Approximations for a Fork/Join Station with Inputs from Finite Populations

Ananth Krishnamurthy
Department of Decision Sciences and Engineering Systems,
Rensselaer Polytechnic Institute,
110 8th Street, Troy, NY 12180, USA

Rajan Suri
Center for Quick Response Manufacturing, University of Wisconsin-Madison
1513 University Avenue, Madison, WI 53706, USA

Mary Vernon
Department of Computer Sciences, University of Wisconsin-Madison
1210 W Dayton Street, Madison, WI-53706, USA

Abstract

Fork/join stations model synchronization constraints in queuing network models of many manufacturing and computer systems. We consider a fork/join station with two input buffers and general inputs from finite populations and derive approximate expressions for throughput and mean queue lengths at the input buffers. We assume that the arrivals to the fork/join stations are renewal, but our approximations only use information about the first two moments of the inter-renewal distributions. Therefore the approximations can be used to predict performance for a variety of systems. We verify the accuracy of these approximations against simulation and report sample results.

Keywords: Parametric decomposition, two-moment approximation, closed queuing networks, fork/join stations.

1. Introduction

Fork/join stations are used to model synchronization constraints between entities in a queuing network. The fork/join station of interest in this paper consists of a server with zero service times and two input buffers. As soon as there is one entity in each buffer, an entity from each of the buffers is removed and joined together. The joined entity exits the fork/join station instantaneously. Subsequent to its departure, the joined entity forks back into the component entities, which then get routed to other parts of the network. Fork/join stations find many applications in queuing models of manufacturing and computer systems. In queuing models of assembly systems, the assembly operation is typically modeled using a fork/join station [2][4][8]. Fork/join stations are also used model the synchronization constraints in models of kanban control strategies [1]. They are also used to model parallel processing, database concurrency control in computer systems analysis [3].

As a starting point for understanding the behavior of queuing networks with fork/join stations, several researchers have analyzed such stations in isolation. For the sake of analytical tractability, a majority of the previous research efforts assume that the fork/join stations have Poisson inputs [2][4][9][10]. Although these results are useful, in many of the applications cited above the input processes are not Poisson. Most studies that assume arrival processes other than Poisson such as those reported in [11], assume infinite populations for each arrival process. However, if the fork/join station is part of a closed queuing network, then once the queue length of an input buffer equals the size of the population that can arrive to the buffer, the arrival process shuts down. The analysis of fork/join stations with general arrival processes from finite populations can become very complex even when the inter-arrival times are independent and have a Coxian distribution [6]. Thus approximations for the performance of the fork/join stations in particular and the network in general can be highly useful.

In this paper, we derive approximate expressions for throughput and mean queue lengths at the input buffers of a fork/join station with general inputs from finite populations. The approximations are based on the assumption that the arrivals to the fork/join stations are renewal, but they only use the first two moments of the inter-renewal distributions and can therefore be used to predict performance for a wide variety of systems. In the literature such approximations are often referred to as two-moment approximations [12]. In addition to providing performance estimates for a fork/join station in isolation, the approximations can also be used as building blocks in parametric decomposition approaches for solving larger queuing networks. The outline of this paper is as follows. Section 2 describes the fork/join station under consideration and an overview of our approach. We derive the general form for the approximations in Section 3 and in Section 4 we derive the detailed form of the approximations and test their accuracy against simulation. Section 5 provides the concluding remarks.

2. System Description and Approach

We describe our model of the fork/join station next. As shown in the Figure 1 the fork/join station has two input buffers, B_1 and B_2. SN_i denotes the rest of the queuing network for entities that arrive to buffer B_i, i=1,2. If an entity arriving in buffer B_1 (B_2) finds buffer B_2 (B_1) empty, it waits for the corresponding entity to arrive in input buffer B_2 (B_1). As soon as there is at least one entity in each queue, one entity is removed from each buffer. The removed entities join together, and immediately depart from the fork/join station. As a result the contents of both input buffers reduce by one. Subsequent to departure from the fork/join station, the joined entity forks back into two entities that are routed back to SN_1 and SN_2 respectively. In SN_1 and SN_2 these entities are subjected to random delays before they revisit the fork/join station. There is a finite population of size K_i for each entity i. Consequently, the number of entities in input buffer B_i and queuing network SN_i always sum up to K_i, i=1,2. Additionally, the arrival process to buffer B_i shuts down when there are K_i units in buffer B_i.

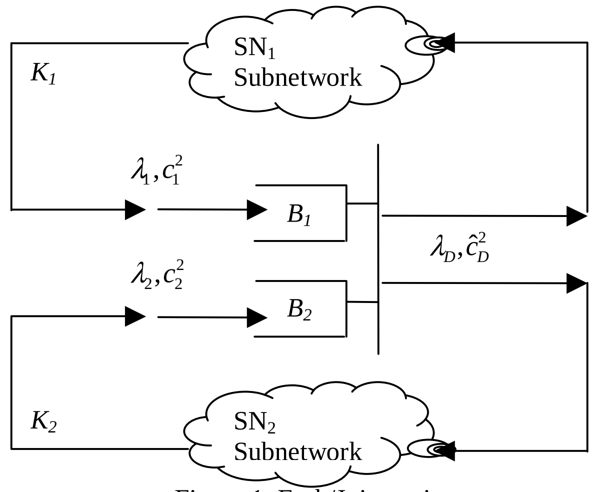

Figure 1. Fork/Join station

Since the sub-networks SN_1 and SN_2 from which entities arrive to input buffers can have different configurations resulting in arbitrary delays, the arrival processes to the fork/join stations can have arbitrary characteristics. However, analysis of fork/join stations for general arrival processes can be quite complicated. To simplify our analysis, and in keeping with other two-moment approximation methods, we will assume that the arrival processes are independent renewal processes and that the inter-arrival times to the input buffers are independent and identically distributed (*i.i.d*) having means $1/\lambda_1$, $1/\lambda_2$, and squared coefficients of variation (SCVs) c_1^2, c_2^2, respectively. Since we assume that the arrival process to buffer B_1 (B_2) shuts down once it has K_1 (K_2) units, the arrival processes are renewal between shut downs. With these assumptions, our model of this fork/join station is completely characterized by the parameter 6-tuple (λ_1, c_1^2, K_1, λ_2, c_2^2, K_2). For a fork/join station characterized thus, we obtain approximations for the throughput, λ_D, and the mean queue length of each buffer, $\overline{L}_i$, i=1,2. In developing the approximations, we assume that the ratio of input rates $\rho = \lambda_1/\lambda_2$ lies in the interval [0.3,3.0]. This is justified for most practical situations, since in a high performance system one would not normally expect the arrivals rates at one input buffer of a synchronization station to be more than three times that of the other. We also assume that both c_1^2 and c_2^2 lie in the interval [0.5,4.0]. These values capture the typical SCVs observed in actual manufacturing systems [5].

To develop these approximations we first study the impact of the mean of the inter-arrival times ($1/\lambda_1, 1/\lambda_2$) and population size (K_1, K_2) on the performance of the fork/join station using the exact expressions reported in [9] and [10]. Although these expressions are exact only for the case of exponentially distributed inter-arrival times, the insights about the impact of arrival rates help us understand behavior for more general arrival processes. However, to study the impact of second moments of the arrival distributions (in particular c_1^2, and c_2^2) on the performance of the fork/join station, a model assuming Poisson inputs is inadequate. In [6] we analyze a fork/join station where the inter-arrival times have a 2-phase Coxian distribution. This permits analysis for input processes with a wide range of means (0,∞) and SCVs [0.5, ∞). Using the results of this analysis we study the impact of both means and SCVs on the performance measures. Using insights from all the above cases, we develop two-moment approximations for a fork/join station characterized by the 6-tuple (λ_1 , c_1^2 , K_1, λ_2 , c_2^2 , K_2).

3. General Form of Approximations

In this section we use the insights from the exact analysis presented in [9], [10], and [6] to develop the general form of the two-moment approximations for the throughput, λ_D , and the mean queue length of each buffer, $\overline{L}_i$, i=1,2. For the sake of clarity, we modify the notation and use a superscript "E" when the performance measures are based on the discussions presented in [9], [10] assuming exponential inputs and a superscript "C" when the station performance measures are based on the discussions presented in [6] assuming Coxian inputs. First we consider the case with exponential inputs. Assuming without loss of generality that $\rho < 1$, Takahashi *et al.* [10] derive the following expressions for the throughput, λ_D^E , and the mean queue lengths, $\overline{L}_i^E$, i=1,2 at the fork/join station:

$$\lambda_D^E = \lambda_1 \left[\frac{1-\rho^{K_1+K_2}}{1-\rho^{K_1+K_2+1}} \right] \tag{1}$$

$$\overline{L}_1^E = \left[\frac{K_1 \rho^{K_1+K_2+1}}{\rho^{K_1+K_2+1}-1} \right] - \left[\frac{\rho^{K_2+1}}{\rho-1} \right] \left[\frac{\rho^{K_1}-1}{\rho^{K_1+K_2+1}-1} \right] \tag{2}$$

$$\overline{L}_2^E = \left[\frac{\rho}{\rho-1} \right] \left[\frac{\rho^{K_2}-1}{\rho^{K_1+K_2+1}-1} \right] - \left[\frac{K_2}{\rho^{K_1+K_2+1}-1} \right] \tag{3}$$

Note that the above expressions are for the case where $\rho \neq 1$. The corresponding expressions for $\rho = 1$ are obtained by taking the limits as $\rho \to 1$. Using these expressions we study the impact of the mean rates of the input processes on the performance of the fork/join station.

To study the impact of higher moments of the arrival distributions on the performance of the fork/join station, we consider the case where the inter-arrival times to the input buffers have 2-Phase Coxian distributions with mean, $1/\lambda_i$ and SCV, c_i^2 , for i=1,2, respectively. Using the additional constraint of balanced means, we derive unique 2-phase Coxian distributions to characterize the inter-arrival times at each input buffer. Then based on the exact analysis presented in [6] we study the impact of the SCVs of the input processes on station performance measures. Since the closed form expressions for mean queue length and throughput are substantially more complicated than those obtained for exponential inputs we obtain the required insights from the numerical results obtained from the exact computations. To obtain these insights we compute the throughput λ_D^C, and mean queue length $\overline{L}_i^C$, i=1,2 for over 900 input parameter settings with ρ in the interval [0.3,3.0], c_i^2 in the interval [0.5,4.0], and K_i, in the interval [2,20]. These numerical results provide the following insights: (i) For any value of SCV, the upper bound of the throughput, λ_D^C from the fork/join station is min (λ_1, λ_2). This throughput is achieved as K_1 or K_2 tends to infinity. (ii) For given values of c_1^2, c_2^2, and K_1+K_2, the value of λ_D^C is insensitive to the choice of K_1, and K_2. (iii) Substantial queues are observed at the buffers of the input processes with higher rates of arrivals, i.e., $\overline{L}_1^C >> \overline{L}_2^C$ when $\lambda_1 > \lambda_2$. (iv) When input rates are equal, i.e. $\rho = 1$, λ_D^C is quite sensitive to c_1^2 and c_2^2 . (v) When $\rho > 1$ or $\rho < 1$, the station performance measures become increasingly less sensitive to the SCVs and are more primarily dependent on ρ . (vi) Let c^2 denote $0.5(c_1^2 + c_2^2)$. Then, we observe that for a given ρ and c^2, $\lambda_D^C > \lambda_D^E$ if

c^2 <1, and $\lambda_D^C < \lambda_D^E$ if c^2 >1. Additionally, when (ρ -1) (c^2 -1)<0, then $\overline{L}_1^C > \overline{L}_1^E$ and $\overline{L}_2^C < \overline{L}_2^E$ while (ρ -1)(c^2 -1)>0, then $\overline{L}_1^C < \overline{L}_1^E$ and $\overline{L}_2^C > \overline{L}_2^E$.

To obtain quantitative insights into the impact of the SCVs of the inter-arrival times, we study the relative differences of these measures for the Coxian and exponential cases. Thus, we define:

$$d_{\lambda_D} = \frac{\lambda_D^C - \lambda_D^E}{\lambda_D^E} \text{ and } d_{\overline{L}_i} = \frac{\overline{L}_i^C - \overline{L}_i^E}{\overline{L}_i^E}, \ i=1,2 \qquad (4)$$

We compute d_{λ_D} , and $d_{\overline{L}_i}$, i=1,2 and plot their values against c^2 -1 for the parameter settings considered in the numerical study for Coxian inputs. The details are listed in [7]. From these graphs we infer that d_{λ_D} , and $d_{\overline{L}_i}$, i=1,2, vary roughly linearly with c^2 -1, with additional (variation depending on ρ , K_1 or K_2).

Based on these observations we propose the following functions as candidates for the approximations:

$$\lambda_D = \lambda_D^E \left(1 + \left(c^2 - 1\right) w_{\lambda_D}\right) \text{ and } \overline{L}_i = \overline{L}_i^E \left(1 + \left(c^2 - 1\right) w_{\overline{L}_i}\right) \ i=1,2 \qquad (5)$$

Note that when $c_1^2 = c_2^2$ =1, the approximation functions above yield $\lambda_D = \lambda_D^E$, $\overline{L}_i = \overline{L}_i^E$, i=1,2. This implies that the approximations are exact for exponential arrivals. Next, based on the insights from the study for Coxian inputs we identify the general form of w_{λ_D} and $w_{\overline{L}_i}$, i=1,2.

In particular, w_{λ_D} needs to satisfy the following properties:

1. $w_{\lambda_D}\left(\lambda_1, K_1, \lambda_2, K_2\right)$ must be a single valued function of $\lambda_1, K_1, \lambda_2,$ and K_2 .
2. We require that $w_{\lambda_D}\left(\lambda_1, K_1, \lambda_2, K_2\right) = w_{\lambda_D}\left(\lambda_2, K_2, \lambda_1, K_1\right)$ due to symmetry.
3. $w_{\lambda_D}\left(\lambda_1, K_1, \lambda_2, K_2\right) < 0$ This is because $\lambda_D^C > \lambda_D^E$ if c^2 <1, and $\lambda_D^C < \lambda_D^E$ if c^2 >1.
4. $w_{\lambda_D}\left(\lambda_1, K_1, \lambda_2, K_2\right) \to 0$ for $\rho \to \infty$ and $\rho \to 0$ since $\lambda_D \to \lambda_D^E$ when $\rho \to \infty$ or $\rho \to 0$.
5. We require that $w_{\lambda_D}\left(\lambda_1, K_1, \lambda_2, K_2\right) = w_{\lambda_D}\left(\lambda_1, K_2, \lambda_2, K_1\right)$.
6. Since $\lambda_D \leq$ min (λ_1, λ_2) we require $\lambda_D^E \left(1 + \left(c^2 - 1\right) w_{\lambda_D}\right) \leq$ min (λ_1, λ_2).

Candidate functions $w_{\lambda_D}\left(\lambda_1, K_1, \lambda_2, K_2\right)$ that satisfies properties 1 through 5 above are:

$$w_{\lambda_D}\left(\lambda_1, K_1, \lambda_2, K_2\right) = -a_0 \, (1-\rho) \left(\frac{\rho^{K_1+K_2}}{\rho^{2(K_1+K_2)+1} - 1} \right) \qquad (6)$$

where a_0 is a positive constant or a positive function of K_1+K_2. Further if $0 \leq a_0 \leq 2$, then property 6 is also satisfied. (For proof see [7]).

Next we note that $w_{\overline{L}_i}$ must satisfy the following properties:

1. $w_{\overline{L}_i}$, i=1,2 must be a single valued function of $\lambda_1, K_1, \lambda_2,$ and K_2
2. $w_{\overline{L}_i} \to 0$ when $\rho \to \infty$ and $\rho \to 0$, since $\overline{L}_i^C \to \overline{L}_i^E$ when $\rho \to \infty$ and $\rho \to 0$, for i=1,2.
3. By symmetry $w_{\overline{L}_1}\left(\lambda_1, K_1, \lambda_2, K_2\right) = w_{\overline{L}_2}\left(\lambda_2, K_2, \lambda_1, K_1\right)$.

4. $w_{\overline{L}_1}(\lambda_1, K_1, \lambda_2, K_2) \times w_{\overline{L}_2}(\lambda_2, K_2, \lambda_1, K_1) \le 0$, since for given values of $\lambda_1, K_1, \lambda_2,\ K_2$ and c^2 -1, $w_{\overline{L}_1}(\lambda_1, K_1, \lambda_2, K_2)$ and $w_{\overline{L}_2}(\lambda_2, K_2, \lambda_1, K_1) \le 0$ have to be of opposite sign.
5. For $\rho \le 1,\ w_{\overline{L}_1}(\lambda_1, K_1, \lambda_2, K_2) \ge 0$ and $w_{\overline{L}_2}(\lambda_1, K_1, \lambda_2, K_2) \le 0$ while for $\rho \ge 1$, $w_{\overline{L}_1}(\lambda_1, K_1, \lambda_2, K_2) \le 0$ and $w_{\overline{L}_2}(\lambda_1, K_1, \lambda_2, K_2) \ge 0$. This is because when (ρ -1) (c^2 -1)<0, then $\overline{L}_1^C > \overline{L}_1^E$ and $\overline{L}_2^C < \overline{L}_2^E$ while (ρ -1)(c^2 -1)>0, then $\overline{L}_1^C < \overline{L}_1^E$ and $\overline{L}_2^C > \overline{L}_2^E$
6. Since $\overline{L}_i \le K_i$, we require $\overline{L}_i^E\ \left(1 + (c^2 - 1)w_{\overline{L}_i}\right) \le K_i$ i=1,2.

Candidate functions $w_{\overline{L}_i}$, i=1,2 that satisfy properties 1 through 5 above are:

$$w_{\overline{L}_1} = \left(\frac{1-\rho^{b_1}}{1+\rho^{b_1}}\right)\left(\frac{\rho^{b_2}}{1+\rho^{2b_2}}\right) \text{ and } w_{\overline{L}_2} = \left(\frac{1-\rho^{d_1}}{1+\rho^{d_1}}\right)\left(\frac{\rho^{d_2}}{1+\rho^{2d_2}}\right) \qquad (7)$$

where b_1, b_2 and d_1, d_2 are constants or positive functions of K_1+K_2. The value of these constants that ensures $w_{\overline{L}_i}$ satisfies property 6 in addition is determined using results from the simulation experiments described below.

4. Detailed Form of Approximations and their Accuracy

In this section we determine the detailed form of the approximations and test their accuracy. To determine the detailed form of the approximations we only need to determine the best values of the constants a_0 in equation (6), b_1, b_2 and d_1, d_2 in equations (7). We use simulations to determine the best values of these constants. In the simulation experiments, we evaluate the approximations for inter-arrival times that have 2-stage Erlang, Shifted exponential, Lognormal and Hyper-exponential distributions. In these experiments, (λ_1, λ_2) take values of (1,1), (0.3, 1) and (0.8, 1.25) respectively, while K_i, i=1,2 take several values in the range [2,10] and c_i^2, i=1,2 take values in the range [0.5, 4.0]. From these experiments we observe that setting $a_0 = 0.5$ in our approximation for λ_D and setting $d_1 = b_1 = 1$, and $d_2 = b_2 = 4$, in our approximations for $\overline{L}_i$, i=1,2 gives the best performance. For additional details see [7]. With this choice, the final expressions for the two-moment approximations for throughput, λ_D, and the mean queue length of each buffer, $\overline{L}_i$, i=1,2 are as follows:

$$\lambda_D = \lambda_1\left[\frac{1-\rho^{K_1+K_2}}{1-\rho^{K_1+K_2+1}}\right]\left[1-\left(\frac{c^2-1}{2}\right)\left(\frac{(1-\rho)\rho^{K_1+K_2}}{\rho^{2(K_1+K_2)+1}-1}\right)\right] \qquad (8)$$

$$\overline{L}_1 = \left[\left(\frac{K_1\rho^{K_1+K_2+1}}{\rho^{K_1+K_2+1}-1}\right)-\left(\frac{\rho^{K_2+1}}{\rho-1}\right)\left(\frac{\rho^{K_1}-1}{\rho^{K_1+K_2+1}-1}\right)\right]\left[1+\left(c^2-1\right)\left(\frac{1-\rho}{1+\rho}\right)\left(\frac{\rho^4}{1+\rho^8}\right)\right] \qquad (9)$$

$$\overline{L}_2 = \left[\left(\frac{\rho}{\rho-1}\right)\left(\frac{\rho^{K_2}-1}{\rho^{K_1+K_2+1}-1}\right)-\left(\frac{K_2}{\rho^{K_1+K_2+1}-1}\right)\right]\left[1-\left(c^2-1\right)\left(\frac{1-\rho}{1+\rho}\right)\left(\frac{\rho^4}{1+\rho^8}\right)\right] \qquad (10)$$

Equations 8 to 10 are for the case where $\rho \ne 1$. The corresponding equations for $\rho = 1$ are obtained by taking the limits as $\rho \to 1$. Finally, we test the performance of the approximations by computing the percentage difference in the estimates given by the approximations and estimates from simulation. Sample results are provided in Table 1. In this table, λ_D^{SL} and $\overline{L}_i^{SL}$ i=1,2 correspond to performance measures from simulation experiments, while the performance measures determined by the approximations are given by λ_D, and $\overline{L}_i$, i=1,2 respectively. In these experiments inter-arrival times with SCV of 0.5 were assumed to have a shifted exponential distribution while inter-arrival times with SCV of 1.0 and 4.0 were assumed to have a lognormal distribution.

Table 1. Percentage difference between approximations and simulation
(Sample results from over 240 test cases)

λ_1	λ_2	$\varepsilon\left(\lambda_D^{SL}\right)=\frac{\lvert \lambda_D-\lambda_D^{SL}\rvert}{\lambda_D^{SL}}\times 100$		$\varepsilon\left(\overline{L}_1^{SL}\right)=\frac{\lvert \overline{L}_1-\overline{L}_1^{SL}\rvert}{K_1}\times 100$		$\varepsilon\left(\overline{L}_2^{SL}\right)=\frac{\lvert \overline{L}_2-\overline{L}_2^{SL}\rvert}{K_2}\times 100$	
		Maximum	*Average*	*Maximum*	*Average*	*Maximum*	*Average*
1.0	1.0	5.7	2.3	8.8	2.8	8.9	2.8
0.8	1.3	6.0	1.4	10.7	3.4	8.9	3.2
0.3	1.0	3.5	0.3	4.0	0.5	3.1	1.1

5. Conclusions and Extensions

In this paper, we have proposed approximations for the throughput and mean queue lengths at the input buffers of a fork/join station with general arrivals from a finite population. From the sample results reported we observe that the maximum difference in the estimates provided by the approximations, compared with simulation was 6% for station throughput and 11% for mean queue lengths. Although these approximations have been developed for a fork/join station in isolation, a principal application is in developing parametric methods for analysis of larger closed queuing networks with fork/join stations. In such applications, we also need two-moment approximations for the variability parameter of the departure process, $\hat{c}_D^2$. Studies such as [13] reports several issues that need to be addressed when determining the variability parameter in the context of simple queues. We are currently working on addressing these issues in the context of fork/join stations and deriving the necessary two-moment approximations. Using these approximations as building blocks we intend to develop new methods to analyze closed queuing network models of single and multi-stage kanban systems.

References

1. Di Mascolo, M., Frein, Y. and Dallery, Y., 1996, "An Analytical Method for Performance Evaluation of Kanban Controlled Production Systems," Operations Research, 44(1), 50-64.
2. Harisson, J.M., 1973, "Assembly-like Queues," Journal of Applied Probability, 10(2), 354-367.
3. Heidelberger, P. and Trivedi, K. S., 1983, "Queuing Network Models for Parallel Processing with Asynchronous Tasks," IEEE Transactions on Computers, C-31 (11), 1099-1109.
4. Hopp, W. and Simon, J.T., 1989, "Bounds And Heuristics For Assembly-Like Queues," Queuing Systems, 4, 137-156.
5. Kamath, M., Suri, R. and Sanders, J. L., 1988, "Analytical Performance Models for Closed Loop Flexible Assembly Systems," The International Journal of Flexible Manufacturing Systems, 1, 51-84.
6. Krishnamurthy, A., Suri, R. and Vernon, M., 2002, "Analysis of a Fork/Join Synchronization Station with Coxian Inputs From a Finite Population," submitted to Annals of Operations Research.
7. Krishnamurthy, A., Suri, R. and Vernon, M., 2002, "Two-Moment Approximations for a Fork/Join Station with General Inputs," Technical Report, Department of Industrial Engineering, UW -Madison.
8. Rao, P.C. and Suri, R., 2000, "Performance Analysis Of An Assembly Station With Input From Multiple Fabrication Lines," Production and Operations Management, 9(3), 283-302.
9. Som, P., Wilhelm, P.E. and Disney, R.L., 1994, "Kitting Process in a Stochastic Assembly System," Queuing Systems, 17, 471-490.
10. Takahashi, M., Osawa, H. and Fujisawa, T., 1996, "A Stochastic Assembly System With Resume Levels," Asia-Pacific Journal of Operations Research, 15, 127-146.
11. Takahashi, M. and Takahashi, Y., 2000, "Synchronization Queue With Two MAP Inputs And Finite Buffers," Proceedings of the Third International Conference on Matrix Analytical Methods in Stochastic Models, Leuven, Belgium.
12. Whitt, W., 1983, "The Queuing Network Analyzer," Bell Systems Technical Journal, 621983, 2779-2815.
13. Whitt, W., 1995, "Variability Functions for Parametric Decomposition Approximations of Queuing Networks," Management Science, 41, 1704-1715.