BOUNDS FOR TWO SERVER FORK-JOIN QUEUEING SYSTEMS

Don Towsley and Shou-Pin Yu

COINS Technical Report 87-123

November 23, 1987

BOUNDS FOR TWO SERVER FORK-JOIN QUEUEING SYSTEMS[1]

Don Towsley[2] and Shou-Pin Yu[3]

Abstract

We consider a two server system that processes a mixture of regular customers and fork/join customers. A regular customer is one that requires service at one server whereas a fork/join customer requires service at both servers. We study the behavior of this system operating under two different scheduling policies for regular customers. The *distributed scheduling policy (DSP)* routes regular customers probabilistically to either server. Customers are served in a First-come First-serve (FCFS) manner at each server. The *centralized scheduling policy (CSP)* places all regular and fork/join customers into a single queue. Customers are scheduled from this queue in a FCFS manner. We use a matrix-geometric formulation to model the system operating under each policy. For the first policy, we obtain the exact distribution of the sojourn time of a regular customer and tight bounds on the sojourn time distribution of a fork/join customer. For the second policy, we obtain tight bounds for the average sojourn time of either class of customer. Last, we compare the performance of both policies to each other.

[1] This work was supported in part by the National Science Foundation under contracts ECS 8406402 and SBD-8418216

[2] Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003.

[3] Digital Equipment Corporation, 110 Spit Brook Rd., Nashua, NH 03060.

1 Introduction

We consider a queueing system consisting of two servers that process two classes of customers. The first class consists of *regular* customers that can be processed at either of the two servers. The second class consists of *fork/join* customers that require service from both servers. A fork/join customer generates two tasks (fork), one for each server. After completion of service, (i.e., completion of both tasks), the customer departs.

Our interest in this queueing system is motivated primarily by the following analysis problem in computer systems. Numerous fault tolerant disk I/O systems (see [2] for an example) require that two copies of each data item be maintained. As a consequence, whenever the data item is updated, both copies must be modified. On the other hand, a request to read a data item can be satisfied by either copy. This system maps into the model of interest to us where the update requests correspond to fork/join customers and read requests correspond to regular customers. Other applications abound in the area of parallel processing where two processors serve a mix of serial and parallel programs.

We are interested in two variations of this system that differ from each other according to the policy used to schedule regular customers. The first policy associates a queue with each server (Figure 1(a)). Regular customers are assigned to each queue according to a Bernoulli process. Fork/join customers generate two tasks; each task entering each of the two queues. Customers are served in a first-come first-serve (FCFS) manner at each queue. We shall refer to this as the *distributed schsduling policy (DSP)*. Under the second policy both regular and fork/join customers enter a *single central queue* (Figure 1(b)). Customers are removed from this queue for service in a FCFS manner. We shall refer to this as the *centralized scheduling policy (CSP)*. In the context of our fault tolerant I/O system, either policy is reasonable since a read request may be satisfied by either copy of the data. Some applications in parallel processing may be handled by both policies whereas others may be only be handled by DSP. This latter case can occur when each server can only provide specialized service unavailable from the other server.

Although the second policy appears to require only one queue, we shall observe that a second queue is required for the server that lags behind while processing fork/join tasks. This second queue contains only fork/join tasks except possibly the request in service. We will also observe that CSP performs better than DSP.

Queueing systems with fork/join customers have only recently received attention. Flatto and Hahn [4] performed an exact analysis of a system with two exponential servers that process only fork/join customers. Fork/join customers arrive according to a Poisson process and the complete system is modeled by a Markov chain containing two state variables each of which is

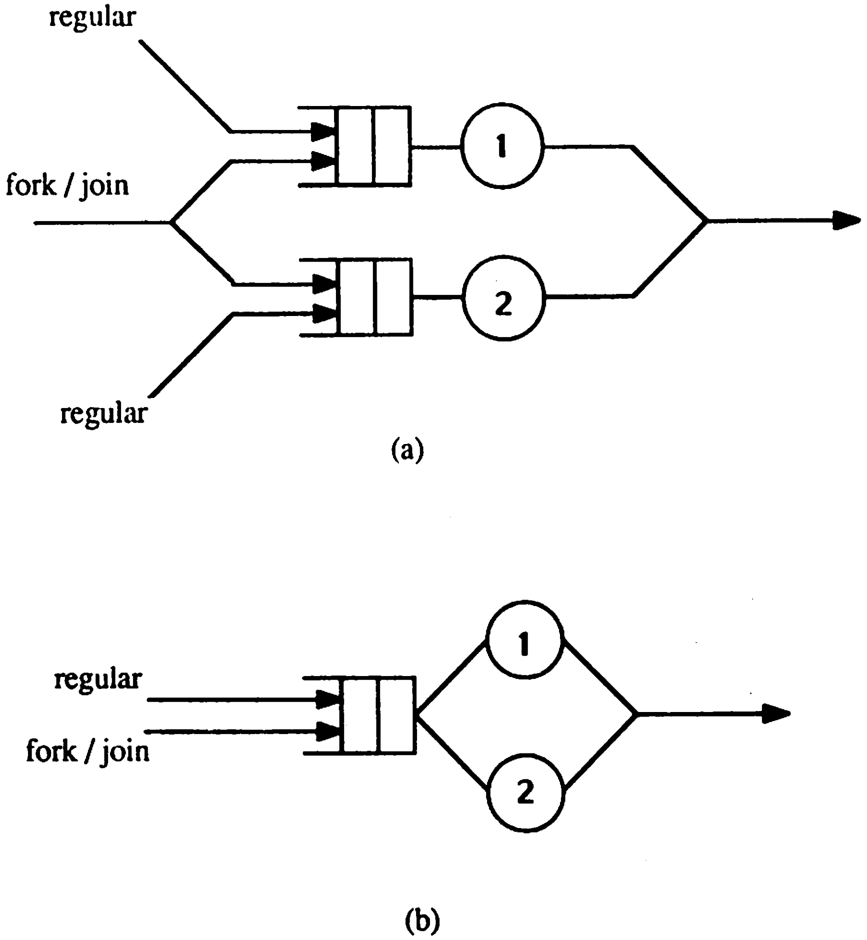

Figure 1: (a) Distributed Scheduling Policy (DSP). (b) Centralized Scheduling Policy (CSP).

unbounded. They obtain the stationary probability distribution for this chain by transforming the problem into a boundary value problem. Rao and Posner, [10], performed an approximate analysis of this model by truncating one of the two state variables and solving the resulting model using the matrix geometric methodology developed by Neuts [9]. We will show in the course of this paper that their approximations to the sojourn time and queue length distributions for fork/join customers provide lower bounds on the correct distributions.

Baccelli and Makowski [3] developed simple computational bounds for a K server system that processes a single class of fork/join customers where each customer requires service at all K servers. Using a different approach, Nelson and Tantawi [8] developed an accurate approximation for the case of K identical servers.

With the exception of the approach used by Rao and Posner [10] none of the approaches appear to generalize to systems that also include regular customers. The first of two contributions made in this paper is to model the behavior of both the DSP and the CSP in the face of regular customers. As described earlier these policies have wide application in computer systems.

The second contribution of this paper is the approach used to model these polcies. We develop tight bounds on the sojourn time distribution of fork/join customers under both policies. This is accomplished by judiciously truncating one of the state variables in the Markov chains underlying both policies. The presence of these bounds provides an analyst the capability of approximating the statistics of the sojourn time of fork/join customers to the degree of accuracy required by changing the truncation parameter. For both policies an increase in accuracy is obtained with an increase in computational cost. We point out that, although the truncated Markov chain for the DSP is similar to that studied by Rao and Posner, [10], they were apparently unaware that their approximation could be used to provide bounds on the sojourn time distribution in the system using DSP. We believe that this bounding technique has wide application to other problems, [5,7].

The remainder of the paper is structured in the following way. Section 2 contains the description of the model for the DSP along with its analysis. Section 3 contains the description and analysis of the CSP. A summary of the paper is contained in Section 4.

2 The Distributed Scheduling Policy

We consider a queueing system containing two servers $i = 1, 2$, each with its own queue. The service time at server i is an exponential random variable with mean $1/\mu_i$, $i = 1, 2$. Regular class customers arrive according to a Poisson process with rate λ. At the time of arrival, a regular customer chooses server i with probability α_i , $i = 1, 2$ ($\alpha_1 + \alpha_2 = 1$) and enters the appropriate queue. Fork/join customers also arrive according to a Poisson process with rate γ. At the time of arrival, a fork/join customer generates two tasks that enter each of the queues. Customers are served in a first come first serve (FCFS) manner at each queue. Let q_r and q_f denote the probability that a customer is a regular customer or a fork/join customer respectively; then $q_r = \lambda/(\lambda+\gamma)$ and $q_f = \gamma/(\lambda+\gamma)$. Last, in order to avoid confusion, we will think of a regular customer as always generating a single task.

Let $W_i^{(r)}$ and $W_i^{(f)}$ denote the sojourn times of the i-th regular and fork/join customers repectively. Let $\mathbf{T}_i = (T_{i,1}, T_{i,2})$ where $T_{i,j}$ denotes the sojourn time of the task generated by the i-th fork/join customer that enters queue j, $j = 1,2$. Here $T_{i,j}$ can be expressed as

$$T_{i,j} = U_{i,j} + X_{i,j} \tag{1}$$

where $U_{i,j}$ is the unfinished work in the queue at the time that the i-th fork/join customer arrives and $X_{i,j}$ is the service time for the task that it generates, $j = 1,2$. The sojourn time of the i-th fork/join customer can be expressed as $W_i^{(f)} = \max\{T_{i,1}, T_{i,2}\}$. Last, let $\hat{\mathbf{T}}_i = (\hat{T}_{i,1}, \hat{T}_{i,2})$ where $\hat{T}_{i,1} = \min\{T_{i,1}, T_{i,2}\}$ and $\hat{T}_{i,2} = \max\{T_{i,1}, T_{i,2}\}$. The sojourn time of the i-th fork/join customer can also be expressed as $W_i^{(f)} = \hat{T}_{i,2}$.

We are interested in the limiting random variables for the above defined random variables when they exist. We shall drop the subscript i when referring to these limiting random variables, i.e., $W^{(r)} = \lim_{i\to\infty} W_i^{(r)}$. We are also interested in the random variables $N^{(r)}$ and $N^{(f)}$ that respectively denote the stationary number of regular customers and fork/join customers in the system.

We first observe that each queue and server can be separately modeled as an M/M/1 system. As a consequence, the system exhibits stationary behavior so long as $\alpha_i\lambda + \gamma < \mu_i$, $i = 1,2$. Since the sojourn time of a regular customer is affected only by the queue that it enters, the distribution of $W^{(r)}$ is given by a weighted sum of the distributions of two independent M/M/1 systems

$$P[W^{(r)} \le w] = 1 - \alpha_1 e^{-(\mu_1-\alpha_1\lambda-\gamma)w} - \alpha_2 e^{-(\mu_2-\alpha_2\lambda-\gamma)w} \tag{2}$$

with mean

$$E[W^{(r)}] = \alpha_1/(\mu_1 - \alpha_1\lambda - \gamma) + \alpha_2/(\mu_2 - \alpha_2\lambda - \gamma). \tag{3}$$

The expected number of regular customers in the system, $E[N^{(r)}]$, can be obtained through an application of Little's rule [6]. Consequently we focus only on the behavior of fork/join customers.

As a further consequence of the fact that each queue behaves as an M/M/1 system, we can write the following expressions for the *marginal* distributions of the sojourn times of the two tasks associated with a fork/join customer

$$P[T_j \le t] = 1 - e^{-(\mu_j - \alpha_j\lambda - \gamma)t}, \;\; j = 1, 2. \tag{4}$$

with means

$$E[T_j] = 1/(\mu_j - \alpha_j\lambda - \gamma), \;\; j = 1, 2. \tag{5}$$

Let us now conduct the following experiment; select a random fork/join customer. Select one of the two tasks associated with this customer with equal probability. Denote the sojourn time of this task by T. Then T has the following distribution,

$$P[T \le t] = (P[T_1 \le t] + P[T_2 \le t])/2. \tag{6}$$

This randomly chosen task is equally likely to be the first or the last of the tasks associated with the customer to complete. Consequently we also have the following identity,

$$P[T \le t] = (F_{min}(t) + F_{max}(t))/2 \tag{7}$$

where $F_{min}(t) = P[\hat{T}_1 \le t]$, and $F_{max}(t) = P[\hat{T}_2 \le t]$. If we are able to obtain the marginal distribution for either $\hat{T}_1$ or $\hat{T}_2$, the above identity allows us to obtain the marginal distribution for the other random variable.

Equation (7) is also useful in obtaining bounds. For example, let us assume that we have a function $F^{(lb)}_{min}(t)$ that lower bounds $F_{min}(t)$, i.e.,

$$F^{(lb)}_{min}(t) \ge F_{min}(t), \;\; 0 \le t.$$

Then equation (7) can be used to obtain the following upper bound for $F_{max}(t)$

$$F_{max}(t) \le 2P[T \le t] - F^{(lb)}_{min}(t), \;\; t \ge 0. \tag{8}$$

In a similar manner, if we have expressions $F^{(ub)}_{min}(t)$ and $F^{(ub)}_{max}(t)$ that bound $F_{min}(t)$ and $F_{max}(t)$ from above, then equation (7) allows us to obtain the following lower bounds on the last two distributions

$$\begin{aligned} F_{min}(t) &\ge 2P[T \le t] - F^{(ub)}_{max}(t), && (9) \\ F_{max}(t) &\ge 2P[T \le t] - F^{(ub)}_{min}(t). && (10) \end{aligned}$$

We shall make use of these relationships in order to obtain bounds on the statistics of the sojourn time of a fork/join customer.

The DSP can be modeled as a Markov chain with state $\mathbf{N}(t) = (N_1(t), N_2(t))$ where $N_1(t)$ and $N_2(t)$ are the number of tasks in the queues associated with servers 1 and 2 respectively at time t. Let $q(i,j) = \lim_{t\to\infty} P[N_1(t) = i,\ N_2(t) = j]$. The stationary probabilities satisfy the following equations,

$$\begin{array}{rcll}
(\gamma+\lambda)q(0,0) & = & \mu_1 q(1,0) + \mu_2 q(0,1), & \\
(\gamma+\lambda+\mu_1))q(i,0) & = & \lambda\alpha_1 q(i-1,0) + \mu_1 q(i+1,0) + \mu_2 q(i,1), & i = 1,\cdots, \\
(\gamma+\lambda+\mu_2)q(0,j) & = & \lambda\alpha_2 q(0,j-1) + \mu_2 q(0,j+1) + \mu_1 q(1,j), & j = 1,\cdots, \\
(\gamma+\lambda+\mu_1+\mu_2)q(i,j) & = & \gamma q(i-1,j-1) + \lambda\alpha_1 q(i-1,j) + \lambda\alpha_2 q(i,j-1) & \\
 & & +\mu_1 q(i+1,j) + \mu_2 q(i,j+1), & i = 1,\cdots;\ j = 1,\cdots.
\end{array} \tag{11}$$

Unfortunately, this model is not amenable to a simple analysis. We focus instead on a modified system in which the second queue (associated with server 2) can hold no more than B tasks. Whenever a fork/join customer arrives to the system at time t and finds $N_2(t) = B$, he generates a *single* task that enters the first queue. The customer completes when this task completes. Similarly, a regular customer that arrives to a full queue at the second server is rejected.

This modified system can be modeled as a Markov chain with the same state defintion. In order to distinguish the modified system from the true system, we shall use the superscript (lb), i.e., $\mathbf{N}^{(lb)}(t)$ instead of $\mathbf{N}(t)$. We define $\mathbf{T}_i^{(lb)}$ according to equation (1) even though this does not produce the correct sojourn time at the second queue.[4] We shall describe an ordering relationship between $\mathbf{T}$ and $\mathbf{T}^{(lb)}$. We first introduce the following definitions [11].

Definition 1 *Let X and Y be two real valued random variables. X stochastically dominates Y $(X \geq_{st} Y)$ iff*

$$P[X \leq x] \leq P[Y \leq x], \quad -\infty < x < \infty. \tag{12}$$

We now introduce the D-ordering relation for vector valued random variables.

Definition 2 *Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ be two real valued vector random variables. We define the relation $\mathbf{X} \geq_D \mathbf{Y}$ to hold iff*

$$P[X_1 \leq x_1, \cdots, X_n \leq x_n] \leq P[Y_1 \leq x_1, \cdots, Y_n \leq x_n], -\infty < x_i < \infty,\ i = 1, \cdots, n. \tag{13}$$

[4] This is because we have defined the sojourn time at queue 2 to be 0 when the queue length is B at the time a customer arrives. Equation (1) produces a non-zero sojourn time for that event.

Note that when $n = 1$, the orderings $\geq_{st}$ and $\geq_D$ are the same. However, the usual definition of stochastic dominance for vector random variables is different than for D-ordering (see [11] for details).

D-ordering exhibits the following useful property.

Property 1 *Let* $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ *and* $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ *and define* $X_{max} = \max\{X_1, \cdots, X_n\}$, $X_{min} = \min\{X_1, \cdots, X_n\}$, $Y_{max} = \max\{Y_1, \cdots, Y_n\}$ *and* $Y_{min} = \min\{Y_1, \cdots, Y_n\}$. *If* $\mathbf{X} \geq_D \mathbf{Y}$ *then* $X_{max} \geq_{st} Y_{max}$ *and* $X_{min} \geq_{st} Y_{min}$.

We now state and prove the following theorem.

Theorem 1 *The following relationships hold between the real system and the modified system.*

1. $\mathbf{N} \geq_D \mathbf{N}^{(lb)}$,

2. $\mathbf{T} \geq_D \mathbf{T}^{(lb)}$.

Proof. In order to prove 1), it is useful to study the queue lengths of the system prior to the arrival of a customer of either class. Let $\mathbf{M}_i^{(lb)} = (M_{i,1}^{(lb)},\ M_{i,2}^{(lb)})$ and $\mathbf{M}_i = (M_{i,1},\ M_{i,2})$ where $M_{i,j}$ and $M_{i,j}^{(lb)}$ are the number in queue j $(j = 1, 2)$ prior to the arrival of customer i in the real system and the modified system respectively. These r.v.'s satisfy the following recurrences,

$$M_{i+1,j} = (M_{i,j} + A_{i,j} - D_{i,j})^+ \; j = 1, 2, \tag{14}$$

$$M_{i+1,1}^{(lb)} = M_{i+1,1}, \tag{15}$$

$$M_{i+1,2}^{(lb)} = (M_{i,2}^{(lb)} + A_{i,2} I_{i,2} - D_{i,2})^+ \tag{16}$$

where

$$A_{i,j} = \begin{cases} 0 & \text{i-th customer does not generate task for j-th queue} \\ 1 & \text{otherwise} \end{cases}$$

$$I_{i,2} = \begin{cases} 0 & \text{queue 2 is full at time of arrival of customer i} \\ 1 & \text{otherwise} \end{cases}$$

and $D_{i,j}$ is the number of departures from queue j between arrivals given an infinite number of tasks in queue j.

If the initial state vectors of the two systems satisfy $\mathbf{M}_0 \geq \mathbf{M}_0^{(lb)}$, an induction argument can be used to show $\mathbf{M}_i \geq \mathbf{M}_i^{(lb)}$ for $i = 0, 1, 2, \cdots$.[5] Consequently we conclude that $\mathbf{M}_i \geq_{st} \mathbf{M}_i^{(lb)}$ for $i = 0, 1, 2, \cdots$. Whenever the real system is ergodic, i.e., the r.v.'s $\mathbf{M}_i$ and $\mathbf{M}_i^{(lb)}$ converge to the limiting r.v.'s $\mathbf{M}$ and $\mathbf{M}^{(lb)}$, then $\mathbf{M} \geq_{st} \mathbf{M}^{(lb)}$. Finally, since arrivals from a Poisson process see time averages, $\mathbf{M}$ and $\mathbf{M}^{(lb)}$ have the same joint distribution as $\mathbf{N}$ and $\mathbf{N}^{(lb)}$ and we conclude that $\mathbf{N} \geq_D \mathbf{N}^{(lb)}$.

The second relationship is proven in a similar manner by focusing on the Lindley equations that must be satisfied by the unfinished work in the system at the time of customer arrivals.
QED

Remark. It is possible to show that $\mathbf{N}(t) \geq_D \mathbf{N}^{(lb)}(t)$ for $0 \leq t$ where $\mathbf{N}(t)$ and $\mathbf{N}^{(lb)}(t)$ are the queue lengths of the two systems at time $t \geq 0$ for any arrival process and i.i.d sequences of service times.

Before we consider the problem of approximating the sojourn time distribution, we obtain the stationary probabilities for the modified system, $P[\mathbf{N}^{(lb)} = (i, j)] = q(i, j)$, $i = 0, 1, \cdots$; $j = 0, 1, \cdots, B$. We define the steady state probability vector $Y = (y(0), y(1), y(2), \cdots)$ where $y(i)$ is a $(B+1)$ element vector, $y(i) = (q(i, 0), q(i, 1), \cdots, q(i, B))$, $i = 0, 1, \cdots$. The infinitesimal generator Q, satisfying $YQ = 0$, is listed below

$$Q = \begin{bmatrix} B_1 & A_0 & 0 & 0 & \cdots \\ B_2 & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{17}$$

where the matrices B_1, B_2, A_0, A_1, and A_2 are defined as $(B+1) \times (B+1)$ matrices,

$$A_0 = \begin{bmatrix} \alpha_1\lambda & \gamma & & & \\ 0 & \alpha_1\lambda & \gamma & & \\ & & \ddots & & \\ & & & & \alpha_1\lambda + \gamma \end{bmatrix}$$

[5] Here two vectors $\mathbf{V} = (v_1, \cdots, v_n)$ and $\mathbf{V}' = (v'_1, \cdots, v'_n)$ satisfy the relation $\mathbf{V} \geq \mathbf{V}'$ iff $v_i \geq v'_i$, $1 \leq i \leq n$.

$$B_1 = \begin{bmatrix} -(\gamma+\lambda) & \lambda\alpha_2 & 0 & & & \\ \mu_2 & -(\gamma+\mu_2+\lambda) & \alpha_2\lambda & & & \\ 0 & \mu_2 & -(\gamma+\mu_2+\lambda) & & & \\ & & \ddots & \ddots & & \\ & & & \mu_2 & -(\gamma+\mu_2+\lambda) & \alpha_2\lambda \\ & & & & \mu_2 & -(\gamma+\mu_2+\alpha_1\lambda) \end{bmatrix}$$

$$B_2 = A_2 = \mu_1 I_{(B+1)\times(B+1)}$$

$$A_1 = \begin{bmatrix} c_0 & \alpha_2\lambda & & & & \\ \mu_2 & c_1 & \alpha_2\lambda & & & \\ 0 & \mu_2 & c_1 & \alpha_2\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mu_2 & c_1 & \alpha_2\lambda \\ & & & & \mu_2 & c_2 \end{bmatrix} \qquad \begin{array}{rcl} c_0 & = & -(\gamma+\mu_1+\lambda) \\ c_1 & = & -(\gamma+\mu_1+\mu_2+\lambda) \\ c_2 & = & -(\gamma+\mu_1+\mu_2+\alpha_1\lambda) \end{array}$$

Define the matrix $A = A_0 + A_1 + A_2$. Neuts [9] showed that Q is positive recurrent if $\pi A_0 e < \pi A_2 e$, where π is the unique solution to $\pi A = 0$, $\pi e = 1$. Here e is a column vector containing $B+1$ ones and π is a $B+1$ element vector containing the stationary queue length distribution for the M/M/1/B queue with arrival rate $\gamma + \alpha_2\lambda$ and service rate μ_2, i.e., the i-th element of π is $(1-u)u^i/(1-u^{(B+1)})$ where $u = (\alpha_2\lambda+\gamma)/\mu_2$. In the modified model, the condition for positive recurrence is $\gamma + \alpha_1\lambda < \mu_1$. Whenever Q is positive recurrent, the stationary probability vector Y satisfies the matrix-geometric form

$$y(i) = y(0)R^i, \;\; i = 0, 1, \cdots.$$

where R is the minimal solution of $A_0 + RA_1 + R^2A_2 = 0$. The matrix R can be obtained iteratively by the following approach. Let $R(n)$ denote the value of R after n iterations.

$$\begin{array}{rcl} R(0) & = & 0, \\ R(n+1) & = & -A_0A_1^{-1} - R^2(n)A_2A_1^{-1}, \quad n > 0. \end{array}$$

Rao and Posner [10] have shown that the vector $y(0)$ takes the form

$$y(0) = \pi(I - R).$$

So far the preceding analysis differs from that presented by Rao and Posner only in the definition of the submatrices B_1, B_2, A_0, A_1, A_2. They derive the following expression for the joint distribution of $\mathbf{T}^{(lb)}$

$$T^{(lb)}(w_1, w_2) = \pi[I - \exp(-\mu_1(I-R)w_1)B(w_2) \tag{18}$$

where I is the identity matrix and $B(w_2)$ is a $(B+1)$ column vector with i-th component $\left[1 - \sum_{r=0}^{i}(\mu_2 w_2)^r/r!\ \exp(-\mu_2 w_2)\right]$ An application of Theorem 1 along with Property 1 of stochastic dominance yields the following bound on $F_{max}(w)$,

$$F_{max}(w) \leq T^{(lb)}(w,w) = \pi[I - \exp(-\mu_1(I-R)w))B(w). \tag{19}$$

Similar arguments can be used to obtain the following bound on the distribution of the time until the first of the two tasks associated with a fork/join customer complete,

$$F_{min}(w) \leq 1 - \pi\ \exp(-\mu_1(I-R)w - \mu_2 w)C(w) \tag{20}$$

where $C(w)$ is a $(B+1)$ element column vector with ordered components $\sum_{j=0}^{i}(\mu_2 w)^j/j!$, $i = 0, 1, \cdots, B$. Substitution of the above lower bound into equation (7) yields the following upper bound on $F_{max}(w)$,

$$\begin{aligned} F_{max}(w) \quad \leq \quad & 2 - 2\exp(-(\mu_1 - \alpha_1\lambda - \gamma) - 2\exp(-\mu_2 - \alpha_2\lambda - \gamma) \\ & + \pi\exp(-\mu_1(I-R)w - \mu_2 w)C(w). \end{aligned} \tag{21}$$

These can be used to obtain bounds on the moments of the sojourn time.

We have been unable to develop similar bounds for the distribution of the number of fork/join customers in the system except for the case that $\lambda = 0$. However, Little's rule can be applied to obtain bounds on the average number of fork/join customers in the system.

We end this section with a derivation of bounds on the fork/join queue length distribution for the case $\lambda = 0$. Let $\hat{\mathbf{N}} = (\hat{N}_1, \hat{N}_2)$ where $\hat{N}_1 = \max\{N_1, N_2\}$ and $\hat{N}_2 = \min\{N_1, N_2\}$. Let $G_{max}(k) = P[\hat{N}_1 \leq k]$ and $G_{min}(k) = P[\hat{N}_2 \leq k]$, $k = 0, 1, \cdots$. Define $\hat{\mathbf{N}}^{(lb)}$, $G_{max}^{(lb)}(k)$ and $G_{min}^{(lb)}(k)$, $k = 0, 1, \cdots$ in a similar manner. Because $\mathbf{N} \geq_{st} \mathbf{N}^{(lb)}$ (Theorem 1), it follows that $G_{max}(k) \leq G_{max}^{(lb)}(k)$ and $G_{min}(k) \leq G_{min}^{(lb)}(k)$, $k = 0, 1, \cdots$ where

$$G_{max}^{lb)}(k) \quad = \quad \begin{cases} \sum_{j=0}^{k}\left[\sum_{i=0}^{j} q(i,j) + \sum_{i=0}^{j} q(j,i)\right] & ,k \leq B \\ \pi(I-R)[R^{B+1} - R^{k+1}](I-R)^{-1}e & ,k > B, \end{cases} \tag{22}$$

$$G_{min}^{(lb)}(k) \quad = \quad \begin{cases} \sum_{j=0}^{k}\left[\pi R^{(j+1)}(I-R)^{-1}e_j + \sum_{i=j}^{B} q(j,i)\right] & ,k < B \\ 1 & ,k \geq B. \end{cases} \tag{23}$$

Here e is a $(B+1)$ column vector consisting of all 1's and e_j is a $(B+1)$ column vector consisting of all 0's except the i-th position which contains a 1. The derivation of the expression for $F_{max}^{(lb)}(k)$

ρ	$B = 4$		$B = 8$		$B = 16$		$B = 32$	
	l.b.	u.b.	l.b.	u.b.	l.b.	u.b.	l.b.	u.b.
.1	1.653	1.653	-	-	-	-	-	-
.2	1.843	1.844	1.844	1.844	-	-	-	-
.3	2.082	2.092	2.089	2.089	-	-	-	-
.4	2.380	2.431	2.415	2.417	2.417	2.417	-	-
.5	2.764	2.925	2.861	2.879	2.875	2.875	-	-
.6	3.287	3.709	3.494	3.586	3.560	3.563	3.562	3.562
.7	4.093	5.103	4.444	4.823	4.676	4.716	4.708	4.708
.8	5.651	8.088	6.122	7.518	6.711	7.102	6.982	7.003
.9	10.37	17.63	10.81	16.55	11.73	15.19	13.05	14.17

Table 1: Performance bounds for different values of B, $\mu_1 = \mu_2 = 1$.

is found in [10]. Finally, the argument used to produce the lower bound for $F_{max}(k)$ (equation (10)) can also be used to obtain the following lower bound on $G^{(lb)}_{max}(k)$,

$$G_{max}(k) \geq 2 - (\gamma_1/\mu_1)^{k+1} - (\gamma_2/\mu_2)^{k+1}) - G^{(lb)}_{min}(k), \quad k = 0, 1, \cdots. \tag{24}$$

Table 1 gives bounds on the average sojourn time of a fork/join customer for different values of B. In this example, no regular customers enter the system and both servers are identical. If we take the average of the bounds for an approximation, the error is less than 3% for $\rho \leq .8$ when $B = 16$. An error of less than 5% can be achieved for $\rho = 0.9$ by taking $B = 32$. Table 2 gives bounds on the average sojourn time of a fork/join customer for different values of μ_1/μ_2 when $\mu_1 = 1$. Here, the value $B = 32$ was used. For values of $\mu_2 > 1.5\mu_1$, the resulting bounds are identical to at least three decimal places. As expected the average sojourn time is a decreasing function of μ_2. Table 3 presents bounds for the average sojourn time of a fork/join customer in a system with identical servers that also serves regular customers. Here we observe that for a fixed server load, the average sojourn time of a fork/join customer increases as the fraction of regular customers increases. This is because the coupling of the arrival processes to the two queues decreases as there are fewer fork/join customers. Again B was chosen to be 32.

3 The Centralized Scheduling Policy

In this system, both regular and fork/join customers enter a single queue at the time of their arrival. We assume that customers are served in a FCFS manner from this queue. Let us

ρ	$\mu_1/\mu_2 = 1$	$\mu_1/\mu_2 = 2/3$	$\mu_1/\mu_2 = 1/2$	$\mu_1/\mu_2 = 1/3$
0.1	1.65	1.38	1.28	1.19
0.2	1.84	1.53	1.41	1.32
0.3	2.09	1.71	1.58	1.50
0.4	2.42	1.95	1.81	1.73
0.5	2.88	2.28	2.14	2.06
0.6	3.56	2.77	2.62	2.55
0.7	4.71	3.58	3.44	3.37
0.8	6.99±.01	5.21	5.08	5.03
0.9	13.61±.56	10.12	10.02	10.01

Table 2: Performance bounds for different values of μ_1/μ_2.

ρ	$q_r = 0$	$q_r = 1/4$	$q_r = 1/2$	$q_r = 3/4$	$q_r = 1$
0.1	1.65	1.65	1.66	1.66	1.67
0.2	1.84	1.85	1.85	1.86	1.88
0.3	2.08	2.10	2.11	2.12	2.14
0.4	2.42	2.43	2.45	2.47	2.50
0.5	2.88	2.89	2.92	2.95	3.00
0.6	3.56	3.59	3.63	3.68	3.75
0.7	4.71	4.76	4.82	4.89	5.00
0.8	6.99±.01	7.08±.01	7.18±.01	7.31±.01	7.50
0.9	13.61±.56	13.77±.55	13.98±.55	14.24±.55	15.00

Table 3: Performance bounds for different workload mixes, $\mu_1 = \mu_2 = 1$.

consider what happens to a customer when it comes to the head of the queue. If it is a regular customer, it begins service at the first server that becomes available. If it is a fork/join customer, it generates two tasks as soon as a server becomes available. One task enters the idle server while the other enters a second queue associated with the remaining server. Thus the system requires a second queue associated with the server that lags behind. This queue only contains fork/join tasks and is served in a FCFS manner.

Regular customers arrive at this system according to a time invariant Poisson process with rate λ. Fork/join customers arrive at this system according to a time invariant Poisson process with rate γ. We assume that the two servers are identical and that the service times at each server are exponential random variables with parameter μ. Let $N_1(t)$ $(0 \leq N_1(t))$ be the number of requests in the common queue, $N_2(t)$ $(0 \leq N_2(t))$ be the number of fork/join requests to the server that lags behind, and $N_3(t)$ $(N_3(t) = 0, 1)$ denote whether the other server is processing a request or not at time t. The state $\mathbf{N}(t) = (N_1(t), N_2(t), N_3(t))$ forms a Markov chain. If the system is stationary and we let $p(m,n,l) = \lim_{t \to \infty} P[\mathbf{N}(t) = (m,n,l)]$, then these probabilities satisfy the following equations,

$$\begin{array}{rcll}
(\lambda+\gamma)p(0,0,0) & = & \mu p(0,1,0), & \\
(\lambda+\gamma+\mu)p(0,1,0) & = & \mu p(0,2,0) + \lambda p(0,0,0) + 2\mu p(0,1,1), & \\
(\lambda+\gamma+\mu)p(0,n,0) & = & \mu p(0,n+1,0) + \mu p(0,n,1), & n = 2,\cdots, \\
(\lambda+\gamma+2\mu)p(0,1,1) & = & \mu p(0,2,1) + 2q_r\mu p(1,1,1) + \gamma p(0,0,0) + \lambda p(0,1,0), & \\
(\lambda+\gamma+2\mu)p(0,2,1) & = & \mu p(0,3,1) + 2q_r\mu p(1,1,1) + q_r\mu p(1,2,1) & \\
 & & + \gamma p(0,1,0) + \lambda p(0,2,0), & \\
(\lambda+\gamma+2\mu)p(0,n,1) & = & \mu p(0,n+1,1) + q_f\mu p(1,n-1,1) & \\
 & & + q_r\mu p(1,n,1) + \gamma p(0,n-1,0) + \lambda p(0,n,0), & n = 3,\cdots \\
(\lambda+\gamma+2\mu)p(i,1,1) & = & \mu p(i,2,1) + 2q_r\mu p(i+1,1,1) + (\lambda+\gamma)p(i-1,1,1), & i = 1,\cdots \\
(\lambda+\gamma+2\mu)p(i,2,1) & = & \mu p(i,3,1) + 2q_f\mu p(i+1,1,1) + q_r\mu p(i+1,2,1) & \\
 & & + (\lambda+\gamma)p(i-1,2,1), & i = 1,\cdots \\
(\lambda+\gamma+2\mu)p(i,n,1) & = & \mu p(i,n+1,1) + q_f\mu p(i+1,n-1,1) & \\
 & & + q_r\mu p(i+1,n,1) + (\lambda+\gamma)p(i-1,n,1), & i = 1,\cdots; n = 3,\cdots
\end{array}$$

We develop bounds on the average sojourn times for regular and fork/join customers by truncating one of the first two state variables, suitably modifying the infinitesimal generator and applying matrix-geometric techniques to the resulting model. Our computational experience indicates that we achieve greater accuracy for the same amount of computation by truncating $N(t)$. Consequently we report on this approach. Unfortunately there is insufficient symmetry in this system to allow us to obtain both optimistic and pessimistic bounds from a single model as we did for DSP. We describe and analyze separate models for each bound.

3.1 Optimistic Bound

In order to obtain optimistic bounds we impose the constraint $N_2(t) \leq B$. Whenever the second queue contains B fork/join requests and a new fork/join request arrives, it passes through without incurring any delay. Consequently, the fork/join customer associated with this request completes as soon as its other task completes at the other server.

This modified system can be modeled as a Markov chain with the same state description, $\mathbf{N}^{(lb)}(t) = (N_1^{(lb)}(t),\ N_2^{(lb)}(t),\ N_3^{(lb)}(t))$. Before we analyze this modified system we prove the following theorem.

Theorem 2 *The true system and the modified system satisfy the following relationships,*

1. $\mathbf{N} \geq_D \mathbf{N}^{(lb)}$,

2. $W^{(f)} \geq_{st} W^{(lb)(f)}$,

3. $W^{(r)} \geq_{st} W^{(lb)(r)}$.

Proof. In order to prove 1) we consider a discrete time Markov chain imbedded at the points in time prior to customer arrivals and service completions. Furthermore, we include events that correspond to the service completions of *fictitious* customers whenever the servers are empty. Let $\mathbf{M}_i = (M_{i,1}, M_{i,2}, M_{i,3})$ and $\mathbf{M}_i^{lb} = (M_{i,1}^{lb}, M_{i,2}^{lb}, M_{i,3}^{lb})$ where $M_{i,1}$ is the number of customers in the common queue, $M_{i,2}$ is the number of requests at the server with the longest queue, and $M_{i,3}$ is the number of customers at the other server immediately prior to the i-th event. $M_{i,j}^{lb}$ has a similar meaning for the model providing the optimistic bound, $j = 1, 2, 3$. These r.v.'s satisfy the following recurrences,

$$\begin{aligned}
M_{i+1,1} &= (M_{i,1} + A_i\chi(\min\{M_{i,2}, M_{i,3}\} > 0) - D_{3,i} - D_{2,i}\chi(M_{i,2} \leq 1))^+, \\
M_{i+1,2} &= (M_{i,2} - D_2 + A_i\chi(M_{i,2} = 0) + A_iF_i\chi(M_{i,3} = 0) + D_{3,i}F_i\chi(M_{i,1} > 0) \\
&\quad + D_{i,2}\chi(M_{i,2} \leq 1))^+ \\
M_{i+1,3} &= (M_{i,3} - D_{i,3} + A_iF_i\chi(M_{i,3} + A_iR_i\chi(M_{i,3} = 0)\chi(M_{i,2} \geq 1) \\
&\quad - D_{i,2}\chi(M_{i,2} = 1)\chi M_i, 1 = 0))^+ \\
M_{i+1,1}^{(lb)} &= (M_{i,1}^{(lb)} + A_i\chi(\min\{M_{i,2}^{(lb)}, M_{i,3}^{(lb)}\} > 0) - D_{3,i} - D_{2,i}\chi(M_{i,2}^{(lb)} \leq 1))^+, \\
M_{i+1,2}^{(lb)} &= (M_{i,2}^{(lb)} - D_2 + A_i\chi(M_{i,2}^{(lb)} = 0) + I_iA_iF_i\chi(M_{i,3}^{(lb)} = 0) + D_{3,i}F_i\chi(M_{i,1}^{(lb)} > 0) \\
&\quad + D_{i,2}\chi(M_{i,2}^{(lb)} \leq 1))^+
\end{aligned}$$

$$\begin{aligned} M_{i+1,3}^{(lb)} &= (M_{i,3}^{(lb)} - D_{i,3} + A_i F_i \chi(M_{i,3}^{(lb)} + A_i R_i \chi(M_{i,3}^{(lb)} = 0)\chi(M_{i,2}^{(lb)} \geq 1) \\ &\quad - D_{i,2}\chi(M_{i,2}^{(lb)} = 1)\chi M_i^{(lb)}, 1 = 0))^+ \end{aligned}$$

where $\chi(\mathit{predicate})$ is the indicator function that takes value 1 if *predicate* is true and 0 otherwise, A_i takes on value 1 if the event corresponds to an arrival, $D_{i,2}$ and $D_{i,3}$ take on value 1 if the event is a service completion from the server with the longest queue or the other queue respectively, F_i and R_i take value 1 if the customer is either a fork/join or regular customer. In addition I_i takes value 1 if the longest server queue exceeds $B - 1$ prior to the time that a fork/join customer begins service. If the state of the system initially satisfies $\mathbf{M}_0 \geq \mathbf{M}_0^{(lb)}$, an induction argument can be used to show $\mathbf{M}_i \geq \mathbf{M}_i^{(lb)}$ for $i = 0, 1, \cdots$. Consequently we conclude that $\mathbf{M}_i \geq_D \mathbf{M}_i^{(lb)}$ for $i = 0, 1, \cdots$. Whenever the real system is ergodic, i.e., the r.v.'s $\mathbf{M}_i$ and $\mathbf{M}_i^{(lb)}$ converge to the limiting r.v.'s $\mathbf{M}$ and $\mathbf{M}^{(lb)}$, then $\mathbf{M} \succ \underline{\mathbf{M}}^{(lb)}$. Last, it follows from the definition of $\mathbf{M}$ and $\mathbf{M}^{(lb)}$ that $\mathbf{N} \geq_D \mathbf{N}^{(lb)}$.

Similar arguments can be used to prove 2) and 3).
QED

We now obtain the stationary probability distribution for the modified system, $P[\mathbf{N}^{(lb)} = (m, n, l)] = q(m, n, l)$, $m = 0, 1, \cdots$; $n = 0, 1, \cdots, B$; $l = 0, 1$. We define the steady state probability vector $Y = (x, y(0), y(1), y(2), \cdots)$ where $x = [q(0,0,0)]$, $y(0)$ is the B element vector $[q(0,1,0), q(0,2,0), \cdots, q(0,B,0)]$, and $y(i)$ is a B element vector $y(i) = (q(i-1,1,1), \cdots, q(i-1,B,1))$, $i = 1, 2, \cdots$. The infinitesimal generator Q satisfying $YQ = 0$ is listed below,

$$Q = \begin{bmatrix} D_2 & C_1 & B_0 & 0 & 0 & \cdots \\ D_3 & C_2 & B_1 & 0 & 0 & \cdots \\ 0 & C_3 & A_2 & A_1 & 0 & \cdots \\ 0 & 0 & A_3 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{25}$$

where D_2 is a one element vector $D_2 = -[\lambda + \gamma]$, D_3 is a B element column vector $D_3 = [\mu, 0, \cdots, 0]^T$, B_0 and C_1 are B element vectors $B_0 = [\mu, 0, \cdots, 0]$, $C_1 = [\lambda, 0, \cdots, 0]$, and C_2, C_3, B_1, A_1, A_2, A_3 are $B \times B$ matrices

$$C_2 = \begin{bmatrix} -(\gamma + \lambda + \mu) & 0 & 0 & & \\ \mu & -(\gamma + \lambda + \mu) & 0 & & \\ 0 & \mu & -(\gamma + \lambda + \mu) & & \\ & & \ddots & \ddots & \\ & & & \mu & -(\gamma + \lambda\mu) \end{bmatrix}$$

$$C_3 = \begin{bmatrix} 2\mu & 0 & & \\ 0 & \mu & & \\ & \ddots & \ddots & \\ & & 0 & \mu \end{bmatrix}$$

$$B_1 = \begin{bmatrix} \lambda & \gamma & & & \\ & \lambda & \gamma & & \\ & & \ddots & \ddots & \\ & & & \lambda & \gamma \\ & & & & \lambda + \gamma \end{bmatrix}$$

$$A_1 = (\gamma + \lambda) I_{B \times B},$$

$$A_2 = \begin{bmatrix} -(\gamma + \lambda + 2\mu) & 0 & & \\ \mu & -(\gamma\lambda + 2\mu) & & \\ & \ddots \ddots & & \\ & & \mu & -(\gamma + \lambda + 2\mu) \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 2\mu\lambda/(\gamma+\lambda) & 2\mu\gamma/(\gamma+\lambda) & 0 & & \\ 0 & \mu\lambda/(\gamma+\lambda) & \mu\gamma/(\gamma+\lambda) & & \\ & \ddots & \ddots & & \\ & & & \mu\lambda/(\gamma+\lambda) & \mu\gamma/(\gamma+\lambda) \\ & & & & \mu \end{bmatrix}$$

Define the matrix $A = A_1 + A_2 + A_3$. The infinitesimal generator Q is positive recurrent if $\pi A_1 e < \pi A - 3e$, where π is the unique solution to $\pi A = 0$, $\pi e = 1$. The stationary probability vector Y satisfies the matrix-geometric form

$$y(i) = y(1) R^{i-1}, \; i = 1, 2, \cdots$$

where R is the minimal solution of $A_1 + RA_2 + RA_3 = 0$. The matrix R can be obtained in a similar manner as for the DSP. The vectors x, $y(0)$, and $y(1)$ are obtained by solving the following set of equations,

$$\begin{aligned} xD_2 + y(0)D_3 &= 0, \\ xC_1 + y(0)C_2 + y(1)C_3 &= 0, \\ xB_0 + y(0)B_1 + y(1)[A_2 + RA_3] &= 0, \\ x + [y(0) + y(1)[I - R]]e &= 1. \end{aligned}$$

In the remainder of this section, we will derive lower bounds on the expected number of regular customers and fork/join customers as well as the expected sojourn time for regular customers and fork/join customers. The expected length of the common queue, $E[N_1^{(lb)}]$, is

$$E[N_1^{(lb)}] = y(1)R[I - R]^{-2}e, \tag{26}$$

The average number of regular customers that are in service is bounded from above by λ/μ. Consequently, a lower bound on the expected number of regular customers is

$$E[N^{(r)}] \geq \lambda E[N_1^{(lb)}]/(\lambda + \gamma) + \lambda/\mu. \tag{27}$$

Little's result yields

$$E[W^{(r)} \geq E[N_1^{(lb)}]/(\lambda + \gamma) + 1/\mu. \tag{28}$$

The expected number of fork/join customers in the common queue of the modified system is $\gamma q E[N_1^{(lb)}]/(\lambda + \gamma)$. A lower bound on the expected number of fork/join customers that are in service is obtained by first determining the expected service delay incurred by a fork/join customer (i.e., time from beginning of processing of first task until completion of both tasks) in the modified system and then applying Little's result. Denote this expected delay by $d_1^{(lb)}$. The time required to service a fork/join customer depends on the length of the longest server queue and whether both servers are busy. If a fork/join customer begins service when the longest queue contains $i - 1$ customers ahead of him, then the time to complete service is denoted by $h(i)$ which satisfies the recurrence

$$\begin{aligned} &h(1) = 3/(2\mu), \\ &h(i) = 1/(2\mu) + h(i-1)/2 + i/(2\mu), \quad i = 2, 3, \cdots. \end{aligned}$$

The first term in the recurrence for $h(i)$ corresponds to the average delay until the first of the two servers completes. If the server with the queue of length i (including the new request) completes, then the fork/join customer observes the system with one less customer in the longest queue. His average delay in this case corresponds to the second term in the above recurrence. Last, observe that when a fork/join customer begins service, one of his requests will begin service in the server with the shortest queue. Consequently, if this server completes, the average delay of the fork/join customer will be due just to his service time and that of the $i - 1$ customers ahead of him in the queue of the other server. This gives rise to the third term. This recurrence has the following solution

$$h(i) = (i + 2^{-i})/\mu, \quad i = 1, \ldots \tag{29}$$

There are three possible scenarios when a fork/join customer arrives to the system. First, both servers may be completely empty; second one of the servers may be empty, and third neither server may be empty. In the first two cases, the customer initiates service immediately. In the last case, the customer begins service only when he is at the head of the common queue. If at this moment neither server has an outstanding queue, the customer begins service as soon as either server completes service of its customer. Otherwise, the customer begins service only if the server without the outstanding queue completes service.

The above observations yield the following expression for $d_1^{(lb)}$,

$$d_1^{(lb)} = q(0,0,0)h(1) + y(0)V_1 + \{y(1)[I-R]^{-1}e\}y(1)R[I-R]^{-1}V_2/\{y(1)R[I-R]^{-1}V_3\}$$

where

$$\begin{aligned} V_1 &= (h(2),h(3),\cdots,h(B),h(B+1))^T, \\ V_2 &= (2h(2),h(3),h(4),\cdots,h(B),h(B+1))^T, \\ V_3 &= (2,\underbrace{1,1,\cdots,1}_{B-1})^T. \end{aligned}$$

The coefficient 2 for the first element of V_2 and V_3 is a consequence of the observation that service of a new customer is initated whenever a departure occurs from either server and there is no outstanding queue.

A lower bound on the expected number of fork/join customers in the system is

$$E[N^{(f)}] \geq \gamma E[N_1^{(lb)}]/(\gamma+\lambda) + \gamma d_1^{(lb)}. \tag{30}$$

The lower bound on the expected fork/join customer sojourn time is given by an application of Little's result:

$$E[W^{(f)}] \geq E[N_1^{(lb)}]/(\gamma+\lambda) + d_1^{(lb)}. \tag{31}$$

Lower bounds on the expected sojourn times are found in Tables 4 and 5 for different mixes of regular and fork/join customers (the entries in Table 5 are both upper and lower bounds accurate to three places). These values were calculated for $B = 16$.

3.2 Pessimistic Bound

A pessimistic bound on the performance of CSP is obtained in a similar manner. The outstanding queue is allowed to have up to B requests. Whenever this queue is full and a request completes at the other server, the completed request is replaced by a fictitious request that initiates a new service period. This avoids the possible event of a fork/join customer at the head of the common queue placing a request in the outstanding queue and thus increasing its length above B. The resulting Markov chain is identical to the one described for the optimistic bound except for the following changes in the transition rates,

1. Remove state $(0, B, 0)$ and all transitions to and from it.
2. Remove the transition from state $(i, B, 1)$ to $(i-1, B, 1), i = 2, \cdots$.

Let $\mathbf{N}^{(ub)}(t) = (N_1^{(ub)}(t), N_2^{(ub)}(t), N_3^{(ub})(t))$ be the state of this new system. Let $W^{(ub)(r)}$ and $W^{(ub)(f)}$ denote the stationary sojourn times for a regular customer and fork/join customer respectively. We state the following theorem without proof.

Theorem 3 *The true system and the modified system satisfy the following relationships,*

1. $\mathbf{N}^{(ub)} \geq_D \mathbf{N}$,

2. $W^{(ub)(f)} \geq_{st} W^{(f)}$,

3. $W^{(ub)(r)} \geq_{st} W^{(r)}$.

Proof. The proof of this theorem is similar to that of theorem 2 and is omitted here.

The procedure for calculating the lower bounds on the average buffer occupancies and sojourn times in the previous section apply without change to the computation of upper bounds. Numerical results for these bounds can be found in Tables 4 and 5 for different mixes of regular and fork/join customers. Again B is taken to be 16. One observes that the bounds are tight for server utilizations less than 0.9 or when the fraction of regular customers exceeds 1/4.

4 Summary

In this paper we have developed models that can be used to bound the performance of fork/join queueing systems consisting of two servers. The modeling approach was to approximate a infinite two-dimensional Markov chain by another Markov chain such that only one dimension is unbounded. These approximations are constructed in such a way as to lead to either upper or lower bounds on the statistics of interest. Numerical results show that the bounds can be

	$q_r = 0$		$q_r = 1/4$			
ρ	$E[W_f^{(lb)}]$	$E[W_f^{(ub)}]$	$E[W_r^{(lb)}]$	$E[W_r^{(ub)}]$	$E[W_f^{(lb)}]$	$E[W_f^{(ub)}]$
0.1	1.65	1.65	1.03	1.03	1.65	1.65
0.2	1.84	1.84	1.07	1.07	1.84	1.84
0.3	2.09	2.09	1.14	1.14	2.06	2.06
0.4	2.42	2.42	1.25	1.25	2.35	2.35
0.5	2.87	2.88	1.42	1.42	2.74	2.74
0.6	3.56	3.56	1.70	1.70	3.28	3.28
0.7	4.68	4.72	2.20	2.20	4.12	4.12
0.8	6.73	7.10	3.26	3.27	5.65	5.66
0.9	11.12	15.34	6.61	6.76	9.64	9.81

Table 4: Performance bounds for CSP, $q_r = 0, 1/4$.

	$q_r = 1/2$		$q_r = 3/4$	
ρ	$E[W_r]$	$E[W_f]$	$E[W_r]$	$E[W_f]$
0.1	1.02	1.65	1.02	1.65
0.2	1.07	1.83	1.06	1.81
0.3	1.14	2.03	1.12	1.99
0.4	1.24	2.28	1.23	2.21
0.5	1.41	2.61	1.39	2.48
0.6	1.69	3.05	1.65	2.84
0.7	2.19	3.73	2.12	3.41
0.8	3.24	4.97	3.09	4.47
0.9	6.51	8.45	6.06	7.52

Table 5: Performance bounds for CSP, $q_r = 1/2, 3/4$.

quite accurate. Consequently, the performance of the two classes of fork/join queuing systems can be studied for different workloads, server speeds, etc...

This approach is of interest because it can be applied to many queueing problems in which the underlying Markov chain has two state variables that may take a countably infinite number of values. It has been our experience that a little thought is required in order to determine which state variable to truncate and in what manner so as to lead to performance bounds in either direction.

Last, the performance of CSP is appreciably better than the performance of CSP. This is illustrated in Figures 2 and 3 for $q_r = 1/4$ and $3/4$. One observes that the difference in performance increases as the fraction q_r of regular customers increases. This is expected since CSP corresponds to two independent M/M/1 queues when $q_r = 1$ whereas DSP is identical an M/M/2 queue.

References

[1] Baccelli, F. 1985. "Two parallel queues created by arrival with two demands: The M/G/2 symmetrical case". Report INRIA No. 426.

[2] Bartlett, J.F. 1981. "A Nonstop* Kernel," *Proc. Eigth Symp. on Operating System Principles*, pp. 22-29.

[3] Baccelli, F. and A. Makowski. 1985. "Simple computable bounds for the Fork-Join queue", *Proc. of the John Hopkins Conf. on Information Sciences and Systems.*

[4] Flatto L. and S. Hahn. 1984. " Two parallel queues created by arrivals with two demands I," *SIAM J. Appl. Math.*, vol. 44, pp. 1041-1053.

[5] Green, L. 1985. "A queueing system with general-use and limited -use servers," *Operations Research*, Vol. 33, pp. 168-182.

[6] Little, J.D.C. 1961. "A proof of the queueing formula $L = \lambda W$," *Operations Research*, Vol. 9, pp. 383-387.

[7] Nelson, R. and B.R. Iyer. 1985. "Analysis of a replicated data base," *Performance Evaluation*, Vol. 5, pp. 133-148.

[8] Nelson, R. and A.N. Tantawi. 1985. "Approximate analysis of fork/join synchronization in parallel queues," IBM Report RC11481.

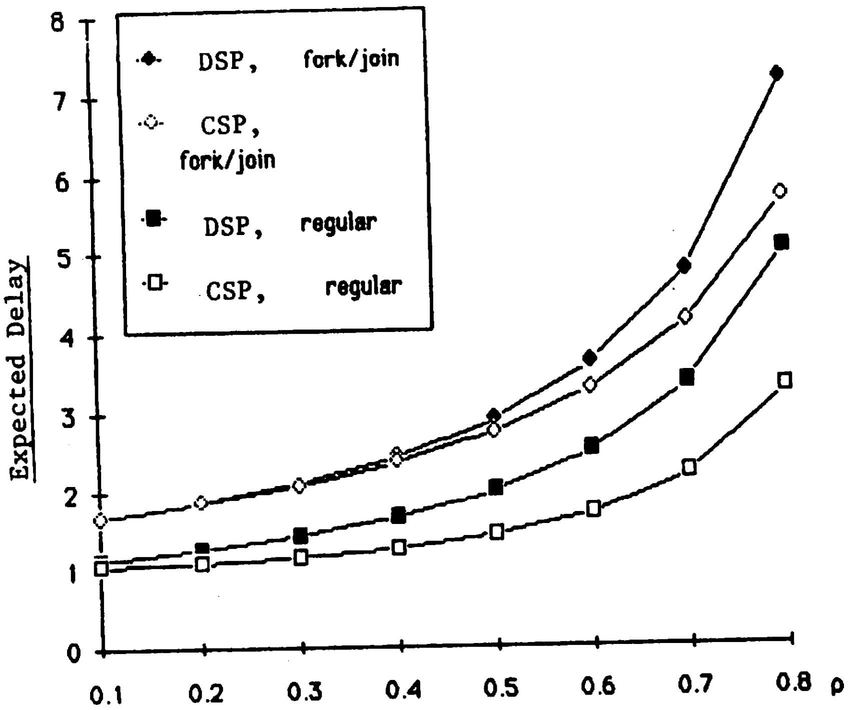

Figure 2: Comparison of CSP and DSP, $q_r = 1/4$.

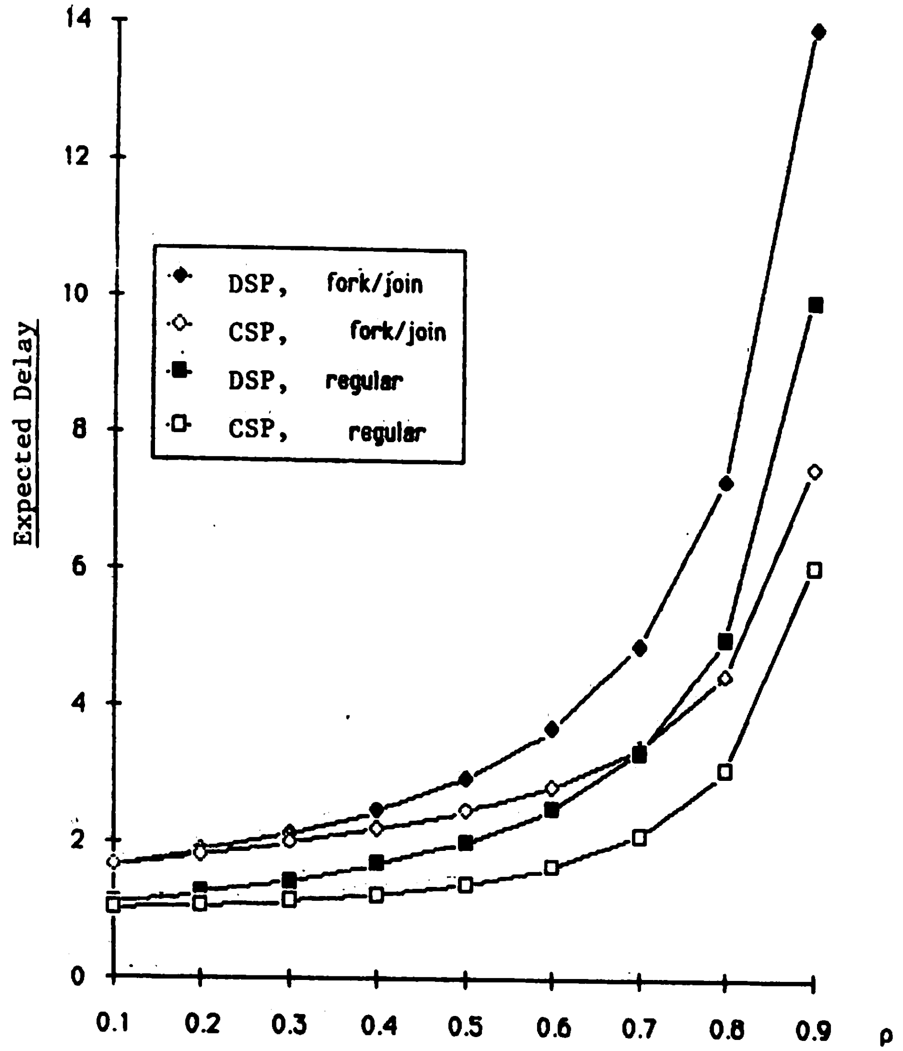

Figure 3: Comparison of CSP and DSP, $q_r = 3/4$.

[9] Neuts, M.F. 1981. *Matrix-Geometric solutions in stochastic models - an algorithmic approach*, John Hopkins University Press.

[10] Rao, B.M. and M.J.M. Posner. 1985. "Algorithmic and Approximation Analyses of the Split and Match Queue," *Stochastic Models*, Vol. 1, pp. 433-456.

[11] Stoyan D. 1983. *Comparison methods for queues and other stochastic models*, John Wiley & Sons, Chichester England.