:

$$\text{(A2.1)}\quad G_1(X_1,X_2) = \{G_1(X_1,X_2)-G_1(0,X_2)-F_K(X_1,X_2)\}\cdot Q_1(X_1)\cdot X_2/X_1 + \{G_2(X_1,0) - \pi_2(0,0)\}\cdot Q_1(X_1)\cdot X_2/X_1 + \pi_2(0,0)\cdot X_1\cdot Q_1(X_1)\cdot X_2/X_1,$$

and

$$\text{(A2.2)}\quad G_2(X_1,X_2) = \{G_2(X_1,X_2) - G_2(X_1,0)\}\cdot Q_2(X_1)/X_2 + \{G_1(0,X_2) + F_K(X_1,X_2)\}\cdot Q_2(X_1)/X_2,$$

where

$$F_K(X_1,X_2) := \sum_{i_1=1}^{\infty} \pi_1(i_1,K)\, X_1^{i_1} X_2^{K}.$$

From analysis of the two-stage tandem queue model with a *NZ* scheme, the following expressions have been obtained [5][8]:

$$G'_{2,x_2}(1,1) = \frac{1}{4}\{K-1 - (1-\rho)\cdot\sum_{r=1}^{K-1}(K-r)\frac{\sigma_r}{(r-1)!}\}, \tag{A2.3}$$

$$\sum_{r=1}^{K-1}\frac{1}{r!}\{\omega_j^{K}\cdot Q_2^{r}(\omega_j) - \omega_j^{r}\cdot Q_1^{K-r}(\omega_j)Q_2^{K}(\omega_j)\}\cdot\sigma_r = (1-\omega_j)\{Q_1(\omega_j)\cdot Q_2(\omega_j)\}^{K}, \quad \text{for } j=1,2,\ldots,K-1. \tag{A2.4}$$

$$\omega_j := \sum_{n=1}^{\infty}\frac{(-\lambda)^{n-1}}{n!}\varepsilon_j^{n}\frac{d^{n-1}}{d\lambda^{n-1}}\{H^*_1(\lambda)\cdot H^*_2(\lambda)\}^{n}, \tag{A2.5}$$

and

$$\varepsilon_j := \exp\{\frac{2\pi j}{K}i\}, \quad \text{where } i^2=-1 \quad \text{for } j=1,2,\ldots,K-1. \tag{A2.6}$$

Appendix 3. Generating function for the gated service

In the *N*-stage tandem queue model with a *GA* scheme, let $q^{(i)}_m$ denote the probability that m customers arrive at Q_0 during service times τ_i, $i=1,2$ and let $Q_i(X)$ denote the generating function. Then, the following functional relationships [6][7] for $G_n(X_0,X_1,X_2)$, $n=1,2$ defined by (3.24) are obtained:

$$\begin{aligned} G_1(X_0,X_1,X_2) = {} & \{G_1(X_0,X_1,X_2)-G_1(X_0,0,X_2)\}\cdot Q_1(X_0)\cdot X_2/X_1 \\ & + \{G_2(X_1,0,0)-\pi_2(0,0,0)\}\cdot Q_1(X_0)\cdot X_2/X_1 \\ & + \pi_2(0,0,0)\cdot X_1\cdot Q_1(X_0)\cdot X_2/X_1, \end{aligned} \tag{A3.1}$$

and

$$\begin{aligned} G_2(X_0,0,X_2) = {} & \{G_2(X_0,0,X_2) - G_2(X_0,0,0)\}\cdot Q_2(X_0)/X_2 \\ & + G_1(X_0,0,X_2)\cdot Q_2(X_0)/X_2. \end{aligned} \tag{A3.2}$$

From analysis of the two-stage tandem queue model with a *GA* scheme, the following expressions have been obtained [6][7]:

$$G'_{2,x_2}(1,0,1) = \frac{1}{1-\rho^2}\{\rho_1\rho_2 + \frac{\lambda^2}{2}(h_1^{(2)}+h_2^{(2)})\}, \tag{A3.3}$$

$$(A3.4) \qquad E(\theta_2)_{GA} = \frac{1-\rho_1^2+\rho_2}{1-\rho^2} \cdot h_1 + h_2 + \frac{\lambda(1+2\rho_1+\rho_2)}{2(1-\rho^2)} (h_1^{(2)}+h_2^{(2)}),$$

and

$$(A3.5) \qquad E(W_2)_{GA} = \frac{\rho_1(1+\rho_1)}{1-\rho^2} \cdot h_2 + \frac{\lambda(1+\rho_1)}{2(1-\rho^2)} (h_1^{(2)}+h_2^{(2)}).$$