The Order Statistics of the Sojourn Times of Customers that Form a Single Batch in the $M^X/M/c$ Queue

R. Nelson, A.N. Tantawi, D. Towsley

COINS Technical Report 91-05

The Order Statistics of the Sojourn Times of Customers that Form a Single Batch in the $M^X/M/c$ Queue

R. Nelson, A.N. Tantawi
IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

D. Towsley*
Dept. of Computer and Information Science
University of Massachusetts
Amherst, MA 01003

August 1989

Abstract

In this paper we study the order statistics of the sojourn times of customers that form a single batch arriving in a $M^X/M/c$ queue. We determine the marginal probability density functions of each of the order statistics, the density function of the range and the joint probability density function of all of the order statistics. In addition, we obtain simple upper and lower bounds on the cumulative distributions of the order statistics of the sojourn times. The results in this paper can be applied to modeling parallel processing and fault-tolerant systems.

*The work of this author was partially supported by the Office of Naval Research under Contract N0014-87-K-0796

1 Introduction

We consider an $M^X/M/c$ queue where customers arrive in batches according to a Poisson process with parameter λ and the number of customers in a batch, X, is a random variable (r.v.) having probability distribution $\beta_n = \Pr[X = n]$, $1 \leq n$ and mean $E[X] < \infty$. The service time requirements for the customers form an independent and identically distributed (i.i.d.) set of exponential r.v.'s with mean $1/\mu$. Batches are scheduled in a first come first serve manner and customers within a batch are served in random order.

Consider a batch containing b customers labeled $1, 2, \cdots, b$ and let T_a, $a = 1, 2, \cdots, b$ denote the sojourn time of the a-th customer in the batch. It is clear that $\{T_a\}_{a=1}^{b}$ form a sequence of identically (but not independent) distributed random variables. Let $T(1,b) \leq T(2,b) \leq \cdots \leq T(b,b)$ denote these sojourn times arranged in increasing order and thus $T(a,b)$ denotes the a-th order statistic of batch customer sojourn time.

In this paper we derive expressions for the probability density function (p.d.f.) of $T(a,b)$, denoted by $f_{T(a,b)}(t)$, $t \geq 0$, the joint p.d.f. of the order statistics $\underline{T}(b) \equiv (T(1,b), T(2,b), \cdots, T(b,b))$, denoted by $f_{\underline{T}(b)}(\underline{t})$, $\underline{t} = (t_1, t_2, \cdots, t_b)$, $t_i \geq 0$, $1 \leq i \leq b$, $b \geq 1$, and the p.d.f. of the *range* of order statistics, $R(b) \equiv T(b,b) - T(1,b)$, denoted by $f_{R(b)}(t)$, $t \geq 0$.

These results are of interest because batch arrival queueing systems of this type can be used to model parallel processing systems. For example, in one model of a fork-join queueing system, jobs consist of independent tasks that can be executed concurrently. The job is considered to be completed only when all of its tasks have finished execution. In the queueing model analyzed in the paper, a job of this type would correspond to a batch and its response time to $T(b,b)$ defined above. We will sometimes refer to $T(b,b)$ as the *batch sojourn time*. In [7] such a fork-join model was considered and an analysis for the mean response time was presented. Our results here can thus be viewed, in this context, as an extension of that analysis.

Another application is to fault tolerant computing systems. Many such systems execute jobs consisting of distinct tasks that perform identical functions. As soon as two tasks (or a majority) agree in their results, the job execution is considered complete. Hence if the job consists initially of b tasks, its response time is $T(2,b)$ (or $T(\lfloor (b+1)/2 \rfloor, b)$ in the case of majority voting).

The problem of determining order statistics of the sojourn times of customers within a batch has not been studied for multiple server queues. In the case of a batch arrival $M^X/G/1$ queue, Whitt [10] and Haflin [4] studied the behavior of the batch sojourn time. The primary purpose of these studies was to compare this statistic with the sojourn time of a randomly chosen customer.

There has been recent interest in the problem of modeling parallel processing behavior. A number of papers [1, 7, 9, 2, 3] have studied the behavior of the maximum order statistic for a class of queueing systems referred to as *fork-join* queues. Here there exist c servers, each with its own queue. At the time of an arrival of a batch, the customers are assigned to different servers. In our model, on the other hand, there is a single queue and customers need not receive service at different servers.

The paper has the following organization. The model is formally defined in Section 2. This section also contains derivations of the probability density functions of the marginal order statistics, the joint statistics and the range Upper and lower bounds on the order statistics are obtained in Section 3 and the results are summarized in Section 4.

2 Analysis

In this section we present the analysis for the derivation of the expressions for the pdf's of $T(a,b)$, $R(b)$, and $\underline{T}(b)$. Our approach is to tag a random batch, B, and enumerate all delays encountered by its customers after arriving to the system. We assume that B has $b \geq 1$ customers in its batch. Customer sojourn times clearly depend on the number of customers in the system at the time of $B's$ arrival to the system. We denote this random variable by K and let its stationary distribution be given by $\pi_k = \Pr[K = k]$, $k \geq 0$. Let $T_k(a,b)$ be a random variable denoting the a-th order statistic ($a = 1, 2, \cdots, b$) of the sojourn times of customers from B given that $K = k, k \geq 0$. We can write

$$T_k(a,b) = \begin{cases} S_{\min(b,c-k)}(a,b), & k < c, \\ W_k + S_1(a,b), & k \geq c. \end{cases} \tag{1}$$

In (1) W_k is a random variable that represents the time B waits in the queue before any of its customers are selected for service given that $K = k$ and $S_i(a,b), i = 1, 2, \ldots, c, 1 \leq a \leq b \leq c$ is a random variable that denotes the time needed to service a customers from a batch of size b given that B is initially allocated i servers. We denote the p.d.f. of $S_i(a,b)$ by $f_{S_i(a,b)}(t)$.

By removing the conditioning, we can write

$$f_{T(a,b)}(t) = \sum_{k=0}^{\infty} \pi_k f_{T_k(a,b)}(t), \quad 1 \le a \le b,\ b \ge 1, \tag{2}$$

where $f_{T_k(a,b)}(t)$ is the p.d.f. of $T_k(a,b)$.

An expression for the joint p.d.f. of the order statistics, $f_{\underline{T}_k(b)}(\underline{t})$, can be written similarly. Let $\underline{T}_k(b)$ denote the value of $\underline{T}(b)$ conditioned on the number of customers in the system at the time of $B's$ arrival being k. We can write

$$\underline{T}_k(b) = \begin{cases} \underline{S}_{\min(b,c-k)}(b), & k < c, \\ W_k \underline{1}_b + \underline{S}_1(b), & k \ge c. \end{cases} \tag{3}$$

where $\underline{1}_b$ is a vector containing b 1's and $\underline{S}_i(b) = (S_i(1,b), S_i(2,b), \cdots, S_i(b,b))$ are the joint order statistics for the times needed to serve all b customers from B. We denote the p.d.f. for this random variable as $f_{\underline{S}_i(b)}(t)$.

Similar to equation 2, we can write

$$f_{\underline{T}(b)}(t) = \sum_{k=0}^{\infty} \pi_k f_{\underline{T}_k(b)}(t), \quad 1 \le a \le b,\ b \ge 1 \tag{4}$$

where $f_{\underline{T}_k(b)}(t)$ is the p.d.f. of $\underline{T}_k(b)$.

Expressions for π_k, the density of W_k, $f_{S_i(a,b)}$ and $f_{\underline{S}_i(b)}(t)$ that are needed to evaluate (2) and (4) are derived in the following sections.

We first define three random variables that will be subsequently used.

- E_k - An Erlang k r.v. with parameter $c\mu$, density function $g(t;k) \equiv c\mu(c\mu t)^{k-1}e^{-c\mu t}/k!$, $k \ge 1$, $0 \le t$, and expectation $\overline{E}_k \equiv k/c\mu$.
- $H_{m,n}$ - The m-th order statistic of a sequence of i.i.d. exponential r.v.'s, each with mean $1/\mu$. The density function and mean are

$$h(t;m,n) \equiv \mu \binom{n}{m} \left(1 - e^{-\mu t}\right)^{m-1} e^{-\mu t(n-m+1)}, \quad 0 \le t;\ 1 \le m \le n,$$

$$\overline{H}_{m,n} \equiv \sum_{l=n-m+1}^{n} 1/(l\mu), \quad 1 \le m \le n$$

respectively.

- $H_{k,m,n} \equiv E_k + H_{m,n}$ (note: $H_{0,m,n} = H_{m,n}$) with density function

$$h(t;k,m,n) \equiv \mu(c\mu)^k \binom{n}{m} \sum_{l=1}^{m} \binom{m-1}{l-1} (-1)^l \left[e^{-\mu ct} \sum_{r=1}^{k} \frac{t^{k-r}}{(k-r)!\mu^r (c-n+m-l)^r} - \frac{e^{-\mu t(n-m+l)}}{\mu^k (c-n+m-l)^k} \right], \ 0 \le t;\ 0 \le k;\ 1 \le m \le n$$

and expectation $\overline{H}_{k,m,n} \equiv k/c\mu + \sum_{l=n-m+1}^{n} 1/(l\mu)$. The density function is obtained by convolving $g(t;k)$ and $h(t;m,n)$ together.

We denote the convolution of two non-negative p.d.f.'s by $[f(t) * g(t)] = \int_0^t f(x)g(t-x)dx$.

2.1 Derivation of Waiting Time Statistics

The stationary probability, $\pi_k, k = 0, 1, \ldots$, that an $M^X/M/c$ queue has k customers is given by (see [11]):

$$\pi_k = \frac{c\rho}{min(c,k)} \sum_{j=0}^{k-1} \pi_j \beta'_{k-j}, \qquad k \ge 1, \tag{5}$$

where $\beta'_j = \sum_{i=j}^{\infty} \beta_i / E[X]$ and $\rho = \lambda E[X]/(c\mu)$. The probabilities π_i, $i = 0, 1, \cdots, c-1$, are obtained from the first $c-1$ equations of (5) along with the normalizing equation

$$(1-\rho) = \sum_{k=0}^{c-1} (1 - k/c)\pi_k.$$

The remaining probabilities can be computed directly from (5) in a recursive manner.

We can write W_k as

$$W_k = \begin{cases} 0, & k < c, \\ E_{k-c+1}, & k \ge c. \end{cases} \tag{6}$$

The mean wait time until a batch begins service, $\overline{W}$, is ([11, 7])

$$\overline{W} = \left(\overline{K} + 1 - c + \sum_{k=0}^{c-1} \pi_k (c-k-1) \right) / (c\mu) \tag{7}$$

where $\overline{K}$, the mean number of customers in the system, is given by

$$\overline{K} = \left[\rho\left(\sum_{k=1}^{\infty} k\beta'_k + 1\right) + \sum_{k=1}^{c-1}(1 - k/c)k\pi_k\right] /(1-\rho).$$

2.2 Derivation of Service Time Statistics

To calculate $f_{S_i(a,b)}$ we define a Markov process called the *service path* which accounts for the number of B's customers in the queue and in the servers during the time it has at least one of its customers in service. Let $X_r = (Q_r, N_r)$ for $r = 0, 1, \ldots$ be a Markov process where Q_r is a random variable denoting the number of B's customers that are in the queue and N_r is the number of servers it holds. The transitions of X_r correspond to particular departure epochs from the system. Specifically, the service path is composed of two disjoint periods, the *increasing period* during which B has at least one customer in the queue ($Q_r > 0$ and thus the number of servers it holds can only increase) and the *decreasing period* during which B has no customers in the queue ($Q_r = 0$ and thus the number of servers it holds can only decrease). The increasing period does not exist if, when scheduled, B obtains b servers.

The state transitions for the service path are defined differently for each of the above periods. In the increasing period, a transition corresponds to a departure from the system of any customer from any batch. In the decreasing period, a transition corresponds to a departure from the system of any customer from $B's$ batch. In the following, we will also refer to state transitions as *steps* of the service path.

Assume that $N_0 = i \geq 1$ and furthermore that the increasing period exists, i.e. $Q_0 = b - i > 0$. This implies that the service path is in the increasing period for $0 \leq r \leq b - i$ and is in the decreasing period for $b - i < r \leq b - i + N_{b-i}$. The number of departures of customers belonging to B by the r-th step during the increasing period is given by $D_r \equiv N_0 + r - N_r$, $0 \leq r \leq b - N_0$.

To write the transition probabilities for the increasing period, assume that $Q_r = q_r > 0$ and that $r = 0, 1, \ldots, b - i$. We then have

$$\Pr[X_{r+1} = (q_{r+1}, n_{r+1}) \mid X_r = (q_r, n_r)] = \begin{cases} n_r/c, & q_{r+1} = q_r - 1, n_{r+1} = n_r, \\ 1 - n_r/c, & q_{r+1} = q_r - 1, n_{r+1} = n_r + 1, n_r < c, \\ 0, & \text{otherwise}, \end{cases} \quad (8)$$

To write the transition probabilities for the decreasing period assume that $Q_r = 0$ and that $r = b - i + 1, \ldots, b - i + N_{b-i}$. We then have

$$\Pr[X_{r+1} = (q_{r+1}, n_{r+1}) \mid X_r = (q_r, n_r)] = \begin{cases} 1, & q_{r+1} = q_r, n_{r+1} = n_r - 1, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

The value of the random variable $S_i(a, b)$ is equal to the first hitting time of state (0,0) of the Markov process X_r. To calcuate its expected value we let

$$\kappa_{r,j} = \begin{cases} 1, & N_r = j \text{ and } N_{r-1} = j, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

for $r = 1, \cdots, b - i$. The indicator random variable $\kappa_{r,j}, 1 \leq r \leq b - i, 1 \leq j \leq c$ is equal to 1 if step r is a departure of a customer belonging to B and customers in B hold j servers. We define $n_j(r) = \sum_{l=1}^{r} \kappa_{l,j}$ to be the number of steps up to step r that occur while B holds j servers.

Let $p(j; i, r) = \Pr[N_r = j | N_0 = i]$, for $j = i, i+1, \cdots, c$, $r = 0, 1, \cdots, b - i$, $i = 1, \cdots, c$ and equal to 0 elsewhere. ¿From (8) we can write

$$p(j; i, r) = \sum_{\sum_{l=i}^{j} n_l(r) = r - j + i} \prod_{l=i}^{j} \left(\frac{l}{c}\right)^{n_l(r)} \prod_{l=i}^{j-1} \left(1 - \frac{l}{c}\right), \quad i \leq j \leq \min(c, i + r). \tag{11}$$

This expression has the following closed form (see [5, 6])

$$p(j; i, r) = \binom{c - i}{j - i} \sum_{l=i}^{j} \binom{j - i}{l - i} \left(\frac{l}{c}\right)^r (-1)^{j-l}, \quad i \leq j \leq \min(c, i + r). \tag{12}$$

Let $\beta_a^{inc}(r \mid i), a \leq r \leq b - i$, be the probability that the a-th departure from B occurs at step r in the increasing period given that B begins its service period with i servers. This event can happen only if there are $a - 1$ departures from B at step $r - 1$ and a departure from B at step r. We thus have, after conditioning on N_0, $\beta_a^{inc}(r|i) = \Pr[D_{r-1} = a - 1 | N_0 = i] N_{r-1}/c$. Rewriting we have

$$\begin{aligned} \beta_a^{inc}(r|i) &= \Pr[N_{r-1} = i + r - a | N_0 = i](i + r - a)/c, \\ &= p(i + r - a; i, r)(i + r - a)/c. \end{aligned} \tag{13}$$

In a similar manner, let $\beta_a^{dec}(r|i)$, $0 < b-i < r$, be the probability that the a-th departure from B occurs at step r in the decreasing period given that B begins its service period with i servers. This event can happen only if there are $a+b-r-i$ departures from B at step $b-i$ in the increasing period. This implies that $N_{b-i} = i+r-a$ and, since $N_{b-i} \leq \min(b,c)$, that $r \leq \min(b,c)+a-i$. We thus have $\beta_a^{dec}(r|i) = \Pr[D_{b-i} = a+b-r-i|N_0 = i]$. Rewriting yields

$$\begin{aligned} \beta_a^{dec}(r|i) &= \Pr[N_{b-i} = i+r-a|N_0 = i], \\ &= p(i+r-a; i, b-i). \end{aligned} \tag{14}$$

The pdf of $S_i(a,b)$ can be expressed by conditioning on the step within the service path that the a-th customer completes and then removing the conditioning by using $\beta_a^{inc}(r|i)$ and $\beta_a^{dec}(r|i)$. This results in

$$f_{S_i(a,b)}(t) = \sum_{r=1}^{b-i} \beta_a^{inc}(r|i) g(t;r) + \sum_{r=b-i+1}^{\min(b,c)+a-i} \beta_a^{dec}(r|i) h(t; b-i, i+r-b, i+r-a). \tag{15}$$

We can write the average value of $S_i(a,b)$ as

$$\overline{S}_i(a,b) = \sum_{r=1}^{b-i} \beta_a^{inc}(r|i) \overline{E}_r + \sum_{r=b-i+1}^{\min(b,c)+a-i} \beta_a^{dec}(r|i) \overline{H}_{b-i,i+r-b,i+r-a}. \tag{16}$$

2.3 Derivation of the a-th Order Statistic

Using equation (1) allows us to write the following expression for $f_{T_k(a,b)}(t)$,

$$f_{T_k(a,b)}(t) = \begin{cases} f_{S_{\min(b,c-k)}(a,b)}(t) & k < c, \\ g(t; k-c+1) * f_{S_1(a,b)}(t), & k \geq c \end{cases} \tag{17}$$

which can be substituted into equation (2) to yield $f_{T(a,b)}(t)$.

The mean sojourn time $E[T(a,b)]$ is given by

$$E[T(a,b)] = \begin{cases} \overline{W} + \sum_{k=0}^{c-b} \pi_k \overline{S}_b(a,b) + \sum_{k=c-b+1}^{c-1} \pi_k \overline{S}_{c-k}(a,b) \\ \quad + (1 - \sum_{k=0}^{c-1} \pi_k) \overline{S}_1(a,b), & b = 1, \cdots, c, \\ \overline{W} + \sum_{k=0}^{c-1} \pi_k \overline{S}_{c-k}(a,b) + (1 - \sum_{k=0}^{c-1} \pi_k) \overline{S}_1(a,b), & b = c+1, \cdots, \end{cases} \tag{18}$$

Similar expressions can also be obtained for the Laplace transform and higher order moments of $T(a,b)$.

2.4 Derivation of Range and Joint Order Statistics

To derive the pdf of the joint order statistics we focus on the joint service statistics. Let r_i^* be the first step at which i tasks of B have completed execution and Δ_i be the number of transitions within the service path that occur between the i^{th} and $i-1^{st}$ departures of tasks in B. Thus r_i^*, $1 \le i \le b$ is the unique value that satisfies $D_{r_i^*-1} = i-1$ and $D_{r_i^*} = i$ and $\Delta_i = r_i^* - r_{i-1}^*$ where we set $r_0^* = 0$. Assume that $N_{b-i} = j$; i.e. that $b-j$ departures from B occurred in the increasing period and j departures occurred in the decreasing period. Let $f_{\underline{S}_i(b)|D_{b-i}}(\underline{t}|j)$ denote the conditional pdf of $\underline{S}_i(b)$ given the value of D_{b-i}. It is given by

$$f_{\underline{S}_i(b)|D_{b-i}}(\underline{t}|j) = E_{\Delta_1}(t_1)\prod_{i=2}^{b-j} g(t_i - t_{i-1};\Delta_i)\prod_{i=0}^{j-1}(j-i)\mu e^{-(j-i)\mu(t_{b-j+i+1}-t_{b-j+i})} \tag{19}$$

where $\underline{t} = (t_1, t_2, \cdots, t_b)$. Removal of the conditioning on D_{b-i} yields

$$f_{\underline{S}_i(b)}(\underline{t}) = \sum_{j=i}^{c} p(j;i,b-i) f_{S_i(b)|D_{b-i}}(\underline{t}|j). \tag{20}$$

For the case where there is no increasing period, i.e. where $b-i=0$, we can easily write the following expressions

$$f_{\underline{S}_b(b)}(\underline{t}) = b\mu e^{-b\mu t_1}\prod_{i=1}^{b-1}(b-i)\mu e^{-(b-i)\mu(t_{i+1}-t_i)}, \quad 0 \le t_1 \le t_2 \le \cdots \le t_b. \tag{21}$$

The above results leads to the following expression for the joint response time order statistics,

$$f_{\underline{T}_k(b)}(\underline{t}) = \begin{cases} f_{\underline{S}_{\min(b,c-k)}(b)}(\underline{t}) & k < c, \\ \int_0^{t_1}[g(x;b-c+k-1) * f_{\underline{S}_1(b)}(\underline{t}-x)]dx, & k \ge c. \end{cases} \tag{22}$$

We now determine the pdf of the range of order statistics. We condition on the number of servers available to B when it begins its service period. Let $R_i(b)$ be the range when i servers are available to $B, i = 1, \cdots, c$. Suppose the first departure from b occurs at step j in the service path, where $1 \leq j \leq c - i + 1$. The range has the same distribution as $S_{j-1+i}(b-1, b-1)$. We can thus write the pdf of the range as

$$f_{R_i(b)}(t) = \sum_{j=1}^{c-i+1} \beta_1^{inc}(j|i)[g(t;j) * f_{S_{j-1+i}(b-1,b-1)}(t)]. \tag{23}$$

Removal of the conditioning on the number of servers available yields

$$f_{R(b)}(t) = \sum_{k=0}^{c-1} \pi_k f_{R_{\min(b,c-k)}(b)}(t) + (1 - \sum_{k=0}^{c-1} \pi_k) f_{R_1(b)}(t). \tag{24}$$

The moments of the range can be obtained in a similar manner.

3 Bounds on the Order Statistics of the Sojourn Times

In this section we derive bounds on the random variables $\{S_i(a, b)\}$ defined in section 3. Specifically we define two sets of r.v.'s one of which contains r.v.'s that stochastically dominate $S_i(a, b)$ and the other containing r.v.'s that are stochastically dominated by $S_i(a, b)$. Here the *stochastic dominance* relation $\geq_{st}$ is defined as

Definition 1 *The random variable X stochastically dominates the random variable Y, written as $X \geq_{st} Y$ iff*

$$\Pr[X \leq t] \leq \Pr[Y \leq t]$$

Definition 2 *Let $Z_{a,b}$ be a r.v. corresponding to the a-th smallest of a set of b independent exponential r.v.'s each with parameter μ.*

The r.v. $Z_{a,b}$ exhibits the following property, $Z_{a,b} \geq_{st} Z_{a-1,b-1}$, $1 < a \leq b$. We have the following theorem.

Theorem 1 *The following stochastic dominance relation holds among the set of r.v.'s $\{S_i(a, b)\}_{1 \leq a \leq b, 1 \leq i \leq c}$*

$$\begin{array}{llllllll} E_{b-1} + S_b(a,b) & \geq_{st} & E_{b-i} + S_b(a,b) & \geq_{st} & S_i(a,b) & \geq_{st} & S_b(a,b) & ,1 \leq b \leq c, \\ E_{c-1} + S_c(a,b) & \geq_{st} & E_{c-i} + S_c(a,b) & \geq_{st} & S_i(a,b) & \geq_{st} & S_c(a,b) & ,c < b. \end{array}$$

Proof. These dominance relations follow from simple path coupling arguments. We will present the proof of the relation $E_{b-i} + S_b(a,b) \geq_{st} S_i(a,b)$ in the case of $1 \leq b \leq c$.
The service time $S_i(a,b)$ can be expressed in the form of either

- $S_i(a,b) = E_k(\mu c)$, $a \leq k \leq b - i$, or
- $S_i(a,b) = E_{b-i}(\mu c) + Z_{k,b-k+a}$, $1 \leq k \leq a$.

As a consequence of the above relations and properties of $Z_{a,b}$, it follows that $E_{b-i}(\mu c) + S_b(a,b) \geq_{st} S_i(a,b)$.

The remaining relations stated in the theorem can be proven in a similar manner. □

This theorem yields simple bounds on the cumulative marginal distribution of the random variables $T(a,b)$, $1 \leq a \leq b$ which can be found in the Appendix. Bounds on $E[T(a,b)]$ are also found in the Appendix.

We observe that the maximum difference between the loosest lower bound and the upper bound is $(c-1)/(\mu c)$. Consequently, the bounds are very tight as the load increases, $\lambda E[D]/(\mu c) \rightarrow 1$, and/or the batch size increases, $b \rightarrow \infty$, while a/b remains constant.

4 Summary

In this paper we have obtained the order statistics of the sojourn times of customers in an $M^X/M/c$ queue. We have derived expressions for the densities of random variables $T(a,b)$, $\underline{T}(b)$, and $R(b)$. In the case of $T(a,b)$, and $\underline{T}(b)$ the expression requires evaluation of an infinite summation. The pdf of $R(b)$, however only requires a finite summation.

These results can be used to analyze models of parallel processing systems. Jobs in such models are assumed to consist of tasks that can be executed in parallel. The value of $T(b,b)$ in such a system corresponds to job response time, an important performance measure. The value of $R(b)$ corresponds to the time that tasks of a job wait for their siblings to complete execution. In systems where system locks are used, this time corresponds to the amount of time that a job holds a lock.

A Bounds on the Sojourn Time Order Statistics

We present the bounds on the cumulative distribution and expected value of $T(a,b)$, $1 \le a \le b$.

$$
\begin{aligned}
&P[K < c]H(t; b-1, a, b) + \sum_{k=c}^{\infty} \pi_k H(t; k-c+b, a, b) \le \\
&\quad \sum_{k=0}^{c-1} \pi_k H(t; \min\{c-k, b\}, a, b) + \sum_{k=c}^{\infty} \pi_k H(t; k-c+b, a, b) \le \\
&\quad \Pr[T(a,b) \le t] \le P[K < c]H(t; 0, a, b) + \sum_{k=c}^{\infty} \pi_k H(t; k-c+1, a, b), \quad 1 \le b \le c,
\end{aligned}
$$

$$
\begin{aligned}
&P[K < c]H(t; b-1, c+a-b, c) + \sum_{k=c}^{\infty} \pi_k H(t; k-c+b, c+a-b, c) \le \\
&\quad \sum_{k=0}^{c-1} \pi_k H(t; c-k, c+a-b, c) + \sum_{k=c}^{\infty} \pi_k H(t; k-c+b, c+a-b, c) \le \Pr[T(a,b) \le t] \\
&\quad \le P[K < c]H(t; b-c, c+a-b, c) + \sum_{k=c}^{\infty} \pi_k H(t; k-c+1, c+a-b, c), \quad 1 \le c \le b;\ b-c < a \le b,
\end{aligned}
$$

and

$$
\begin{aligned}
&P[K < c]G(t; c+a-1) + \sum_{k=c}^{\infty} \pi_k G(t, k+a) \le \\
&\quad \sum_{k=c}^{\infty} \pi_k G(t; k+a) + \sum_{k=0}^{c-1} \pi_k G(t; c+a-k) \le \Pr[T(a,b) \le t] \le \\
&\quad P[K < c]G(t; a) + \sum_{k=c}^{\infty} \pi_k G(t; k-c+a+1), \quad 1 \le c \le b;\ 0 < a \le b-c,
\end{aligned}
$$

where

$$
\begin{aligned}
G(t; i) &= \int_0^t g(x; i)\,dx, \\
H(t; i, j, k) &= \int_0^t h(x; i, j, k)\,dx,
\end{aligned}
$$

These inequalities can be used to establish the following bounds on $E[T(a,b)]$,

$$E[W]+\frac{b-1}{\mu c}+\sum_{k=b-a+1}^{b}1/(k\mu)\geq E[W]+\sum_{k=0}^{c-1}\pi_k\frac{\min(c-k,b)}{\mu c}+\sum_{k=b-a+1}^{b}1/(k\mu)$$

$$\geq E[T(a,b)]\geq E[W]+\sum_{k=b-a+1}^{b}1/(k\mu),\qquad 1\leq b\leq c,$$

$$E[W]+\frac{c-1}{\mu c}+\sum_{k=b-a+1}^{c}1/(k\mu)\geq$$

$$E[W]+\sum_{k=0}^{c-1}\pi_k\frac{c-k}{\mu c}+\sum_{k=b-a+1}^{c}1/(k\mu)\geq$$

$$E[T(a,b)]\geq E[W]+\frac{b-c}{\mu c}+\sum_{k=b-a+1}^{c}1/(k\mu),\quad 1\leq c\leq b;\ b-c<a\leq b,$$

and

$$E[W]+\frac{c-1+a}{\mu c}\geq$$

$$E[W]+\sum_{k=0}^{c-1}\pi_k\frac{c-k+a}{\mu c}+P[L\geq c]\frac{c-1+a}{\mu c}\geq$$

$$E[T(a,b)]\geq E[W]+a/(\mu c),\qquad 1\leq c\leq b;\ 0<a\leq b-c.$$

References

[1] F. Baccelli, A.M. Makowski, A. Shwartz, "The Fork-Join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds",to appear in *Advances in Applied Probability.*

[2] F. Baccelli, W.A. Massey, D. Towsley, "Acyclic Fork-Join queueing networks", *J. Assoc. Comp. Mach.*, **36**,3, 615-642, July 1989.

[3] F. Baccelli, Z. Liu, "On the Execution of Parallel Programs on Multiprocessor Systems - A Queueing Theory Approach," INRIA Rapport de Recherche, No. 833, 1988. *to appear in J. Assoc. Comp. Mach.*

[4] S. Halfin, "Batch Delays Versus Customer Delays", *Bell System Technical Journal*, **62**, 7, pp. 2011-2016, September 1983.

[5] Harrison, P.G., "On Normalizing Constants in Queueing Networks", *Operations Research*, **33**, 2, 464-468, March-April 1985.

[6] Koenigsberg, E., "Comments on Normalizing Constants in Queueing Networks", *Operations Research*, **34**, 2, 330, March-April 1986.

[7] Nelson, R., Towsley, D., Tantawi, A.N., "Performance Analysis of Parallel Processing Systems," *IEEE Transactions on Software Engineering*, **14**, 4, 532-540, April 1988.

[8] R. Nelson, A.N. Tantawi, "Approximate analysis of Fork/Join synchronization in parallel queues", *IEEE Trans. Computers*, C-37, pp. 739-743, 1988.

[9] D. Towsley and S.P. Yu, "Bounds on the Performance of Two Server Fork-Join Queueing Systems", COINS Technical Report 87-123, Nov. 1987.

[10] W. Whitt, "Comparing Batch Delays and Customer Delays", *Bell System Technical Journal*, **62**, 7, pp. 2001-2009, September 1983.

[11] Yao, D.D., "Some Results for the Queues $M^X/M/c$ and $GI^X/G/c$," *Operations Research Letters*, **4**, 79-83, July 1985.