Probability in the Engineering and Informational Sciences, **24**, 2010, 473–483.
doi:10.1017/S0269964810000112

ON THE PROBABILITY DISTRIBUTION OF JOIN QUEUE LENGTH IN A FORK-JOIN MODEL

JUN LI
Communications Research Centre (CRC) Canada
Ottawa, ON, Canada K2H 8S2
E-mail: jun.li@crc.gc.ca

YIQIANG Q. ZHAO
School of Mathematics and Statistics
Carleton University, Ottawa, ON, Canada K1S 5B6
E-mail: zhao@math.carleton.ca

In this article, we consider the two-node fork-join model with a Poisson arrival process and exponential service times of heterogeneous service rates. Using a mapping from the queue lengths in the parallel nodes to the join queue length, we first derive the probability distribution function of the join queue length in terms of joint probabilities in the parallel nodes and then study the exact tail asymptotics of the join queue length distribution. Although the asymptotics of the joint distribution of the queue lengths in the parallel nodes have three types of characterizations, our results show that the asymptotics of the join queue length distribution are characterized by two scenarios: (1) an exact geometric decay and (2) a geometric decay with the prefactor $n^{-1/2}$.

1. INTRODUCTION

In this article, we consider the $M/M/1$ fork-join queueing system, referred to as the Flatto–Hahn model, which was first studied by Flatto and Hahn [7] and is depicted in Figure 1. In this model, jobs arrive at the system according to a Poisson process with rate λ. Each job arriving at the system creates two tasks simultaneously, which are assigned to two separate nodes, queue 1 and queue 2, respectively, for processing. Each node operates like an $M/M/1$ queuing system using the first-come–first-served (FCFS) discipline. The service rate in queue i is μ_i $(i = 1, 2)$. When a task is completed by the server, it is put in the join queue if the other task of the same job is still in

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

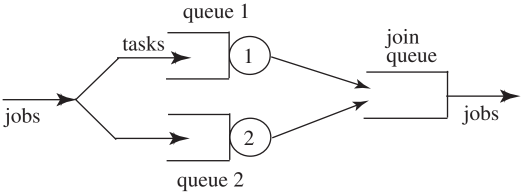

FIGURE 1. A two-node fork-join network.

service; otherwise, the task immediately departs from the system, together with the other belonging to the same job of the task being removed from the join queue.

Fork-join queuing systems have received much research attention in the literature, as they model parallel processing in computer and manufacturing engineering (see, e.g., Chao and Zheng [4], Chen [5], Heidelberger and Trivedi [8], Nelson and Tantawi [10], Nelson and Towsley [11], and Song, Xu, and Liu [13]). In computer and communication networking, a fork-join model represents the processing of computer programs, data packets, database systems, and so forth that involve parallel multitasking through splitting/joining information. In manufacturing, a fork-join network, referred to as an assembly network, represents the assembly of a product or system requiring two or more parts to be processed simultaneously at separate locations. A supply-chain fork-join network typically represents filling an order and obtaining items or products simultaneously from vendors.

Studies on fork-join queuing systems have been conducted for three decades. For a general m-node fork-join model, for which an intractable m-dimensional Markov chain is involved, only upper/lower bounds and approximation results of performance metrics, such as the job response time and the queue length, were derived (see, e.g., Ayhan and Kim [1], Baccelli, Makowski, and Shwartz [3], Chen [5], Ko and Serfozo [9], Nelson and Tantawi [10], Nelson and Towsley [11], Varki [16], Varma and Makowski [17]). (When the arrival process is deterministic in an m-node fork-join model, a product form solution was obtained by Pinotsi and Zazanis [12].) On the contrary, for a two-node fork-join model, exact analysis has been carried out, which was initiated by Flatto [6] and Hahn [7]. They derived the generating function of both queue lengths and studied the asymptotics in the distribution for an $M/M/1$ fork-join model having two nodes with heterogeneous service rates. The asymptotics of the joint distribution of the queue lengths of the model were also studied by Shwartz and Weiss [14] using a completely different approach (large deviations and time reversibility). The probability distribution of the join queue length for the model with the homogeneous service rate was obtained by Nelson and Tantawi [10], who also obtained an exact expression of the expected response time. Explicit or approximate expressions of the joint queue length distribution for the model were derived by Tan and Knessl [15] under a heavy traffic condition. A discrete-time version of the model was studied by Zhang [19], who computed the generating function of the queue lengths and the Laplace transform of the joint waiting times. By adding a situation in which jobs

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

cannot be split into two tasks to the Flatto–Hahn model, Wright [18] generalized the results derived by Flatto and Hahn [7]. An extension to the $M/G/1$ fork-join queuing system was made by Baccelli [2].

Most of these studies were concerned with either performance at the system level, such as the response time and the number of jobs in the system, or the queue lengths of the parallel queues. However, the join queue size is another key performance metric for the management of inventories in supply chain fork-join networks and of production capabilities in assembly fork-join networks. To the best of our knowledge, the only result of the join queue length distribution reported is for the Flatto–Hahn model with a homogeneous service rate, but no studies on the join queue length distribution for a heterogeneous system have been carried out and no asymptotic results have been reported in the literature. In this article, we generalize the result obtained by Nelson and Tantawi [10] through an investigation of the stationary join queue length distribution for the Flatto–Hahn model with heterogeneous service rates. Moreover, we derive the exact tail asymptotics of the join queue length distribution under two different scenarios of the service rates.

The rest of this article is organized as follows: Section 2 states the main results and provides relevant results in the literature; Section 3 derives explicit expressions of two probability distribution functions based on which the join queue length distribution is obtained; and asymptotic analysis of the join queue length distribution is carried out in Section 4.

2. MAIN RESULTS

In this section we first recall some relevant results of the Flatto-Hahn model, which are necessary for stating the main results of this article. The main results are stated in Theorem 2.2 and Theorem 2.4, respectively. The former, which will be proved in Section 3, provides an expression for the probability distribution function of the join queue length and the later, which will be proved in Section 4, provides a complete characterization of the exact tail asymptotics of the join queue length distribution.

Without loss of generality, we assume $\mu_1 \leq \mu_2$ and $\lambda + \mu_1 + \mu_2 = 1$. The fork-join network is stable if and only if $\lambda < \min\{\mu_1, \mu_2\} = \mu_1$. Under the stability condition, we denote by $P_{i,j}$ the probability that there are i tasks in queue 1 and j tasks in queue 2 and by q_k we denote the probability that there are k tasks in the join queue, for $i, j, k = 0, 1, \ldots$. We define the joint probability generating function of $P_{i,j}$ by $P(z,w) = \sum_{i,j} z^i w^j p_{ij}$. In [7], Flatto and Hahn determined $P(z,w)$ through explicitly deriving the following two marginal probability generating functions:

$$P(z,0) \equiv \sum_{i=0}^{\infty} z^i P_{i,0} = \frac{\mu_2 - \lambda}{\mu_2} \frac{\phi(z)}{\phi(1)}, \qquad |z| \leq 1, \tag{2.1}$$

$$P(0,w) \equiv \sum_{j=0}^{\infty} w^j P_{0,j} = \frac{\mu_1 - \lambda}{\mu_1} \frac{\psi(w)}{\psi(1)}, \qquad |w| \leq 1. \tag{2.2}$$

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

In Eq. (2.1), $\phi(z)$ is given by

$$\phi(z) = \frac{\sqrt{a_3 - z} + \sqrt{a_3 - 1}}{\left(\sqrt{a_3 - z} + \sqrt{a_3 - \mu_1/\mu_2}\right)\left(\sqrt{a_3 - z} - \sqrt{a_3 - \mu_1/\lambda}\right)}, \tag{2.3}$$

where a_3 is the largest zero of $D_1(z) = (z - \mu_1)^2 - 4\lambda\mu_2 z^3$ and $a_3 > \mu_1/\lambda$, as proved by Flatto and Hahn [7]. In Eq. (2.2), $\psi(w)$ is given by

$$\psi(w) = \frac{\sqrt{\hat{a}_3 - w} + \sqrt{\hat{a}_3 - 1}}{\left(\sqrt{\hat{a}_3 - w} + \sqrt{\hat{a}_3 - \mu_2/\mu_1}\right)\left(\sqrt{\hat{a}_3 - w} + \sqrt{\hat{a}_3 - \mu_2/\lambda}\right)}, \tag{2.4}$$

where $\hat{a}_3 = (\mu_2/\mu_1)a_3$. We note that $\psi(w)$ is well defined for $|w| \le \mu_2/\mu_1$. In particular, if $\mu_1 = \mu_2 = \mu$, the two generating functions are equal and

$$P(z,0) = P(0,z) = \frac{(1-\rho)^{3/2}}{\sqrt{1-\rho z}}, \tag{2.5}$$

where $\rho = \lambda/\mu$.

Definition 2.1: *For sequences $\{x_n, n \in \mathbb{Z}_+\}$ and $\{y_n, n \in \mathbb{Z}_+\}$ of real numbers, where $\mathbb{Z}_+$ is the set of all nonnegative integers, x_n is asymptotic to y_n, denoted by $x_n \sim y_n$ as n approaches infinity, if* $\lim_{n\to\infty}(x_n/y_n) = 1$.

With the specified assumption $\mu_1 < \mu_2$, the asymptotics of $P_{i,0}$ and $P_{0,j}$ have also been derived by Flatto and Hahn [7] and are given by

$$P_{i,0} \sim \frac{(\mu_2 - \mu_1)\left(\sqrt{\lambda(\lambda a_3 - \mu_1)} - \lambda\right)}{\mu_1\mu_2}\rho_1^i, \qquad \text{as } i \longrightarrow \infty \tag{2.6}$$

and

$$\begin{aligned} P_{0,j} \sim {} & \frac{(\mu_1 - \lambda)\sqrt{a_3}\,(c_1 + c_3)(c_2 + c_3)}{4\sqrt{\pi}\,\mu_1 c_1 c_2}\left(\frac{1}{c_1} + \frac{1}{c_2} - \frac{1}{c_3}\right) j^{-3/2} \\ & \times \left(\frac{\mu_1}{\mu_2 a_3}\right)^j \quad \text{as } j \longrightarrow \infty, \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} & \rho_1 = \frac{\lambda}{\mu_1}; \qquad \rho_2 = \frac{\lambda}{\mu_2}; \qquad c_1 = \sqrt{a_3 - \mu_1/\lambda}; \qquad c_2 = \sqrt{a_3 - 1}; \\ & c_3 = \sqrt{a_3 - \mu_1/\mu_2}. \end{aligned} \tag{2.8}$$

In the following, we denote by q_k^+ (respectively q_k^-) the limiting probability that the number of tasks in queue 1 (respectively queue 2) exceeds the number of tasks in queue 2 (respectively queue 1) by k.

THEOREM 2.2: *For the Flatto–Hahn model with heterogeneous service rates satisfying $\lambda + \mu_1 + \mu_2 = 1$ and $\lambda < \min\{\mu_1, \mu_2\} = \mu_1$, the stationary distribution function q_n of the join queue length is given by*

$$q_0 = P(\mu_1/\mu_2, 0), \tag{2.9}$$

and for $n = 1, 2, \ldots,$

$$q_n = q_n^+ + q_n^-, \tag{2.10}$$

where

$$q_n^+ = \sum_{i=n}^{\infty} \left(\frac{\mu_1}{\mu_2}\right)^{i-n} P_{i,0}, \tag{2.11}$$

$$q_n^- = \sum_{j=n}^{\infty} \left(\frac{\mu_2}{\mu_1}\right)^{j-n} P_{0,j}. \tag{2.12}$$

Remark 2.3: Indeed, it is also true that $q_0 = P(0, \mu_2/\mu_1)$, from which we have $P(\mu_1/\mu_2, 0) = P(0, \mu_2/\mu_1)$. This equality can be verified through the expressions given in Eq (2.1) and Eq (2.2).

THEOREM 2.4: *For the Flatto–Hahn model with heterogeneous service rates satisfying $\lambda + \mu_1 + \mu_2 = 1$ and $\lambda < \min\{\mu_1, \mu_2\}$, the exact tail asymptotics of the join queue length distribution are given as follows:*

1. *If $\mu_1 = \mu_2 = \mu$,*

$$q_n \sim \frac{2\sqrt{1-\rho}}{\sqrt{\pi}} n^{-\frac{1}{2}} \rho^n \quad \text{as } n \longrightarrow \infty. \tag{2.13}$$

2. *If $\mu_1 < \mu_2$,*

$$q_n \sim \frac{(\mu_2 - \mu_1)\left(\sqrt{\lambda(\lambda a_3 - \mu_1)} - \lambda\right)}{\mu_1\mu_2(1-\rho_2)} \rho_1^n \quad \text{as } n \longrightarrow \infty. \tag{2.14}$$

Remark 2.5: We note that the asymptotics of the join queue length distribution is characterized by either a geometric decay with the prefactor $n^{-1/2}$ (when the service rates are the same) or an exact geometric decay (when the service rates are unequal).

When the service rates of the parallel queues are equal, there is an equal probability that the number of tasks in one queue exceeds the other by n. Intuitively, one might think that the asymptotic behavior of q_n has the same trend as that of $P_{n,0}$, which is confirmed in Theorem 2.2. However, when the service rate of queue 1 is less than that of queue 2, more tasks in queue 1 than queue 2 are waiting for required services; that is, in the equilibrium regime, the probability that the number of tasks in queue 1 exceeds queue 2 by n is larger than the probability that the number of tasks in queue 2 exceeds queue 1 by n, for a fixed n. Therefore, the asymptotic property of q_n is dominated by that of q_n^+, which has an exact geometric decay.

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

3. STATIONARY DISTRIBUTION

It is clear that

$$q_k^+ = \sum_{i=0}^{\infty} P_{i+k,i} \quad \text{and} \quad q_k^- = \sum_{i=0}^{\infty} P_{i,i+k}, \qquad k = 0, 1, 2, \ldots, \tag{3.1}$$

and

$$q_k = q_k^+ + q_k^-, \qquad k = 1, 2, \ldots. \tag{3.2}$$

We define

$$q_0 = q_0^+ = q_0^-. \tag{3.3}$$

LEMMA 3.1: *For $k \geq 1$,*

$$(\mu_1 + \mu_2) q_0 = \mu_1 q_1^+ + \mu_2 q_1^- + (\mu_1 + \mu_2) P_{0,0}, \tag{3.4}$$

$$(\mu_1 + \mu_2) q_k^+ = \mu_1 q_{k+1}^+ + \mu_2 q_{k-1}^+ - \mu_2 (P_{k-1,0} - P_{k,0}), \tag{3.5}$$

$$(\mu_1 + \mu_2) q_k^- = \mu_1 q_{k-1}^- + \mu_2 q_{k+1}^- - \mu_1 (P_{0,k-1} - P_{0,k}). \tag{3.6}$$

PROOF: The balance equations for $P_{i,j}$ are as follows:

State	Equation	
$(0,0)$	$\lambda P_{0,0} = \mu_1 P_{1,0} + \mu_2 P_{0,1},$	**(3.7)**
$(i,0); i \geq 1$	$(\lambda + \mu_1) P_{i,0} = \mu_1 P_{i+1,0} + \mu_2 P_{i,1},$	**(3.8)**
$(0,j); j \geq 1$	$(\lambda + \mu_2) P_{0,j} = \mu_1 P_{1,j} + \mu_2 P_{0,j+1},$	**(3.9)**
$(i,j); i,j \geq 1$	$(\lambda + \mu_1 + \mu_2) P_{i,j} = \mu_1 P_{i+1,j} + \mu_2 P_{i,j+1} + \lambda P_{i-1,j-1}.$	**(3.10)**

Hence,

$$\begin{aligned}
q_0 &= P_{0,0} + \sum_{i=1}^{\infty} P_{i,i} \\
&= P_{0,0} + \mu_1 \sum_{i=1}^{\infty} P_{i+1,i} + \mu_2 \sum_{i=1}^{\infty} P_{i,i+1} + \lambda \sum_{i=1}^{\infty} P_{i-1,i-1} \\
&= P_{0,0} + \mu_1 \left(\sum_{i=0}^{\infty} P_{i+1,i} - P_{1,0} \right) + \mu_2 \left(\sum_{i=0}^{\infty} P_{i,i+1} - P_{0,1} \right) + \lambda \sum_{i=0}^{\infty} P_{i,i} \\
&= P_{0,0} + \mu_1 \left(q_1^+ - P_{1,0} \right) + \mu_2 \left(q_1^- - P_{0,1} \right) + \lambda q_0 \\
&= \mu_1 q_1^+ + \mu_2 q_1^- - (\mu_1 P_{1,0} + \mu_2 P_{0,1}) + P_{0,0} + \lambda q_0 \\
&= \mu_1 q_1^+ + \mu_2 q_1^- - \lambda P_{0,0} + P_{0,0} + \lambda q_0,
\end{aligned}$$

in which Eq. (3.10) was used for the second equality and Eq. (3.7) was used for the last one. Equation (3.4) is now derived by using the assumption that $\lambda + \mu_1 + \mu_2 = 1$.

For $k \geq 1$,

$$
\begin{aligned}
q_k^+ &= P_{k,0} + \sum_{i=1}^{\infty} P_{i+k,i} \\
&= P_{k,0} + \mu_1 \sum_{i=1}^{\infty} P_{i+k+1,i} + \mu_2 \sum_{i=1}^{\infty} P_{i+k,i+1} + \lambda \sum_{i=1}^{\infty} P_{i+k-1,i-1} \\
&= P_{k,0} + \mu_1 \left(\sum_{i=0}^{\infty} P_{i+k+1,i} - P_{k+1,0} \right) + \mu_2 \left(\sum_{i=0}^{\infty} P_{i+k-1,i} - P_{k,1} - P_{k-1,0} \right) \\
&\quad + \lambda \sum_{i=0}^{\infty} P_{i+k,i} \\
&= \mu_1 q_{k+1}^+ + \mu_2 q_{k-1}^+ - (\mu_1 P_{k+1,0} + \mu_2 P_{k,1}) + P_{k,0} - \mu_2 P_{k-1,0} + \lambda q_k^+ \\
&= \mu_1 q_{k+1}^+ + \mu_2 q_{k-1}^+ - (\lambda + \mu_1) P_{k,0} + P_{k,0} - \mu_2 P_{k-1,0} + \lambda q_k^+ \\
&= \mu_1 q_{k+1}^+ + \mu_2 q_{k-1}^+ + \mu_2 P_{k,0} - \mu_2 P_{k-1,0} + \lambda q_k^+,
\end{aligned}
$$

in which Eq. (3.10) was used for the second equality and Eq. (3.8) was used for the second last equality. Equation (3.5) follows immediately. By symmetry, Eq. (3.6) can be easily obtained. ■

LEMMA 3.2: *For $n = 0, 1, 2, \ldots,$*

$$q_{n+1}^+ = \frac{\mu_2}{\mu_1}(q_n^+ - P_{n,0}), \tag{3.11}$$

$$q_{n+1}^- = \frac{\mu_1}{\mu_2}(q_n^- - P_{0,n}). \tag{3.12}$$

PROOF: By summing up both sides of Eq. (3.5) over k from 1 to n and rearranging the terms, we have

$$\mu_1 q_{n+1}^+ = \mu_2 q_n^+ - \mu_2 P_{n,0} + \mu_2 C, \tag{3.13}$$

where $C = \mu_1 q_1^+ - \mu_2 q_0 + \mu_2 P_{0,0}$ is a constant. By letting n approach infinity, we have that C must be zero, which gives

$$q_1^+ = \frac{\mu_2}{\mu_1}(q_0 - P_{1,0}), \tag{3.14}$$

and

$$\mu_1 q_{n+1}^+ = \mu_2 q_n^+ - \mu_2 P_{n,0}, \qquad n \geq 1. \tag{3.15}$$

We obtain Eq. (3.11) by combining Eqs. (3.14) and (3.15). Using the same argument on Eq. (3.6), we can obtain Eq. (3.12). ■

PROOF OF THEOREM 2.4: The proof follows from Lemma 3.2. ■

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

4. EXACT TAIL ASYMPTOTICS

Under the specified assumption that $\mu_1 < \mu_2$, we first derive the exact tail asymptotics of q_n^+ and q_n^- in the following lemma.

LEMMA 4.1: *If $\mu_1 < \mu_2$, then, as $n \to \infty$,*

$$q_n^+ \sim \frac{(\mu_2 - \mu_1)(\sqrt{\lambda(\lambda a_3 - \mu_1)} - \lambda)}{\mu_1 \mu_2 (1 - \rho_2)} \rho_1^n, \tag{4.1}$$

$$q_n^- \sim \frac{c a_3}{a_3 - 1} \left(\frac{\mu_1}{\mu_2 a_3}\right)^n n^{-3/2}, \tag{4.2}$$

where

$$c = \frac{(\mu_1 - \lambda)\sqrt{a_3}\,(c_1 + c_3)(c_2 + c_3)}{4\sqrt{\pi}\,\mu_1 c_1 c_2} \left(\frac{1}{c_1} + \frac{1}{c_2} - \frac{1}{c_3}\right) \tag{4.3}$$

with c_i given in Eq. (2.8) *and a_3 being the largest zero of $D_1(z) = (z - \mu_1)^2 - 4\lambda\mu_2 z^3$.*

PROOF:

$$\begin{aligned} q_n^+ &= \sum_{i=n}^{\infty} \left(\frac{\mu_1}{\mu_2}\right)^{i-n} P_{i,0} \\ &= \sum_{i=0}^{\infty} \left(\frac{\mu_1}{\mu_2}\right)^{i} P_{i+n,0} \\ &= \left(\frac{\lambda}{\mu_1}\right)^{n} \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu_2}\right)^{i} P_{i+n,0} \left(\frac{\mu_1}{\lambda}\right)^{i+n}. \end{aligned}$$

This implies

$$\begin{aligned} \lim_{n\to\infty} \frac{q_n^+}{\rho_1^n} &= \lim_{n\to\infty} \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu_2}\right)^{i} P_{i+n,0} \left(\frac{\mu_1}{\lambda}\right)^{i+n} \\ &= \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu_2}\right)^{i} \left\{\lim_{n\to\infty} P_{i+n,0} \left(\frac{\mu_1}{\lambda}\right)^{i+n}\right\} \\ &= \frac{(\mu_2 - \mu_1)\left(\sqrt{\lambda(\lambda a_3 - \mu_1)} - \lambda\right)}{\mu_1 \mu_2} \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu_2}\right)^{i} \\ &= \frac{(\mu_2 - \mu_1)(\sqrt{\lambda(\lambda a_3 - \mu_1)} - \lambda)}{\mu_1 \mu_2 (1 - \rho_2)}, \end{aligned}$$

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

where the second equality is due to the dominated convergence theorem and the third equality is due to Eq. (2.6).

$$\begin{aligned}
q_n^- &= \sum_{j=n}^{\infty}\left(\frac{\mu_2}{\mu_1}\right)^{j-n} P_{0,j}\\
&= \sum_{j=0}^{\infty}\left(\frac{\mu_2}{\mu_1}\right)^{j} P_{0,j+n}\\
&= \sum_{j=0}^{\infty}\left(\frac{\mu_2}{\mu_1}\right)^{j} P_{0,j+n}(j+n)^{3/2}\left(\frac{\mu_2 a_3}{\mu_1}\right)^{j+n}(j+n)^{-3/2}\left(\frac{\mu_1}{\mu_2 a_3}\right)^{j+n}\\
&= \left(\frac{\mu_1}{\mu_2 a_3}\right)^{n} n^{-3/2}\sum_{j=0}^{\infty}\left(\frac{1}{a_3}\right)^{j} P_{0,j+n}(j+n)^{3/2}\left(\frac{\mu_2 a_3}{\mu_1}\right)^{j+n}\left(\frac{n}{j+n}\right)^{3/2},
\end{aligned}$$

which implies

$$\begin{aligned}
&\lim_{n\to\infty}\frac{q_n^-}{(\mu_1/\mu_2 a_3)^n n^{-3/2}}\\
&\qquad= \lim_{n\to\infty}\sum_{j=0}^{\infty}\left(\frac{1}{a_3}\right)^{j} P_{0,j+n}(j+n)^{3/2}\left(\frac{\mu_2 a_3}{\mu_1}\right)^{j+n}\left(\frac{n}{j+n}\right)^{3/2}\\
&\qquad= \sum_{j=0}^{\infty}\left(\frac{1}{a_3}\right)^{j}\left\{\lim_{n\to\infty} P_{0,j+n}(j+n)^{3/2}\left(\frac{\mu_2 a_3}{\mu_1}\right)^{j+n}\left(\frac{n}{j+n}\right)^{3/2}\right\}\\
&\qquad= c\sum_{j=0}^{\infty}\left(\frac{1}{a_3}\right)^{j}\\
&\qquad= \frac{ca_3}{a_3-1},
\end{aligned}$$

where the second equality is due to the dominated convergence theorem and the third equality is due to Eq. (2.7). ■

Proof of Theorem 2.4: When $\mu_1 = \mu_2 = \mu$, $q_0 = P(1,0) = 1-\rho$ and $q_n = 2\sum_{i=n}^{\infty} P_{i,0}$ $(n = 1,2,\dots)$ by Theorem 2.2. Meanwhile, $P_{i,0}$ is the coefficient of z^i in the Taylor expansion of $P(z,0)$ given in Eq. (2.5). Therefore, for $m = 1,2,\dots,$

$$\begin{aligned}
P_{m,0} &= (1-\rho)^{3/2}\left\{[z^m](1-\rho z)^{-1/2}\right\}\\
&= (1-\rho)^{3/2} g_m \rho^m,
\end{aligned}$$

where $[z^m]f(z)$ is the coefficient of z^m in the Taylor expansion of $f(z)$ and

$$g_m = \frac{(2m)!}{2^{2m}(m!)(m!)} \sim \frac{1}{\sqrt{\pi m}} \quad \text{as } m \longrightarrow \infty.$$

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

Now, for $n = 1, 2, \ldots,$

$$\begin{aligned} q_n &= 2\sum_{i=n}^{\infty} P_{i,0} \\ &= 2(1-\rho)^{3/2}\sum_{i=n}^{\infty} g_i\rho^i \\ &= 2(1-\rho)^{3/2}\rho^n\sum_{m=0}^{\infty} g_{m+n}\rho^m \\ &= 2(1-\rho)^{3/2}\rho^n n^{-1/2}\sum_{m=0}^{\infty} g_{m+n}n^{1/2}\rho^m. \end{aligned}$$

This implies

$$\begin{aligned} \lim_{n\to\infty}\frac{q_n}{2(1-\rho)^{3/2}\rho^n n^{-1/2}} &= \lim_{n\to\infty}\sum_{m=0}^{\infty} g_{m+n}n^{1/2}\rho^m \\ &= \lim_{n\to\infty}\sum_{m=0}^{\infty} g_{m+n}(m+n)^{1/2}\left(\frac{n}{m+n}\right)^{1/2}\rho^m \\ &= \sum_{m=0}^{\infty}\left\{\lim_{n\to\infty} g_{m+n}(m+n)^{1/2}\left(\frac{n}{m+n}\right)^{1/2}\right\}\rho^m \\ &= \sum_{m=0}^{\infty}\left\{\frac{1}{\sqrt{\pi}}\right\}\rho^m \\ &= \frac{1}{(1-\rho)\sqrt{\pi}}, \end{aligned}$$

where the third equality is due to the dominated convergence theorem. We proved Eq. (2.13) of the theorem.

When $\mu_1 < \mu_2$, the exponential decay rates of q_n^+ and q_n^- are ρ_1 and $\mu_1/\mu_2 a_3$, respectively, by Lemma 4.1. Since $\mu_2 a_3\rho_1 > \mu_2(\mu_1/\lambda)(\lambda/\mu_1) = \mu_2 > \mu_1$,

$$\rho_1 > \frac{\mu_1}{\mu_2 a_3}; \tag{4.4}$$

that is, when $\mu_1 < \mu_2$, the decay rate of q_n^+ is always larger than that of q_n^- (or q_n is always dominated by q_n^+ for sufficiently large n). Therefore, $q_n \sim q_n^+$ as $n \to \infty$, when $\mu_1 < \mu_2$. From Eq. (4.1) in Lemma 4.1, we obtain Eq. (2.14) of the theorem. ■

Acknowledgement

This research was supported in part by an NSERC Discovery Grant.

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press

References

1. Ayhan, H. & Kim, J.K. (2007). A general class of closed fork and join queues with subexponential service times. *Stochastic Models* 23: 523–535.
2. Baccelli, F. (1985). Two parallel queues created by arrivals with two demands: The $M/G/2$ symmetrical case. Technical report 426, INRIA-Rocquencourt.
3. Baccelli, F., Makowski, A.M., & Shwartz, A. (1989). The fork-join queue and related systems with synchronization constraints: stochastic ordering and computable bounds. *Advances in Applied Probability* 21: 629–660.
4. Chao, X. & Zheng, S. (2000). Triggered concurrent batch arrivals and batch departures in queueing networks. *Discrete Event Dynamic Systems* 10: 115–129.
5. Chen, R.J. (2001). A hybrid solution of fork/join synchronization in parallel queues. *IEEE Transactions on Parallel and Distributed Systems* 12: 829–845.
6. Flatto, L. (1985). Two parallel queues created by arrivals with two demands II. *SIAM Journal on Applied Mathematics* 45: 861–878.
7. Flatto, L. & Hahn, S. (1984). Two parallel queues created by arrivals with two demands I. *SIAM Journal on Applied Mathematics* 44: 1041–1053.
8. Heidelberger, P. & Trivedi, K.S. (1983). Queueing network models for parallel processing with asynchronous tasks. *IEEE Transactions on Computers* 32: 73–82.
9. Ko, S.S. & Serfozo R.F. (2004). Response times in $M/M/s$ fork-join networks. *Advances in Applied Probability* 36: 854–871.
10. Nelson, R. & Tantawi, A.N. (1988). Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers* 37: 739–743.
11. Nelson, R. & Towsley, D. (1993). A performance evaluation of several priority policies for parallel processing systems. *Journal of the ACM* 40: 714–740.
12. Pinotsi, D. & Zazanis, M.A. (2005). Synchronized queues with deterministic arrivals. *Operations Research Letters* 33: 560–566.
13. Song, J.S., Xu, S.H., & Liu, B. (1999). Order-fulfillment performance measures in an assemble-to-order system with stochastic leadtimes. *Operations Research* 47: 131–149.
14. Shwartz, A. & Weiss, A. (1993). Induced rare events: analysis via large deviations and time reversal. *Advances in Applied Probability* 25: 667–689.
15. Tan, X. & Knessl, C. (1996). A fork-join queueing model: Diffusion approximation, integral representations and asymptotics. *Queueing Systems* 22: 287–322.
16. Varki, E. (1999). Mean value technique for closed fork-join networks. *ACM SIGMETRICS Performance Evaluation Review* 27: 103–112.
17. Varma, S. & Makowski, A.M. (1994). Interpolation approximation for symmetric fork-join queues. *Performance Evaluation* 20: 245–265.
18. Wright, P.E. (1992). Two parallel queues with coupled inputs. *Advances in Applied Probability* 24: 986–1007.
19. Zhang, Z. (1990). Analytical results for waiting time and system size distributions in two parallel queueing systems. *SIAM Journal on Applied Mathematics* 50: 1176–1193.

https://doi.org/10.1017/S0269964810000112 Published online by Cambridge University Press