Chapter 1

TWO-MOMENT APPROXIMATIONS FOR THROUGHPUT AND MEAN QUEUE LENGTH OF A FORK/JOIN STATION WITH GENERAL INPUTS FROM FINITE POPULATIONS

Ananth Krishnamurthy
Department of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180.
krisha@rpi.edu

Rajan Suri
Center for Quick Response Manufacturing
University of Wisconsin-Madison, 1513 University Avenue, Madison, WI 53706.
suri@engr.wisc.edu

Mary Vernon
Department of Computer Sciences
University of Wisconsin-Madison, 1210 W. Dayton Street, Madison, WI 53706.
vernon@cs.wisc.edu

Chapter to appear in *Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains*, J.G. Shanthikumar, D.D. Yao and W.H.M. Zijm (Eds.), Kluwer International Series in Operations Research and Management Science (in press).

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 chapter 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 each 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 station is typically modeled using a fork/join station (Harrison [7], Latouche [17], Hopp and Simon [8], Rao and Suri [23, 24]). Fork/join stations are also used model the synchronization constraints in models of material control strategies for multi-stage manufacturing systems (Buzacott and Shanthikumar [5], Di Mascolo *et al.* [6], Krishnamurthy *et al.* [11]). In computer systems analysis, queuing networks with fork/join stations have been studied in the context of parallel processing, database concurrency control, and communication protocols (Baccelli *et al.* [3], Prabhakar *et al.* [21], Varki [37]).

As a starting point for understanding the behavior of queuing networks with fork/join stations, several researchers have analyzed such fork/join stations in isolation. The typical inputs for such an analysis are the capacity of each input buffer and a description of the arrival process of entities to each input buffer. Performance measures of interest include synchronization delays, queue length distributions at the different input buffers, and in the case of finite customer populations, station throughput.

For the sake of analytical tractability, a majority of the previous research efforts assume that the fork/join stations have Poisson inputs (Harrison [7], Bhat [4], Lipper and Sengupta [19], Hopp and Simon [8], Som *et al.* [25], Takahashi *et al.* [28]). Although these results are useful, in many of the applications cited above the input processes are not Poisson. In fact, often the input processes have variability quite different from that of a Poisson process. Most approaches that analyze more general arrival processes such as those reported in Takahashi *et al.* [29], and Takahashi and Takahashi [30], 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.

For fork/join stations that are a part of a closed queuing network, exact analysis can be computationally prohibitive. The analysis of fork/join stations with general arrival processes from finite populations can become very complex even when the inter-arrival times have a Coxian distribution [12]. Thus approximations for the performance of the fork/join stations in particular and the network in general can be highly useful. In this chapter, 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.

The two-moment approximations developed in this research find immediate applications in analyzing closed queuing network models of many manufacturing and computer systems. For instance, the approximations can be used to analyze closed queuing network models of fabrication/assembly systems, and material control strategies such as kanban systems. Alternatively, the approximations can be used to analyze closed queuing network models of parallel and distributed computing systems. In addition to these immediate applications, these approximations can be used as building blocks in parametric decomposition approaches for solving larger closed queuing networks with multiple synchronization constraints.

The outline of this chapter is as follows. Section 2 provides some background on two-moment approximations and describes the fork/join station under consideration. The approach to developing two-moment approximations consists of two parts. First, insights obtained from the literature on exact analyses for the cases of exponential and Coxian inter-arrival times are discussed in Section 3. Next, based on these insights, a general form for the approximations is proposed in Section 4. In Section 5 the detailed form of the approximation equations is identified and the accuracy of the approximations is tested against simulation. Section 6 discusses a numerical example and Section 7 provides the concluding remarks.

2. Background

2.1. Two-Moment Approximations

Two-moment approximations have been used extensively to obtain performance measures of simple queues. For instance, Marshall [20], Kuehn [15], Whitt [32, 35] and Kimura [16] assume that the arrival and service processes at the queues are renewal processes characterized by their mean and squared coefficient of variation (SCV) and develop two-moment approximations for performance measures such as mean waiting times, mean queue lengths and SCV of the inter-departure times from M/G/1, GI/G/1, and GI/G/s queues.

Apart from their use in analysis of queues in isolation, two-moment approximations form an integral part of the parametric decomposition approach to estimating the steady state performance measures of non-product form queuing networks. See Whitt [32], Suri *et al.* [26, 27] and the references therein for applications of this approach. The main idea behind the parametric decomposition approach is to first characterize the arrival and service processes approximately by renewal processes and then analyze the individual queues in the network separately. The inter-renewal times are characterized by two parameters, one to represent the rate and the other to represent the variability. In most cases the variability parameter is the SCV of the inter-renewal times. Using this information about the traffic processes, the individual queues in the network are analyzed separately using the two-moment approximations for simple queues and equations that link the output process from a node to the input process of the subsequent nodes in the network are developed. These equations are then used to analyze the networks more efficiently.

Although the parametric decomposition method has several attractive features, thus far it has not been applied to analyzing closed queuing networks with fork/join synchronization stations. The primary reason for this is that two-moment approximations characterizing performance measures at a fork/join station with inputs from finite populations have not been developed. Previous approaches to analyzing queuing networks with fork/join synchronization stations have less efficient solution methods such as product form approximation (Di Mascolo *et al.* [6]) and mean value analysis approaches (Varki [37]). The aim of this research is to develop such two-moment approximations for performance measures at a fork/join station so that the parametric decomposition approach can be used to analyze a larger class of closed queuing networks. Apart from their use in analyzing the performance of queuing networks with several synchronization stations, the approximations for the fork/join station in isolation finds direct applications in the performance analysis of several fabrication/assembly systems and parallel/distributed computing systems. We discuss this in detail in the next section.

2.2. System Description

We describe our model of the fork/join station and explain how it could be used to represent the synchronization behavior in particular manufacturing and computer systems. The model, is illustrated in Figure 1.1, and Table 1.1 summarizes the notation used in the model and throughout the remainder of the chapter.

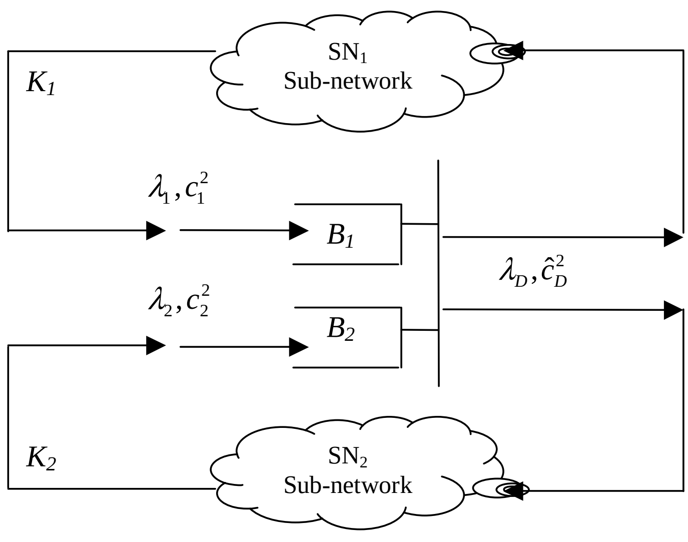


Figure 1.1. Fork/Join Station illustrated as part of two sub-networks

As shown in the figure 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 are reduced 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.

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 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 $(\lambda_1, c_1^2, K_1, \lambda_2, c_2^2, K_2)$. For a fork/join station characterized thus, our

Table 1.1. Notation

Notation	*Description*
λ_i	Rate of arrivals to buffer $B_i, i = 1, 2$
c_i^2	SCV of inter-arrival times at buffer $B_i, i = 1, 2$
K_i	Finite population of entities arriving at buffer $B_i, i = 1, 2$
λ_D	Throughput as computed for general inputs
L_i	Mean queue length at buffer $B_i, i = 1, 2$ for general inputs
c^2	The average of SCVs of inter-arrival times at the two buffers i.e., $c^2 = (1/2)(c_1^2 + c_2^2)$
$\hat{c}_D^2$	Variability parameter of departure process for general inputs
λ_D^x	Throughput as computed for arrival process x where $x = E$(Exponential), C(Coxian)
L_i^x	Mean queue length at buffer $B_i, i = 1, 2$ as computed for arrival process x where $x = E$(Exponential), C(Coxian)
EH	A 2 Stage Erlang/Hyper-exponential combination
SL	A Shifted exponential/Lognormal combination

goal is to obtain approximations for the throughput, λ_D, the variability parameter, $\hat{c}_D^2$ and the mean queue length of each buffer, $\bar{L}_i, i = 1, 2$.

The variability parameter $\hat{c}_D^2$ is beyond the scope of this chapter. In fact characterizing the variability of the departure process from the fork/join station as well as developing two-moment approximations for the variability parameter ($\hat{c}_D^2$) of the departure process, is quite challenging. In particular, the correlations between the successive inter-departure times need to be accounted for so as to capture the impact of such correlations on performance measures (such as queue lengths) at the subsequent queues in the network. These issues require substantial investigation and discussion, and are addressed in a separate paper (Krishnamurthy *et al.* [14]).

Next we provide three practical examples of the queuing system described above.

1 First, the fork/join station described above can represent a synchronization station before an assembly operation in a fabrication/assembly system (Rao and Suri [23, 24]). In this case $K_1(K_2)$ could correspond to the fixed number of automated guided vehicles (AGVs) circulating in the fabrication sub-network $SN_1(SN_2)$ feeding the assembly station. Entities in buffers B_1 and B_2 correspond to fabricated parts that are to be assembled. The join operation corresponds to the kitting operation, while the fork operation corresponds to the release of free AGVs carrying the parts required for assembly. These AGVs would go back to the fabrication sub-networks SN_1 and SN_2 to be restocked with parts.

2 As a second example, the model could represent the synchronization constraint in a kanban control system. In modern manufacturing, kanban systems are a popular form of material control (Buzacott and Shanthikumar [5], Di Mascolo *et al.* [6], Liberopoulos and Dallery [18]). If the fork/join station models the synchronization constraint in a multi-stage kanban system, SN_1 and SN_2 could correspond to upstream and downstream manufacturing stages respectively and $K_1(K_2)$ would be the number of kanbans in stage $SN_1(SN_2)$. Each entity in buffer B_1 corresponds to a part with an upstream kanban attached to it. Each entity in buffer B_2 corresponds to a free kanban returning from the downstream stage. During the join operation a part and upstream kanban are joined

with a downstream kanban and during the fork operation, the upstream kanban is sent back to SN_1, while the part and downstream kanban are sent to SN_2.

3 Finally, this fork/join station model can also be applied to represent the synchronization behavior of parallel programs contending for shared resources in a parallel or distributed computer system. See Heidelberger and Trivedi [9].

2.3. System Assumptions and Approach

As discussed in Section 1, our goal is to obtain two-moment approximations for throughput, λ_D, and mean queue lengths $\bar{L}_1$ and $\bar{L}_2$ at the input buffers B_1 and B_2 for a fork/join station specified by the parameter 6-tuple $(\lambda_1, c_1^2, K_1, \lambda_2, c_2^2, K_2)$. To do so we first study the impact of the mean rates of the input processes (λ_1, λ_2), and population size (K_1, K_2) on the performance of the fork/join station using the exact expressions reported in Som *et al.* [25] and Takahashi *et al.* [28]. 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. 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 Krishnamurthy *et al.* [12], 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 $(0,\infty)$ and SCVs $[0.5,\infty)$. From this analysis we observe the impact of both means and SCVs on the performance measures. Using insights from all the above cases, we develop two-moment approximations for the more general case.

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 cover a significant portion of the range of SCVs observed in practice. In addition many of the prior research on parametric decomposition and two-moment approximations have focused on a similar range of parameters (Albin [2], Buzacott and Shanthikumar [5], Kamath *et al.*[10]).

3. Insights from Exact Analysis

3.1. Impact of Mean Arrival Rates

We will use the notation in Table 1.1 noting that a superscript 'E' denotes the station performance measures that are estimated assuming exponential inter-arrival times. Assuming without loss of generality that $\lambda_1 \leq \lambda_2$, Takahashi *et al.* [28] derive the following expressions for the throughput λ_D^E and mean queue lengths $\bar{L}_i^E, i = 1, 2$ at the fork/join station:

If $\rho = \frac{\lambda_1}{\lambda_2} \neq 1$,

$$\lambda_D^E = \lambda_1 \left(\frac{1 - \rho^{K_1+K_2}}{1 - \rho^{K_1+K_2+1}} \right) \tag{1.1}$$

$$\bar{L}_1^E = \left(\frac{\rho^{K_2+1}}{1-\rho} \right) \left(\frac{1 - \rho^{K_1}}{1 - \rho^{K_1+K_2+1}} \right) - \left(\frac{K_1 \rho^{K_1+K_2+1}}{1 - \rho^{K_1+K_2+1}} \right) \tag{1.2}$$

$$\bar{L}_2^E = \left(\frac{K_2}{1 - \rho^{K_1+K_2+1}} \right) - \left(\frac{\rho}{1-\rho} \right) \left(\frac{1 - \rho^{K_2}}{1 - \rho^{K_1+K_2+1}} \right) \tag{1.3}$$

and if $\rho = \frac{\lambda_1}{\lambda_2} = 1$,

$$\lambda_D^E = \lambda_1 \left(\frac{K_1 + K_2}{K_1 + K_2 + 1} \right) \tag{1.4}$$

$$\bar{L}_i^E = \frac{K_i(K_i + 1)}{2(K_1 + K_2 + 1)} \qquad \text{for} \qquad i = 1, 2. \tag{1.5}$$

Based on these expressions we obtain the following insights about the performance of the fork/join station for the case of exponential inputs.

1 The upper bound of the throughput, λ_D^E from the fork/join station is $\min(\lambda_1, \lambda_2)$. If $\rho = 1$, the bound is nearly achieved for moderate K_1 and K_2.

2 The throughput λ_D^E, depends on the values of K_1 and K_2 only through their sum $(K_1 + K_2)$.

3 When the input rates are unequal, substantial queues are observed at the buffers of the input processes with higher rates of arrivals, i.e., $\bar{L}_1^E \gg \bar{L}_2^E$ when $\lambda_1 > \lambda_2$.

4 When $\rho = 1$, we have $\lambda_D^E < \lambda_1 = \lambda_2$ and $\bar{L}_i^E < \frac{K_i}{2}$ for $K_i < \infty, i = 1, 2$.

Note that taking limits as K_i, $i = 1, 2$ tend to infinity in the equations above we can confirm three of the key observations made in Harrison [7] with regard to the performance of fork/join stations in open networks of queues. In particular, we observe from equation 1.1 that the slower of the two arrival processes controls the throughput from the fork/join station, *i.e.*, when $\lambda_1 < \lambda_2$ the throughput λ_D^E from the fork/join station is λ_1. Further, in this case the mean queue length $\bar{L}_2^E$ at buffer B_2 grows without bound (See equation 1.3). Finally, as shown in Harrison [7], when $\lambda_1 = \lambda_2$ the mean queue lengths at both the buffers grow without bound (See equation 1.5).

3.2. Impact of SCVs of Inter-arrival Times

Next we study the impact of higher moments of the arrival distributions on the performance of the fork/join station. Let the inter-arrival times to the input buffers have means, $1/\lambda_i$, and SCV, c_i^2, for $i = 1, 2$, respectively. Then using this information and the additional constraint of balanced means, one can derive a unique 2-phase Coxian distribution to characterize the inter-arrival times at each input buffer. In Krishnamurthy *et al.* [12] we present an exact analysis of such a system and compute the performance measures. We use the insights gained from the exact analysis to understand the impact of the SCVs of the input processes on station performance measures.

The decision to use these 2-phase Coxian distributions to model the inter-arrival times for given values of means and SCVs is motivated by the following factors. First, exact analytical results are available for the case where the inter-arrival times have 2 phased Coxian distribution. Second, this analysis is valid for the entire range of means and SCVs for which we intend to derive the approximations. Third, choosing other two parameter distributions such as Gamma, shifted exponential or log-normal distributions would imply that new exact analysis would have to be carried out in order to obtain estimates of performance measures. Since our main aim is to develop two-moment approximations for general inputs, we study the impact of higher moments of the arrival distributions on the performance of the fork/join station using the analysis available for the case of Coxian inputs to gain the necessary insights, develop the approximations and then test their performance against simulation for different input distributions. One of the observations from the analysis for Coxian inputs is that the closed form expressions for mean queue length and throughput are substantially more complicated than those obtained for exponential inputs. Therefore we present the insights gained from the numerical results obtained from the exact computations.

The parameter values $(\lambda_1, c_1^2, K_1, \lambda_2, c_2^2, K_2)$ used to obtain insights into the impact of c_i^2, $i = 1, 2$ on the throughput and mean queue lengths are summarized in Table 1.2. Note that each of the SCVs ranges from 0.5 to 4.0. Furthermore, we let the finite population of each entity vary from 2 to 20, in order to determine whether throughput under the more general 2-phase Coxian arrival processes is significantly affected by each finite population size, or whether it depends primarily on the sum of the populations (as it did for the exponential case). We vary these key parameters across the given range of values, for each of several pairs of input arrival rates such that the ratio of arrival rates varies is between 0.3 and 3. For a given pair of arrival rates, we could multiply each arrival rate by some constant, α, which could correspond to changing the arbitrary time unit for specifying the rate by a factor of α. In this case, we would obtain a system throughput that is scaled appropriately. Thus, the ratio of the arrival rates is key determinant of station performance, rather than the magnitude of each arrival rate (e.g., the choice of 1.0 for the equal rates is arbitrary; both could be 0.5 or both could be 2.0).

Table 1.2. Parameters for analysis for Coxian inputs

Parameter	*Values*
(λ_1, λ_2)	(0.3, 1.0), (0.4, 1), (0.5, 1), (0.83, 1.25), (0.9, 1.1), (1.0,1.0), (1.1, 0.9), (1.25,0.83), (2.0, 1.0), (3.0, 1.0)
c_1^2	0.5, 0.8, 1.0, 2.0, 4.0
c_2^2	0.5, 0.8, 1.0, 2.0, 4.0
K_1	2, 4, 6, 8, 10, 20
K_2	2, 4, 6, 8, 10, 20

For each pair of arrival rates, there are 900 system configurations representing different combinations of the other parameter values. It would be tedious to present results for all these cases. Instead we focus on the main insight obtained from the set of results. Sample results are summarized in Tables 1.3 and 1.4, and plotted in Figures 1.2 and 1.3. For additional results, the reader is referred to Krishnamurthy *et al.* [12] and Krishnamurthy *et al.* [14]. The values of λ_D^C, $\bar{L}_1^C$ and $\bar{L}_2^C$ were obtained from the exact analysis for Coxian inputs. The superscript 'C' serves as a reminder that the inter-arrival times have a 2-phase Coxian distribution. Interestingly, one of the insights from the exponential case seems to apply here as well: Tables 1.3 and 1.4 show that, for fixed values of the SCVs, the throughput appears to depend on K_1 and K_2 only through their sum $K_1 + K_2$. Figure 1.2 shows the impact of the SCVs on the values of K_1 and K_2 required to obtain a given throughput while figure 1.3 shows the sensitivity of mean queue lengths to SCVs for different ratios of the arrival rates. The results from the numerical computations lead to the following set of insights:

1 For any values of the SCVs, the upper bound of the throughput, λ_D^C from the fork/join station is $\min(\lambda_1, \lambda_2)$. This throughput is achieved as $(K_1, K_2) \to \infty$.

2 For given values of c_1^2, c_2^2 and $(K_1 + K_2)$ the value of throughput, λ_D^C is insensitive to the choice of K_1 and K_2. See Tables 1.3 and 1.4.

3 As in the case of Poisson inputs, substantial queues are observed at the buffers of the input processes with higher rates of arrivals, i.e., $\bar{L}_2^C \gg \bar{L}_1^C$ when $\lambda_2 > \lambda_1$. See Table 1.3.

4 When input rates are equal, i.e. $\rho = 1$, λ_D^C is quite sensitive to c_1^2, and c_2^2. See Table 1.4 and Figure 1.2.

5 When ρ increases above 1 or decreases below 1, the station performance measures become increasingly less sensitive to the SCVs and are more primarily dependent on ρ. See Table 1.3 and Figure 1.3.

6 Define $c^2 = \frac{c_1^2 + c_2^2}{2}$, that is c^2 is the average of the two arrival SCVs. We observe that for a given ρ and c^2:
If $c^2 < 1$, then $\lambda_D^C > \lambda_D^E$.
If $c^2 > 1$, then $\lambda_D^C < \lambda_D^E$.
If $(\rho - 1)(c^2 - 1) < 0$, then $\bar{L}_1^C > \bar{L}_1^E$ and $\bar{L}_2^C < \bar{L}_2^E$.
If $(\rho - 1)(c^2 - 1) > 0$, then $\bar{L}_1^C < \bar{L}_1^E$ and $\bar{L}_2^C > \bar{L}_2^E$.

Finally we observe from Tables 1.3 and 1.4, and Figures 1.2 and 1.3 that when $0.5 < \rho < 2$, $c_i^2 \neq 1$, for $i = 1, 2$, the performance measures

Table 1.3. Analysis of impact of K_1 and K_2 on system performance for Coxian inputs ($\lambda_1 = 0.83, \lambda_2 = 1.25$)

c_1^2	c_2^2	K_1	K_2	$K_1 + K_2$	$\bar{L}_1^C$	$\bar{L}_2^C$	λ_D^C
0.5	0.5	2	2	4	0.13	1.06	0.81
		4	4	8	0.05	2.80	0.83
		4	6	10	0.01	4.75	0.83
		6	4	10	0.05	2.79	0.83
		2	10	12	0.00	8.74	0.83
		6	6	12	0.01	4.75	0.83
		10	2	12	0.27	1.01	0.83
		4	10	14	0.00	8.74	0.83
		8	6	14	0.01	4.75	0.83
		10	4	14	0.06	2.79	0.83
		6	10	16	0.00	8.74	0.83
		8	8	16	0.00	6.74	0.83
		10	6	16	0.01	4.75	0.83
0.5	4	2	2	4	0.36	1.09	0.72
		4	4	8	0.47	2.40	0.78
		4	6	10	0.32	3.94	0.80
		6	4	10	0.70	2.32	0.80
		2	10	12	0.07	7.44	0.81
		6	6	12	0.48	3.85	0.81
		10	2	12	1.54	0.90	0.81
		4	10	14	0.15	7.32	0.82
		8	6	14	0.62	3.78	0.82
		10	4	14	1.06	2.22	0.82
4	1	2	2	4	0.41	1.01	0.72
		4	4	8	0.58	2.16	0.79
		4	6	10	0.39	3.60	0.80
		6	4	10	0.86	2.08	0.80
		2	10	12	0.08	7.01	0.81
		6	6	12	0.59	3.52	0.81
		10	2	12	1.89	0.82	0.81
		4	10	14	0.18	6.89	0.82
		8	6	14	0.75	3.46	0.82
		10	4	14	1.29	2.00	0.82
4	4	2	2	4	0.47	1.07	0.67
		4	4	8	0.73	2.22	0.74
		4	6	10	0.56	3.59	0.76
		6	4	10	1.10	2.13	0.76
		2	10	12	0.16	6.78	0.78
		6	6	12	0.86	3.48	0.78
		10	2	12	2.25	0.86	0.78
		4	10	14	0.35	6.60	0.79
		8	6	14	1.15	3.39	0.79
		10	4	14	1.77	2.01	0.79
		6	10	16	0.55	6.46	0.80
		8	8	16	0.91	4.83	0.80
		10	6	16	1.41	3.32	0.80

Table 1.4. Analysis of impact of K_1 and K_2 on system performance for Coxian inputs ($\lambda_1 = \lambda_2 = 1$)

c_1^2	c_2^2	K_1	K_2	$K_1 + K_2$	$\bar{L}_1^C$	$\bar{L}_2^C$	λ_D^C
0.5	0.5	2	2	4	0.51	0.51	0.87
		4	4	8	0.99	0.99	0.93
		4	6	10	0.79	1.79	0.95
		6	4	10	1.79	0.79	0.95
		2	10	12	0.16	4.16	0.96
		6	6	12	1.49	1.49	0.96
		10	2	12	4.16	0.16	0.96
		4	10	14	0.56	3.56	0.96
		8	6	14	2.27	1.27	0.96
		10	4	14	3.56	0.56	0.96
		6	10	16	1.11	3.11	0.97
		8	8	16	1.98	1.98	0.97
		10	6	16	3.11	1.11	0.97
0.5	4	2	2	4	0.64	0.73	0.74
		4	4	8	1.17	1.43	0.82
		4	6	10	1.01	2.33	0.85
		6	4	10	1.92	1.24	0.85
		2	10	12	0.33	4.69	0.87
		6	6	12	1.69	2.05	0.87
		10	2	12	4.02	0.38	0.87
		4	10	14	0.79	4.19	0.88
		8	6	14	2.43	1.83	0.88
		10	4	14	3.58	0.97	0.88
4	1	2	2	4	0.73	0.67	0.73
		4	4	8	1.42	1.22	0.81
		4	6	10	1.23	1.98	0.84
		6	4	10	2.31	1.05	0.84
		2	10	12	0.39	4.09	0.85
		6	6	12	2.04	1.75	0.85
		10	2	12	4.64	0.35	0.85
		4	10	14	0.98	3.65	0.87
		8	6	14	2.89	1.56	0.87
		10	4	14	4.16	0.83	0.87
4	4	2	2	4	0.76	0.76	0.68
		4	4	8	1.42	1.42	0.76
		4	6	10	1.26	2.26	0.79
		6	4	10	2.26	1.26	0.79
		2	10	12	0.45	4.45	0.81
		6	6	12	2.03	2.03	0.81
		10	2	12	4.45	0.45	0.81
		4	10	14	1.04	4.04	0.83
		8	6	14	2.85	1.85	0.83
		10	4	14	4.04	1.04	0.83
		6	10	16	1.71	3.71	0.84
		8	8	16	2.62	2.62	0.84
		10	6	16	3.71	1.71	0.84

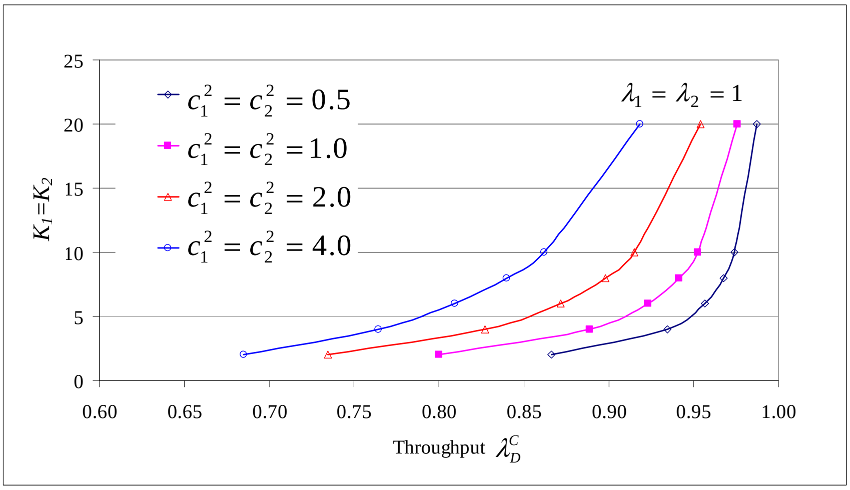


Figure 1.2. Impact of SCVs on $K_1 + K_2$ required to obtain a given λ_D^C

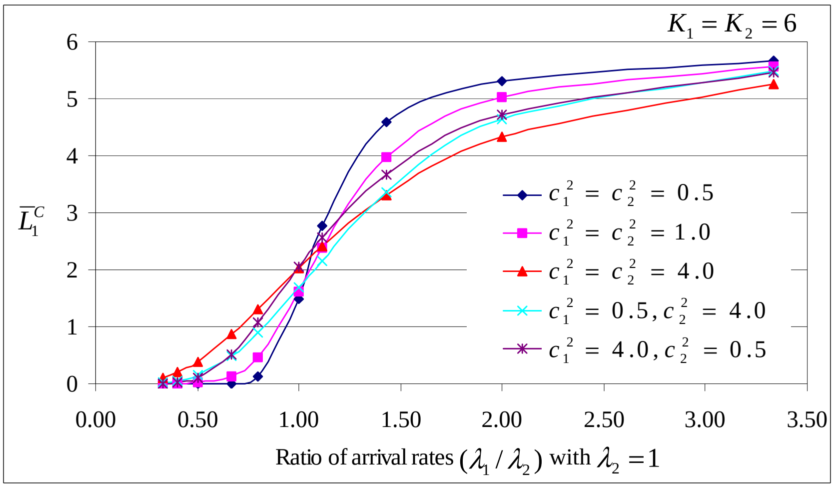


Figure 1.3. Impact of unequal input rates and variabilities on $\bar{L}_1^C$

are quite different from those when $c_i^2 = 1$, for $i = 1, 2$. This implies that the variabilities in the input processes do have a significant impact on the performance measures at the fork/join station. For example Figure 1.2 shows that depending on the input variables significantly different K_i values might be needed to achieve a desired throughput. In other words, there is a substantial impact of variability parameters on system design values. Therefore it is important to incorporate the impact of second moments of inter-arrival time distributions while developing approximations for the performance measures at a fork/join station with more general inputs.

4. General Form of the Two-Moment Approximations

In this section we use the insights gained from the exact analysis presented in the previous section to develop two-moment approximations of fork/join station performance for the case of general inputs. We note that there is a set of probability distributions consistent with every choice of two moments and correspondingly there is a set of possible performance measures associated with any parameter 6-tuple $(\lambda_1, c_1^2, K_1, \lambda_2, c_2^2, K_2)$. The performance measures (mean queue lengths and throughput) derived using the two-moment approximations are therefore regarded as representative of this set. In this section, we develop a general form for these approximations. In subsequent sections we derive the final expressions for the approximations.

The analysis for the case of exponential inputs helps to elucidate the impact of mean arrival rates on the performance measures of the fork/join station. 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:

$$d_{\lambda_D} = \frac{\lambda_D^C - \lambda_D^E}{\lambda_D^E} \tag{1.6}$$

$$d_{\bar{L}_i} = \frac{\bar{L}_i^C - \bar{L}_i^E}{\bar{L}_i^E} \qquad \text{for} \qquad i = 1, 2. \tag{1.7}$$

We compute d_{λ_D}, $d_{\bar{L}_1}$ and $d_{\bar{L}_2}$ for the parameter settings listed in Table 1.2. For a given value of ρ, K_1 and K_2, λ_D^E, $\bar{L}_1^E$, and $\bar{L}_2^E$ are uniquely defined while the values of λ_D^C, $\bar{L}_1^C$, and $\bar{L}_2^C$, depend additionally on the particular values of c_1^2 and c_2^2. The aim of computing d_{λ_D}, $d_{\bar{L}_1}$ and $d_{\bar{L}_2}$ is to gain insight on the impact of c_1^2 and c_2^2. Figures 1.4, 1.5, and 1.6 plot the above quantities against $c^2 - 1$. Figure 1.4 illustrates the variation of d_{λ_D} with c^2-1 for $\rho=1$ and different values of K_1 and K_2. Figure 1.5 illustrates the variation of d_{λ_D} with c^2-1 for $\rho=0.66$ and different values of K_1 and K_2. From these graphs we infer that d_{λ_D} varies roughly linearly with c^2-1, with additional variation depending on ρ, K_1 and K_2. Figure 1.6 illustrates the variation of $d_{\bar{L}_2}$ with c^2-1 for $\rho=0.66$, and different values of K_1 and K_2. From these graphs and symmetry we infer that $d_{\bar{L}_1}$ and $d_{\bar{L}_2}$ vary roughly linearly with c^2-1, with additional variation depending on ρ, K_1 and K_2. Although not illustrated here, we analyzed the behavior for several other values of ρ and found similar trends.

Based on these observations we propose that suitable candidates for the approximation functions would be the following:

$$\lambda_D = \lambda_D^E\left[1+(c^2-1)w_{\lambda_D}\right] \tag{1.8}$$
$$\bar{L}_i = \bar{L}_i^E\left[1+(c^2-1)w_{\bar{L}_i}\right] \qquad \text{for} \qquad i=1,2. \tag{1.9}$$

where w_{λ_D}, $w_{\bar{L}_1}$, and $w_{\bar{L}_2}$ are functions of $(\lambda_1, K_1, \lambda_2, K_2)$.

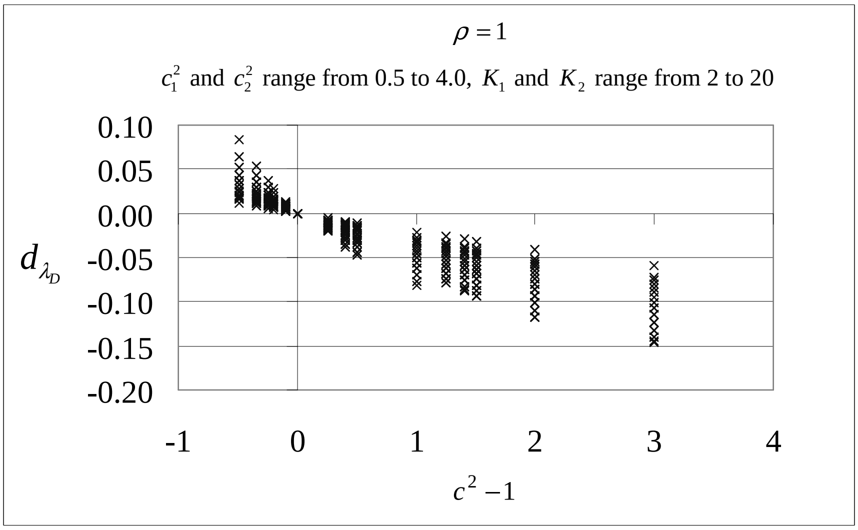


Figure 1.4. Variation of d_{λ_D} with c^2-1 when arrival rates are equal

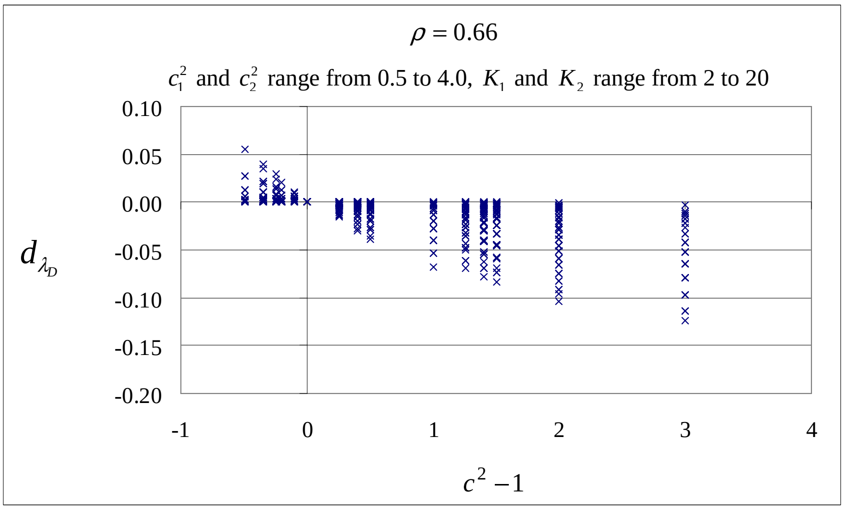


Figure 1.5. Variation of d_{λ_D} with $c^2 - 1$ when arrival rates are unequal

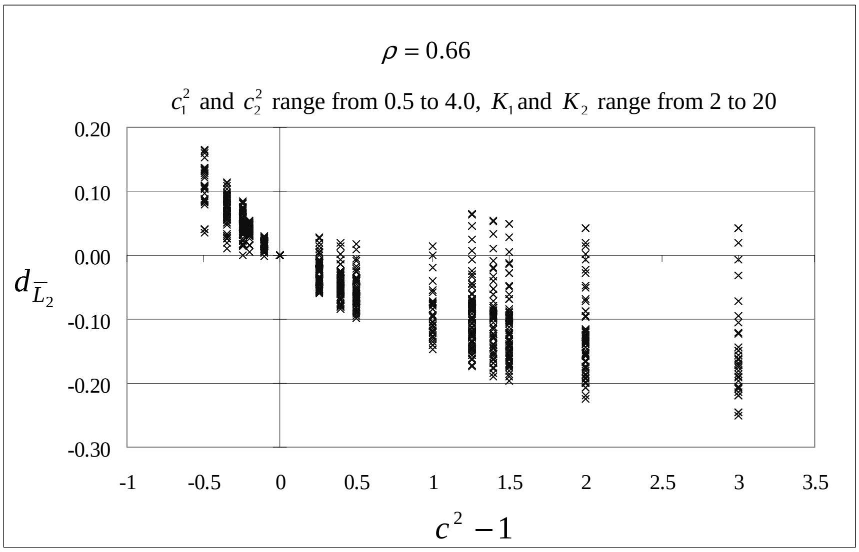


Figure 1.6. Variation of $d_{\bar{L}_2}$ with $c^2 - 1$ when arrival rates are unequal

Note that, when $c_1^2 = c_2^2 = 1$, the approximation functions above yield $\lambda_D = \lambda_D^E$, $\bar{L}_1 = \bar{L}_1^E$ and $\bar{L}_2 = \bar{L}_2^E$. This implies that the approximations are exact for exponential arrivals. The approximation functions need to satisfy all the properties of λ_D^C, $\bar{L}_1^C$ and $\bar{L}_2^C$, stated in Section 3.2. Using the insights obtained from the study for Coxian inputs we identify the general form of w_{λ_D}, $w_{\bar{L}_1}$, and $w_{\bar{L}_2}$. This is discussed in the following sections.

4.1. General Form of w_{λ_D}

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

1 $w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2)$ must be a single valued function of λ_1, λ_2, K_1, and K_2.

2 We require that $w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2) = w_{\lambda_D}(\lambda_2, K_2, \lambda_1, K_1)$ due to symmetry.

3 $w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2) < 0$. This is because when $c^2 < 1$, $\lambda_D^C > \lambda_D^E$ and when $c^2 > 1$, $\lambda_D^C < \lambda_D^E$.

4 $w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2) \to 0$ for $\rho \to \infty$ and $\rho \to 0$. This is because $\lambda_D \to \lambda_D^E$ when $\rho \to 0$ and when $\rho \to \infty$.

5 We require that $w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2) = w_{\lambda_D}(\lambda_1, K_2, \lambda_2, K_1)$. This is evident from the results of Section 3.2.

6 Since $\lambda_D \le \min(\lambda_1, \lambda_2)$ we require $\lambda_D^E(1 + w_{\lambda_D}(c^2 - 1)) \le \min(\lambda_1, \lambda_2)$, implying
$w_{\lambda_D} \le \left[\frac{\min(\lambda_1,\lambda_2)}{\lambda_D^E} - 1\right]\left[\frac{1}{c^2-1}\right]$.

A candidate function $w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2)$ that satisfies properties 1 through 5 above is:

$$w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2) = -a^{'} \rho^{a^{''}} \left[\frac{1-\rho}{1-\rho^{2a^{''}+1}}\right] \tag{1.10}$$

where $a^{'}$ and $a^{''}$ are positive functions of $K_1 + K_2$. Now, we need to determine $a^{'}$ and $a^{''}$ such that $w_{\lambda_D}(\lambda_1, K_1, \lambda_2, K_2)$ also satisfies property 6. We will use the following proposition.

Proposition 1.1 For the range of parameters, $0.5 \le c_i^2 \le 4$, $\frac{1}{3} \le \frac{\lambda_1}{\lambda_2} \le 3$, and $K_i \le 20$,i=1,2:
if $a^{''} = K_1 + K_2$, and $0 \le a^{'} \le 2$, then $\lambda_D \le \min(\lambda_1, \lambda_2)$.

Proof. See Appendix. □

Since properties 1-6 are satisfied from equations 1.1, 1.8 and Proposition 1.1, we suggest the following candidate expression for throughput from a fork/join station characterized by the 6-tuple $(\lambda_1, c_1^2, K_1, \lambda_2, c_2^2, K_2)$: if $\rho \neq 1$,

$$\lambda_D = \lambda_1 \left[\frac{1-\rho^{K_1+K_2}}{1-\rho^{K_1+K_2+1}}\right]\left[1-a^{'}(c^2-1)\left(\frac{(1-\rho)\rho^{K_1+K_2}}{1-\rho^{2(K_1+K_2)+1}}\right)\right] \quad (1.11)$$

Taking limits as $\rho \to \infty$, we obtain the expression when $\rho = 1$. Hence, if $\rho = 1$,

$$\lambda_D = \lambda_1 \left[\frac{K_1+K_2}{K_1+K_2+1}\right]\left[1-a^{'}(c^2-1)\left(\frac{1}{2(K_1+K_2)+1}\right)\right] \quad (1.12)$$

A suitable value of $a^{'}$ will be determined from simulation results in Section 5.

4.2. General Form of $w_{\bar{L}_1}$ and $w_{\bar{L}_2}$

To derive the two-moment approximations for mean queue lengths $\bar{L}_1$ and $\bar{L}_2$ we note that $w_{\bar{L}_i}, i = 1, 2$ must satisfy the following properties:

1 $w_{\bar{L}_i}, i = 1, 2$ is a single valued function of $(\lambda_1, K_1, \lambda_2, K_2)$.

2 $w_{\bar{L}_i}(\lambda_1, K_1, \lambda_2, K_2) \to 0$ when $\rho \to \infty$ and when $\rho \to 0$, since $\bar{L}_i^C \to \bar{L}_i^E$ when $\rho \to \infty$ and when $\rho \to 0$, for $i = 1, 2$.

3 By symmetry $w_{\bar{L}_1}(\lambda_1, K_1, \lambda_2, K_2) = w_{\bar{L}_2}(\lambda_2, K_2, \lambda_1, K_1)$.

4 $\left[w_{\bar{L}_1}(\lambda_1, K_1, \lambda_2, K_2)\right] \times \left[w_{\bar{L}_2}(\lambda_1, K_1, \lambda_2, K_2)\right] \leq 0$, since for a given value of $\lambda_1, K_1, \lambda_2, K_2$ and $c^2 - 1$, $w_{\bar{L}_1}(\lambda_1, K_1, \lambda_2, K_2)$ and $w_{\bar{L}_2}(\lambda_1, K_1, \lambda_2, K_2)$ have to be of opposite sign.

5 For $\rho \leq 1$, $w_{\bar{L}_1}(\lambda_1, K_1, \lambda_2, K_2) \geq 0$, and for $\rho \geq 1$, $w_{\bar{L}_1}(\lambda_1, K_1, \lambda_2, K_2) \leq 0$. This is because as seen from the results of Section 3.2 when $(\rho - 1)(c^2 - 1) \leq 0$, then $\bar{L}_1^C \geq \bar{L}_1^E$, and when $(\rho - 1)(c^2 - 1) \geq 0$, then $\bar{L}_1^C \leq \bar{L}_1^E$.

Similarly, for $\rho \leq 1$, $w_{\bar{L}_2}(\lambda_1, K_1, \lambda_2, K_2) \leq 0$, and for $\rho \geq 1$, $w_{\bar{L}_2}(\lambda_1, K_1, \lambda_2, K_2) \geq 0$. This is because when $(\rho - 1)(c^2 - 1) \leq 0$, then $\bar{L}_2^C \leq \bar{L}_2^E$, and when $(\rho - 1)(c^2 - 1) \geq 0$, then $\bar{L}_2^C \leq \bar{L}_2^E$.

6 Since $\bar{L}_i \leq K_i$, $\bar{L}_i^E \left[1 + (c^2 - 1) w_{\bar{L}_i}\right] \leq K_i$, implying $w_{\bar{L}_i} \leq \left[\frac{K_i}{L_i^E} - 1\right] \left[\frac{1}{c^2 - 1}\right]$, for $i = 1, 2$.

Candidate functions $w_{\bar{L}_i}(\lambda_1, K_1, \lambda_2, K_2), i = 1, 2$ that satisfy properties 1 through 5 above are:

$$w_{\bar{L}_1}(\lambda_1, K_1, \lambda_2, K_2) = \left[\frac{1 - \rho^{b_1'}}{1 + \rho^{b_1'}}\right] \left[\frac{\rho^{b_1''}}{1 + \rho^{2b_1''}}\right] \tag{1.13}$$

and

$$w_{\bar{L}_2}(\lambda_1, K_1, \lambda_2, K_2) = -\left[\frac{1 - \rho^{b_2'}}{1 + \rho^{b_2'}}\right] \left[\frac{\rho^{b_2''}}{1 + \rho^{2b_2''}}\right] \tag{1.14}$$

where b_1', b_1'' and b_2', b_2'' are positive functions of $K_1 + K_2$. Using these functions, the updated expressions for mean queue lengths from a fork/join station characterized by the parameter 6-tuple $(\lambda_1, c_1^2, K_1, \lambda_2, c_2^2, K_2)$ are: if $\rho \neq 1$,

$$\begin{aligned}
\bar{L}_1 \quad = \quad & \left[\left(\frac{\rho^{K_2+1}}{1-\rho}\right)\left(\frac{1-\rho^{K_1}}{1-\rho^{K_1+K_2+1}}\right) - \left(\frac{K_1 \rho^{K_1+K_2+1}}{1-\rho^{K_1+K_2+1}}\right)\right] \times \\
& \left[1 + \left(\frac{1-\rho^{b_1'}}{1+\rho^{b_1'}}\right)\left(\frac{\rho^{b_1''}}{1+\rho^{2b_1''}}\right)(c^2 - 1)\right]
\end{aligned} \tag{1.15}$$

$$\begin{aligned}
\bar{L}_2 \quad = \quad & \left[\left(\frac{K_2}{1-\rho^{K_1+K_2+1}}\right) - \left(\frac{\rho}{1-\rho}\right)\left(\frac{1-\rho^{K_2}}{1-\rho^{K_1+K_2+1}}\right)\right] \times \\
& \left[1 - \left(\frac{1-\rho^{b_2'}}{1+\rho^{b_2'}}\right)\left(\frac{\rho^{b_2''}}{1+\rho^{2b_2''}}\right)(c^2 - 1)\right]
\end{aligned} \tag{1.16}$$

Taking limits as $\rho \to \infty$, we obtain the expression when $\rho = 1$. We have: if $\rho = 1$,

$$\bar{L}_i \quad = \quad \frac{K_i(K_i + 1)}{2(K_1 + K_2 + 1)} \qquad \text{for} \qquad i = 1, 2. \tag{1.17}$$

Now we need to determine b_1', b_1'' and b_2', b_2'' such that $w_{\bar{L}_i}(\lambda_1, K_1, \lambda_2, K_2)$, for $i = 1, 2$ satisfy property 6. These will be determined using the simulation results presented next.

5. Detailed Approximations and their Accuracy

In this section we use simulations to determine the best values of the constants in the approximations. The constants to be determined are the value of $a^{'}$ in equations 1.11 and 1.12 and the values of $b_1^{'}, b_1^{''}$ and $b_2^{'}, b_2^{''}$ in equations 1.15 and 1.16. After determining these constants, we report the percentage difference in the estimates given by the approximations compared with the simulation estimates for a wide range of system parameters. In the simulation experiments, we use distributions of the inter-arrival times to the fork/join station that are different than the 2-phase Coxian distributions. We now describe these simulation experiments.

5.1. Simulation Experiments used for Determining the Approximations

The distributions and ranges of parameter values used in the simulation experiments are summarized in Tables 1.5 and 1.6. As seen from Table 1.5, we evaluate the approximations with inter-arrival times that have 2-stage Erlang, Shifted exponential, Lognormal and Hyper-exponential distributions. We consider two sets of validation experiments. In the 'EH' set of experiments, the inter-arrival times to the two buffers have either a 2-stage Erlang or a Hyper-exponential distribution. Specifically, inter-arrival times with SCV equal to 0.5 are generated using the 2-stage Erlang distribution while inter-arrival times with SCV equal to 1 or 4 are generated using the Hyper-exponential distribution. In the 'SL' experiment, the inter-arrival times to the two buffers have either a Shifted exponential or a Lognormal distribution. Specifically, inter-arrival times with SCV equal to 0.5 are generated using the Shifted exponential distribution while the inter-arrival times with SCV equal to 1 or 4 are generated using the Lognormal distribution. In each set of experiments, the other parameters take on all the values listed in Table 1.6.

The Shifted exponential distribution has a density function given by

$$f(x) = \mu \exp[-\mu(x - d)], \qquad x \geq d. \tag{1.18}$$

For Hyper-exponential distribution, we assume the balanced mean Hyper-exponential distribution. Such a distribution has the density function

Table 1.5. Distributions settings chosen in validation experiments

Input combination	*Inter-arrival time distributions*
EH	2-Stage Erlang for $c_i^2 = 0.5$, Hyper-exponential for $c_i^2 = 1, 4, i = 1, 2$
SL	Shifted Exponential for $c_i^2 = 0.5$, Lognormal for $c_i^2 = 1, 4, i = 1, 2$

Table 1.6. Parameter values in validation experiments

Parameter	*Values*
(λ_1, λ_2)	(0.3, 1.0), (0.83, 1.25), (1.0,1.0)
K_1	2, 6, 10
K_2	2, 6, 10
c_1^2	0.5, 1.0, 4.0
c_2^2	0.5, 1.0, 4.0

$$f(x) = p\mu_1 \exp(-\mu_1 x) + (1-p)\mu_2 \exp(-\mu_2 x), \qquad x \geq 0, \tag{1.19}$$

with the balanced mean constraint $p/(\mu_1) = (1-p)/(\mu_2)$. The constraint implies that the three parameters are uniquely determined from the mean and SCV of the distribution. The Lognormal distribution has the density function

$$f(x) = \left(x\sigma\sqrt{2\pi}\right)^{-1} \exp\left[-(\ln x - \mu)^2/2\sigma^2\right], \qquad x \geq 0. \tag{1.20}$$

The Lognormal distribution has a fatter tail than the Hyper-exponential distribution and the mode of the Lognormal distribution is greater than zero.

Others such as Whitt [31] and Albin [1], have also used all the above distributions to develop and validate two-moment approximations.

The simulation experiments for the validation study were conducted using PROMODEL [22]. We considered 50,000 departures and 5 replications for each run. In addition, for each run a warm up period corresponding to 10,000 departures was chosen. From each run, throughput and mean queue lengths were recorded and 95% confidence intervals were computed for each performance measure of interest obtained from simulation. These were all found to be within 1 percent of the estimates for the mean values. Let λ_D^{EH} and $\bar{L}_i^{EH}, i = 1, 2$ be the estimates of throughput and mean queue lengths obtained using the combination of the 2-stage Erlang and Hyper-exponential distributions. Let λ_D^{SL} and $\bar{L}_i^{SL}, i = 1, 2$ be the simulation estimates of throughput and mean queue lengths obtained using the combination of the Shifted exponential and Lognormal distributions. These values were compared against values from the estimates λ_D and $\bar{L}_i, i = 1, 2$ obtained using the two-moment approximations derived in equations 1.11, 1.12, 1.15, 1.16, and 1.17 respectively.

5.2. Approximation for Throughput and its Accuracy

For a given set of input parameters, $(\lambda_1, c_1^2, K_1, \lambda_2, c_2^2, K_2)$, let $\varepsilon(\lambda_D^{EH})$ and $\varepsilon(\lambda_D^{SL})$ be the absolute percentage difference in the estimate of throughput given by the two-moment approximation and simulation. That is, $\varepsilon(\lambda_D^{EH})$ and $\varepsilon(\lambda_D^{SL})$ are given by:

$$\varepsilon(\lambda_D^{EH}) = \frac{\left|\lambda_D - \lambda_D^{EH}\right|}{\lambda_D^{EH}} \times 100 \qquad \text{and} \qquad \varepsilon(\lambda_D^{SL}) = \frac{\left|\lambda_D - \lambda_D^{SL}\right|}{\lambda_D^{SL}} \times 100$$

Note that this percentage difference depends on the choice of $a^{'}$ in the approximation. For each value of $a^{'}$ in the range [0,2] in increments of 0.5, we compute the percentage difference and then select the value of $a^{'}$ that provides best overall accuracy when both the average and maximum percentage difference is considered. Table 1.7 provides the average and maximum percentage difference for the different values of $a^{'}$ and ρ = 1, 0.67, and 0.3. We note that the performance of the approximations deteriorates as $a^{'}$ deviates from 0.5. In fact the deterioration in performance is more significant when the input processes to the fork/join station have equal rates. Additional experiments for $a^{'}$ values of 0.4 and 0.6 indicated no significant improvement in the approximations.

We select the value of $a^{'} = 0.5$ for our approximation. With this choice, the final expression for the two-moment approximations for throughput is as follows:

If $\rho \neq 1$,

Table 1.7. Maximum (Average) value of $\varepsilon(\lambda_D^{EH})$ and $\varepsilon(\lambda_D^{SL})$ for different a'

For EH Experiments

λ_1	λ_2	$a' = 0$	$a' = 0.5$	$a' = 1.0$	$a' = 1.5$	$a' = 2.0$
1	1	20.5(8.3)	10.9(5.7)	19.7(4.1)	39.8(3.9)	59.9(5.6)
0.83	1.25	17.07(3.9)	11.2(3.4)	9.4(3.0)	18.5(3.1)	30.4(3.3)
0.3	1	5.8(0.6)	4.9(0.6)	4.1(0.6)	3.7(0.5)	3.2(0.5)

For SL Experiments

λ_1	λ_2	$a' = 0$	$a' = 0.5$	$a' = 1.0$	$a' = 1.5$	$a' = 2.0$
1	1	13.9(4.7)	5.7(2.3)	24.3(1.9)	43.2(4.0)	62.1(6.3)
0.83	1.25	11.5(2.0)	6.0(1.4)	11.1(1.5)	22.4(1.7)	33.7(1.9)
0.3	1	4.4(0.3)	3.5(0.3)	2.6(0.3)	1.7(0.2)	1.9(0.2)

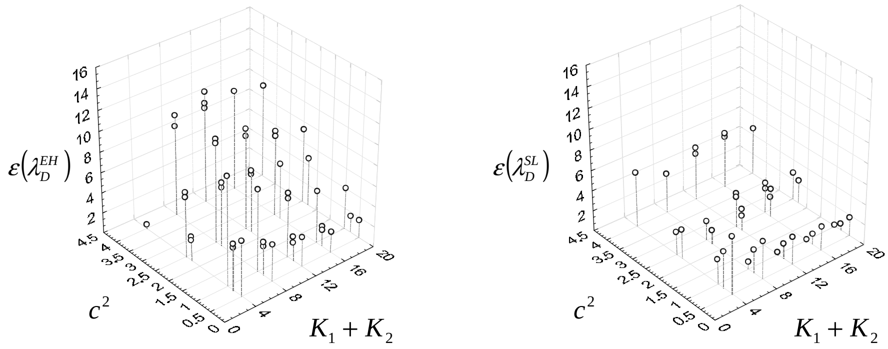
$$\lambda_D = \lambda_1 \left[\frac{1-\rho^{K_1+K_2}}{1-\rho^{K_1+K_2+1}} \right] \left[1 - 0.5(c^2-1) \left(\frac{(1-\rho)\rho^{K_1+K_2}}{1-\rho^{2(K_1+K_2)+1}} \right) \right] \quad (1.21)$$

If $\rho = 1$,

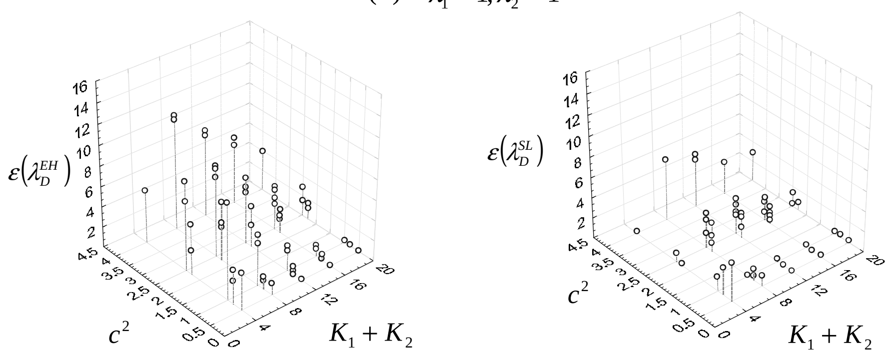
$$\lambda_D = \lambda_1 \left[\frac{K_1+K_2}{K_1+K_2+1} \right] \left[1 - 0.5(c^2-1) \left(\frac{1}{2(K_1+K_2)+1} \right) \right] \quad (1.22)$$

To judge the accuracy of these expressions, in Figure 1.7 the percentage differences $\varepsilon(\lambda_D^{EH})$ and $\varepsilon(\lambda_D^{SL})$ are plotted on a three-dimensional figure against c^2 and $K_1 + K_2$ respectively. In these figures, the parameters c_1^2, c_2^2, K_1 and K_2 take values listed in Table 1.6. The values of λ_D^{EH} and λ_D^{SL} are estimated using simulation, the value of λ_D is computed using the two-moment approximations given by equations 1.21 and 1.22. Figure 1.7 allows us to identify the impact of these parameters on the percentage difference. Figure 1.7(a) presents the results for the case $\rho = 1$, while Figure 1.7(b) presents results for the case $\rho = 0.67$, and Figure 1.7(c) presents results for the case $\rho = 0.3$.

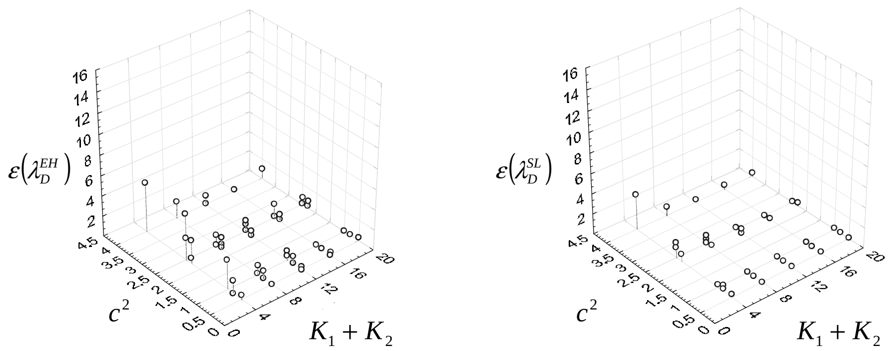
From these figures we observe that the approximations yield fairly accurate estimates for moderately high values of c^2 and $K_1 + K_2$. In addition, the approximation performs better when the input rates are significantly different. Table 1.8 provides additional details on the performance of the approximation. First, we observe that the average difference between the estimates given by the proposed approximation and simulation is under 6% and that the maximum difference in the approx-



$(a) \quad \lambda_1 = 1, \lambda_2 = 1$



$(b) \quad \lambda_1 = 0.83, \lambda_2 = 1.25$



$(c) \quad \lambda_1 = 0.3, \lambda_2 = 1.0$

Figure 1.7. Accuracy of approximations for throughput (λ_D) for different system parameters

imation is less than 12%. Second, we observe that for over 96% of the cases considered in the validation experiment, the percentage difference in estimates of throughput is less than 10%.

Table 1.8. Performance of approximation for λ_D for 'EH' and 'SL' inputs

For EH Experiments

λ_1	λ_2	Percentage of observations for which		Average	Maximum
		$\varepsilon(\lambda_D^{EH}) \leq 5\%$	$\varepsilon(\lambda_D^{EH}) \leq 10\%$	$\varepsilon(\lambda_D^{EH})$	$\varepsilon(\lambda_D^{EH})$
1	1	39.5	96.3	5.7	10.9
0.8	1.25	72.8	97.5	3.4	11.2
0.3	1	100.0	100.0	0.6	4.9

For SL Experiments

λ_1	λ_2	Percentage of observations for which		Average	Maximum
		$\varepsilon(\lambda_D^{SL}) \leq 5\%$	$\varepsilon(\lambda_D^{SL}) \leq 10\%$	$\varepsilon(\lambda_D^{SL})$	$\varepsilon(\lambda_D^{SL})$
1	1	93.8	100.0	2.3	5.7
0.8	1.25	95.1	100.0	1.4	6.0
0.3	1	100.0	100.0	0.3	3.5

5.3. Approximations for Average Queue Length and their Accuracy

Let $\varepsilon(\bar{L}_i^{EH})$ and $\varepsilon(\bar{L}_i^{SL})$, $i = 1, 2$ be the absolute percent differences in the estimates of average queue lengths, between the two-moment approximations and those obtained from the simulation experiments. That is, $\varepsilon(\bar{L}_i^{EH})$ and $\varepsilon(\bar{L}_i^{SL})$, $i = 1, 2$ are given by:

$$\begin{aligned} \varepsilon(\bar{L}_i^{EH}) &= \frac{\left|\bar{L}_i - \bar{L}_i^{EH}\right|}{K_i} \times 100, i = 1, 2, \qquad \text{and} \\ \varepsilon(\bar{L}_i^{SL}) &= \frac{\left|\bar{L}_i - \bar{L}_i^{SL}\right|}{K_i} \times 100, i = 1, 2. \end{aligned}$$

The percentage differences computed depend on the choice of $b_1^{'}, b_1^{''}$ and $b_2^{'}, b_2^{''}$ used in the approximations. Properties 1 through 6 in Section 4.2 imply that $b_1^{'} = b_2^{'}$, and $b_1^{''} = b_2^{''}$. For $b_i^{'}, i = 1, 2$ we choose values in the range [0,2] in increments of 0.5. Further for simplicity, we choose values of $b_i^{''}, i = 1, 2$ in the range [0,4] and restrict our choice to integer

Table 1.9. Maximum (Average) value of $\varepsilon(\bar{L}_2^{EH})$ and $\varepsilon(\bar{L}_2^{SL})$ for different $b_2^{'}$ and $b_2^{''}$

For EH Experiments

λ_1	λ_2	$b_2^{'} = 0.5$, $b_2^{''} = 1.0$	$b_2^{'} = 1.0$, $b_2^{''} = 2.0$	$b_2^{'} = 1.0$, $b_2^{''} = 4.0$	$b_2^{'} = 2.0$, $b_2^{''} = 2.0$	$b_2^{'} = 2.0$, $b_2^{''} = 4.0$
0.8	1.25	12.8 (5.2)	14.2 (4.6)	13.8 (5.4)	24.8 (6.4)	14.1 (4.6)
0.3	1	17.9 (5.9)	8.9 (2.9)	5.6 (2.3)	16.2 (5.4)	4.9 (2.1)

For SL Experiments

λ_1	λ_2	$b_2^{'} = 0.5$, $b_2^{''} = 1.0$	$b_2^{'} = 1.0$, $b_2^{''} = 2.0$	$b_2^{'} = 1.0$, $b_2^{''} = 4.0$	$b_2^{'} = 2.0$, $b_2^{''} = 2.0$	$b_2^{'} = 2.0$, $b_2^{''} = 4.0$
0.8	1.25	8.8 (3.0)	13.1 (3.5)	8.9 (3.2)	24.4 (7.3)	12.9 (3.5)
0.3	1	20.2 (6.5)	10.9 (3.6)	3.1 (1.1)	18.5 (5.9)	3.0 (0.9)

values. For each pair of values for $b_i^{'}, b_i^{''}, i = 1, 2$ we first check whether $w_{\bar{L}_1}$, and $w_{\bar{L}_2}$ satisfy properties 1 through 6 identified in Section 4.2 and if so, we also compute the percentage differences $\varepsilon(\bar{L}_i^{EH})$ and $\varepsilon(\bar{L}_i^{SL})$ for $i = 1, 2$. Using these values for the percentage differences, we determine the values of $b_1^{'}, b_1^{''}$ and $b_2^{'}, b_2^{''}$ to be used in the final approximation.

This course grained search for good rather than optimal values for the approximation parameters is consistent with other efforts to develop two-moment approximations such as those reported in Albin [1] and Albin [2], and achieves a reasonable balance between efficiency in deriving the parameters and accuracy of the final approximations. More robust techniques could be employed if greater accuracy is desirable. Table 1.9 provides the maximum and average percentage difference, $\varepsilon(\bar{L}_2^{EH})$ and $\varepsilon(\bar{L}_2^{SL})$ for the choices of $b_2^{'}$ and $b_2^{''}$ that gave reasonably good performance. In these tables, we focus on the queue length at buffer B_2 since this buffer has more significant queues when $\rho < 1$. Among the feasible set of values, we observe that setting $b_1^{'} = b_2^{'} = 1$, and $b_1^{''} = b_2^{''} = 4$, gives high accuracy.

Therefore our final expressions for the mean queue lengths from a fork/join station are:

If $\rho \neq 1$,

$$\bar{L}_1 = \left[\left(\frac{\rho^{K_2+1}}{1-\rho}\right)\left(\frac{1-\rho^{K_1}}{1-\rho^{K_1+K_2+1}}\right) - \left(\frac{K_1\rho^{K_1+K_2+1}}{1-\rho^{K_1+K_2+1}}\right)\right] \times \left[1 + \left(\frac{1-\rho}{1+\rho}\right)\left(\frac{\rho^4}{1+\rho^8}\right)(c^2-1)\right] \tag{1.23}$$

$$\bar{L}_2 = \left[\left(\frac{K_2}{1-\rho^{K_1+K_2+1}}\right) - \left(\frac{\rho}{1-\rho}\right)\left(\frac{1-\rho^{K_2}}{1-\rho^{K_1+K_2+1}}\right)\right] \times \left[1 + \left(\frac{1-\rho}{1+\rho}\right)\left(\frac{\rho^4}{1+\rho^8}\right)(c^2-1)\right] \tag{1.24}$$

If $\rho = 1$,

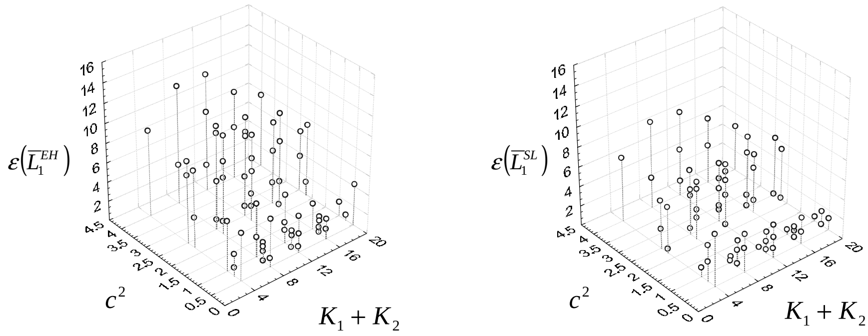
$$\bar{L}_i = \frac{K_i(K_i+1)}{2(K_1+K_2+1)} \qquad \text{for } i = 1, 2. \tag{1.25}$$

Figures 1.8 and 1.9 plot the percentage differences $\varepsilon(\bar{L}_i^{EH})$ and $\varepsilon(\bar{L}_i^{SL})$, for $i = 1, 2$ computed using the two-moment approximations given by equations 1.23, 1.24 and 1.25 against c^2 and $K_1 + K_2$. Figure 1.8 compares $\varepsilon(\bar{L}_1^{EH})$ and $\varepsilon(\bar{L}_1^{SL})$, while Figure 1.9 compares $\varepsilon(\bar{L}_2^{EH})$ and $\varepsilon(\bar{L}_2^{SL})$ for different values of ρ.

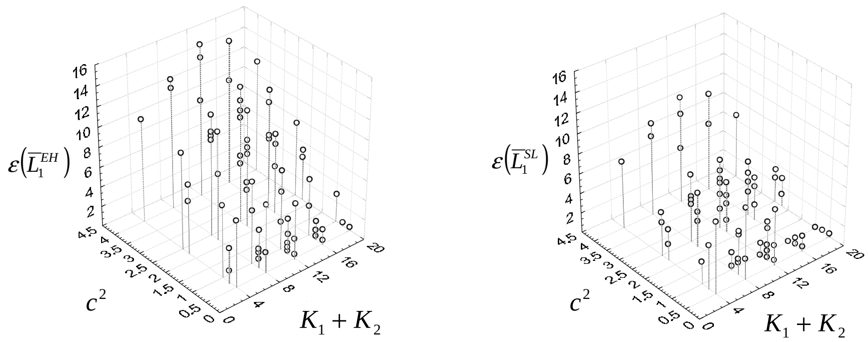
From these figures we observe that like the approximations for the throughput, the approximations for mean queue lengths yield quite accurate estimates (i.e. under 15% error) over the entire ranges of c^2 and $K_1 + K_2$ considered. The approximation performs better when $c^2 < 1$ or the input rates are significantly different. Additional details are presented in Tables 1.10 and 1.11. First, we observe that the average difference between the estimates given by the proposed approximation and simulation is under 7% and that the maximum difference in the approximation is roughly 15%. Second, we observe that for around 90% of the cases considered in the validation experiments, the percentage difference in the estimates of the more significant queue $\bar{L}_2$ is less than 10%.

6. A Numerical Example

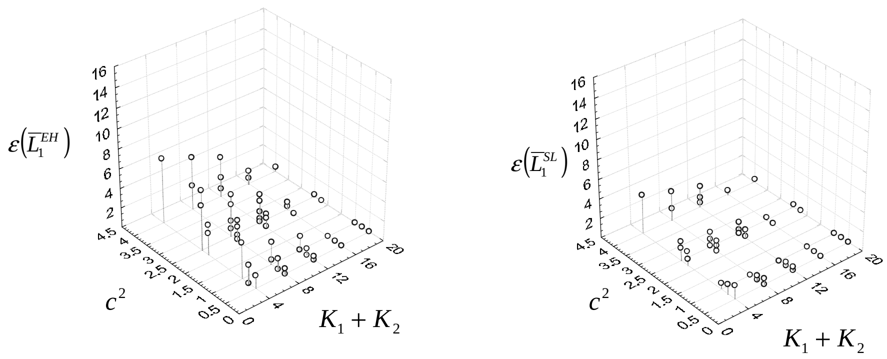
We present a numerical example to demonstrate the usefulness of the proposed approximations. Consider an application where the fork/join



$(a) \quad \lambda_1 = 1, \lambda_2 = 1$

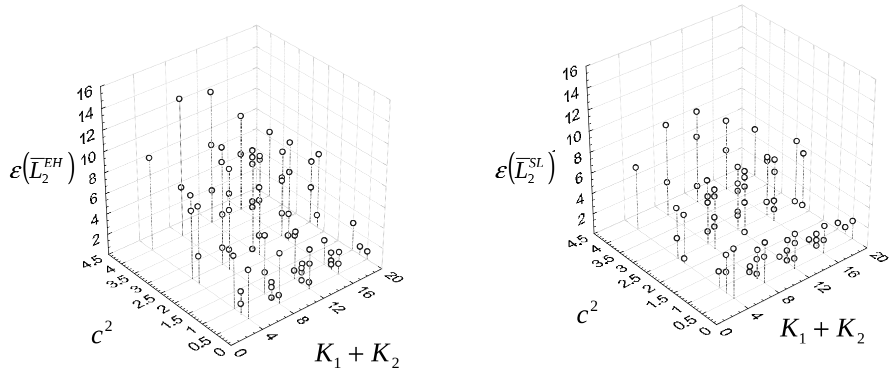


$(b) \quad \lambda_1 = 0.83, \lambda_2 = 1.25$

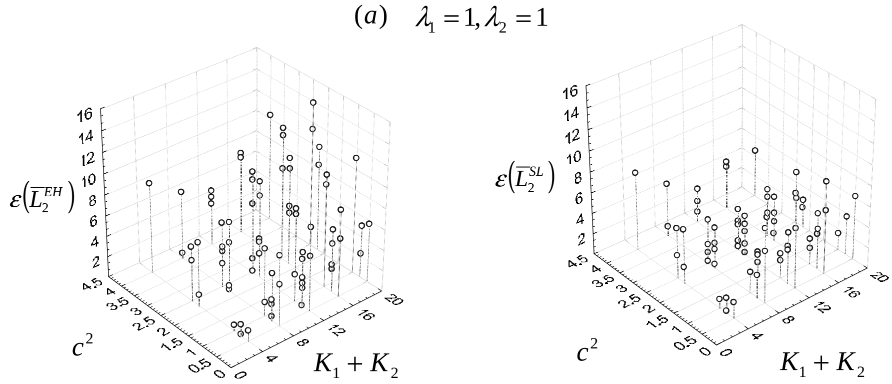


$(c) \quad \lambda_1 = 0.3, \lambda_2 = 1.0$

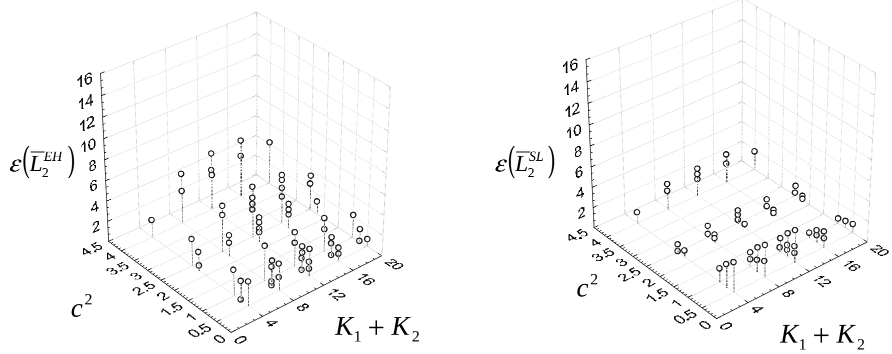
Figure 1.8. Accuracy of approximations for mean queue length ($\bar{L}_1$) at buffer B_1 for different system parameters



$(a)\quad \lambda_1 = 1, \lambda_2 = 1$



$(b)\quad \lambda_1 = 0.83, \lambda_2 = 1.25$



$(c)\quad \lambda_1 = 0.3, \lambda_2 = 1.0$

Figure 1.9. Accuracy of approximations for mean queue length ($\bar{L}_2$) at buffer B_2 for different system parameters

Table 1.10. Performance of approximation for $\bar{L}_1$ for 'EH' and 'SL' inputs

For EH Experiments

λ_1	λ_2	Percentage of observations for which		Average	Maximum
		$\varepsilon(\bar{L}_1^{EH}) \leq 5\%$	$\varepsilon(\bar{L}_1^{EH}) \leq 10\%$	$\varepsilon(\bar{L}_1^{EH})$	$\varepsilon(\bar{L}_1^{EH})$
1	1	61.7	96.3	4.3	13.3
0.8	1.25	39.5	76.5	6.3	15.2
0.3	1	96.3	100.0	1.0	6.6

For SL Experiments

λ_1	λ_2	Percentage of observations for which		Average	Maximum
		$\varepsilon(\bar{L}_1^{SL}) \leq 5\%$	$\varepsilon(\bar{L}_1^{SL}) \leq 10\%$	$\varepsilon(\bar{L}_1^{SL})$	$\varepsilon(\bar{L}_1^{SL})$
1	1	79.0	100.0	2.8	8.8
0.8	1.25	76.5	98.8	3.4	10.7
0.3	1	100.0	100.0	0.5	4.0

Table 1.11. Performance of approximation for $\bar{L}_2$ for 'EH' and 'SL' inputs

For EH Experiments

λ_1	λ_2	Percentage of observations for which		Average	Maximum
		$\varepsilon(\bar{L}_2^{EH}) \leq 5\%$	$\varepsilon(\bar{L}_2^{EH}) \leq 10\%$	$\varepsilon(\bar{L}_2^{EH})$	$\varepsilon(\bar{L}_2^{EH})$
1	1	61.7	96.3	4.3	13.3
0.8	1.25	46.9	88.9	5.4	13.8
0.3	1	97.5	100.0	2.3	5.6

For SL Experiments

λ_1	λ_2	Percentage of observations for which		Average	Maximum
		$\varepsilon(\bar{L}_2^{SL}) \leq 5\%$	$\varepsilon(\bar{L}_2^{SL}) \leq 10\%$	$\varepsilon(\bar{L}_2^{SL})$	$\varepsilon(\bar{L}_2^{SL})$
1	1	79.0	100.0	2.8	8.9
0.8	1.25	84.0	100.0	3.2	8.9
0.3	1	100.0	100.0	1.1	3.1

station models the synchronization constraint in a closed loop fabrication/assembly system. Let sub-networks SN_1 and SN_2 supply components to the synchronization station and let K_1 and K_2 be the number of fixed AGVs circulating in these sub-networks. Then, depending upon the congestion and the random delays in the sub-networks SN_1 and SN_2, the supply of components to the synchronization station could have different variability. Consider the specific situation where the input processes to the two buffers have inter-arrival times with a Lognormal distribution having mean equal to 1 and SCV equal 2.3.

In such a situation, a typical managerial decision is to choose the appropriate number of AGVs, K_1, K_2 in these sub-networks such that they adequately buffer against the variability to obtain a target throughput. Since an exact solution for this system is not available, the tradeoff between the number of AGVs and system throughput can be obtained either using simulation or using the two-moment approximations proposed here. Figure 1.10 shows the results as computed using simulation and two-moment approximations. On the same graph we also plot this curve for exponentially distributed inter-arrival times.

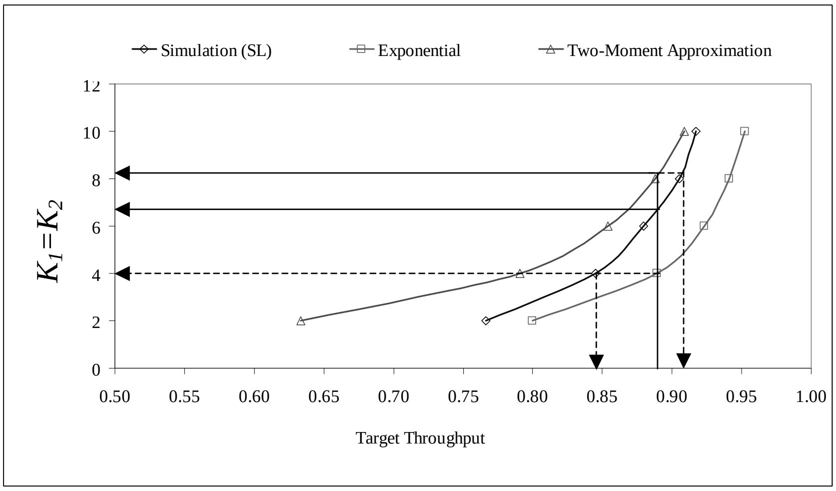


Figure 1.10. Numerical example illustrating use of two-moment approximations

We make several observations from this graph. First, the throughput increases monotonically with K_1 and K_2. Second, the value of K_1 $(= K_2)$ required to obtain a target throughput significantly increases with variability. Consider the number of AGVs required to obtain a target throughput of 0.88. The simulation results indicate that we need $K_1 = K_2 = 7$ AGVs to achieve a target throughput of 0.88. However, the curve obtained assuming exponential inputs would suggest that the required number of AGVs is 4. For closed loop fabrication/assembly systems, this difference would imply significant investment in AGVs to buffer against variability of input processes. Ignoring the impact of variability in the input processes would result in significantly lower throughput than anticipated otherwise. This implies that analyzing fork/join stations for inputs more general than the Poisson process has important practical implications. Third, we note that the two-moment approximations yield values of $K_1 = K_2 = 8$ that are very close to that obtained from the simulation model. In the real system, setting the number of AGVs to 4 would have resulted in throughput much lower than the target throughput. In the presence of high variability, the two-moment approximations on the other hand appear to provide reasonable (and in this case slightly conservative) estimates for making decisions on system design.

Using the two-moment approximations has at least two key advantages over simulation. First, it is computationally far more efficient. Second, the simulation models require the detailed inter-arrival time distributions. In many practical situations, such detailed information is not available. In contrast, the two-moment approximations requires only a basic measure of variability, namely the SCV to obtain the required system design insights.

7. Conclusions and Extensions

In this chapter, 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. For the test cases presented here, we observed that the maximum difference in the estimates provided by the approximations, compared with simulation was 12% for station throughput and 15% for mean queue lengths. These approximations for the fork/join station finds direct applications in the performance analysis of several manufacturing and computer systems.

Although these approximations can be used to analyze a fork/join station in isolation, a principal application of these approximations 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, $\hat{c}_D^2$. Whitt [31, 33, 34, 36] reports several issues that need to be addressed when determining the variability parameter in the context of simple queues. He observes that the squared coefficient of variation (c_D^2) of the inter-departure times is a good estimate of the variability parameter $\hat{c}_D^2$ only if the successive inter-departure times are not correlated. Our preliminary simulation experiments show that this is not true for the departure process from a fork/join station with inputs from finite populations. Consequently, in addition to determining the squared coefficient of variation of the inter-departure times from the fork/join station, we also need to analyze the impact of correlations on the different performance measures. Based on the insights obtained from this analysis, suitable two-moment approximations for the variability parameter need to be derived. In Krishnamurthy *et al.* [14] we address these issues and develop the necessary two-moment approximations.

The approximations developed in this research are for fork/join stations with two inputs. However, in queuing models of many manufacturing and computer applications, we need to model the synchronization of more than than two arrival processes. The analysis presented here can be used to analyze such synchronization constraints in many ways. One possible approach would be to model the fork/join station with N arrival processes as a series arrangement of N-1 fork/join stations, each with two arrival processes. In this arrangement the two arrival processes at a typical fork/join station would be (i) one of the N original arrival processes and (ii) the departure process from the preceding fork/join station. Then by successive application of the approximations developed here, approximations for performance measures of a fork/join station with inputs from several merging processes could be derived. Testing the accuracy of this approach is part of our ongoing research.

The two-moment approximations proposed in this chapter are key building blocks in the effort to develop parametric decomposition methods to analyze queuing networks with fork/join stations. As an example of the application of these building blocks, Krishnamurthy [13] analyzes single and multi-stage kanban systems. Using the approximations proposed here, the queuing networks are analyzed by solving a system of non-linear equations in the set of unknown parameters. The preliminary results reported therein indicate that using these approximations as a building block results in reasonably accurate estimates of network performance. Based on these results, we conjecture that the approximations proposed here will be useful in analyzing queuing networks models of several other systems.

8. Appendix

Proof of Proposition 1.1. For simplicity of notation let $K = K_1 + K_2$. Since $\rho = \lambda_1/\lambda_2$, we have:

$$\lambda_D = \lambda_1 \left[\frac{1-\rho^K}{1-\rho^{K+1}}\right] \left[1 - a^{'}(c^2-1)\left(\frac{(1-\rho)\rho^{a^{''}}}{1-\rho^{2a^{''}+1}}\right)\right] \tag{1.26}$$

Let:

$$f(a^{'}, a^{''}) = \lambda_1 \left[\frac{1-\rho^K}{1-\rho^{K+1}}\right] \left[1 - a^{'}(c^2-1)\left(\frac{(1-\rho)\rho^{a^{''}}}{1-\rho^{2a^{''}+1}}\right)\right] - \min(\lambda_1, \lambda_2) \tag{1.27}$$

Then:

$$f(a^{'}, K) = \lambda_1 \left[\frac{1-\rho^K}{1-\rho^{K+1}}\right] \left[1 - a^{'}(c^2-1)\left(\frac{(1-\rho)\rho^{K}}{1-\rho^{2K+1}}\right)\right] - \min(\lambda_1, \lambda_2) \tag{1.28}$$

Case 1 $(\rho < 1)$:

$$\begin{aligned} & f(a^{'}, K) \\ = \ & \lambda_1 \left[\frac{1-\rho^K}{1-\rho^{K+1}}\right] \left[1 - a^{'}(c^2-1)\left(\frac{(1-\rho)\rho^{K}}{1-\rho^{2K+1}}\right)\right] - \lambda_1 \\ & \left[1 - a^{'}(c^2-1)\left(\frac{(1-\rho)\rho^{K}}{1-\rho^{2K+1}}\right) - \frac{1-\rho^{K+1}}{1-\rho^K}\right] \\ = \ & -\lambda_1 \rho^K (1-\rho) \left[\frac{1-\rho^K}{1-\rho^{K+1}}\right] \left[\frac{1}{1-\rho^K} + \left(\frac{a^{'}(c^2-1)}{1-\rho^{2K+1}}\right)\right] \end{aligned}$$

It is obvious from the above that for $c^2 \geq 1$, $f(a^{'}, K) \leq 0$.
Simplifying further, for $c^2 \leq 1$ we have:

$$\begin{aligned} & f(a^{'}, K) \\ = \ & \frac{-\lambda_1 \rho^K (1-\rho)}{1-\rho^{2K+1}} \left[\frac{1-\rho^K}{1-\rho^{K+1}}\right] \left[1 + \frac{\rho^K \sum_0^K \rho^i}{\sum_0^{K-1} \rho^i} - a^{'}(1-c^2)\right] \end{aligned}$$

From the above it is obvious that $f(a^{'}, K) \leq 0$ for $a^{'} \leq \frac{1}{1-c^2}$, i.e. $a^{'} \leq 2$.

Case 2 ($\rho > 1$):

$$\begin{aligned}
& f(a^{'}, K) \\
= \;& \lambda_1 \left[\frac{\rho^K - 1}{\rho^{K+1} - 1}\right] \left[1 - a^{'}(c^2 - 1)\left(\frac{(\rho - 1)\rho^K}{\rho^{2K+1} - 1}\right)\right] \\
& -\lambda_2 \left[1 - a^{'}(c^2 - 1)\left(\frac{(\rho - 1)\rho^K}{\rho^{2K+1} - 1}\right) - \frac{\rho^{K+1} - 1}{\rho(\rho^K - 1)}\right] \\
= \;& \lambda_1 \left[\frac{\rho^K - 1}{\rho^{K+1} - 1}\right] \left[\frac{(\rho - 1)}{\rho(\rho^K - 1)} - a^{'}(c^2 - 1)\left(\frac{(\rho - 1)\rho^K}{\rho^{2K+1} - 1}\right)\right] \\
= \;& -\lambda_1(\rho - 1) \left[\frac{\rho^K - 1}{\rho^{K+1} - 1}\right] \left[\frac{1}{\rho(\rho^K - 1)} + \left(\frac{a^{'}(c^2 - 1)\rho^K}{\rho^{2K+1} - 1}\right)\right]
\end{aligned}$$

It is obvious from the above that for $c^2 \geq 1, f(a^{'}, K) \leq 0$
Simplifying further, for $c^2 \leq 1$ we have:

$$\begin{aligned}
& f(a^{'}, K) \\
= \;& -\lambda_1 \left[\frac{\rho^K - 1}{\rho^{K+1} - 1}\right] \left[\frac{\rho^K(\rho - 1)}{\rho^{2K+1} - 1}\right] \times \\
& \left[1 + \frac{\rho^{K+1} - 1}{\rho^{2K+1} - \rho^{K+1}} - a^{'}(1 - c^2)\right]
\end{aligned}$$

Hence it is obvious that $f(a^{'}, K) \leq 0$ for $a^{'} \leq \frac{1}{1-c^2}$, i.e. $a^{'} \leq 2$. □

References

[1] Albin, S.L., On Poisson Approximations for Superposition of Arrival Processes in Queues. *Management Science*, **28(2)**(1982), 126-137.

[2] Albin, S.L., Approximating a Point Process by a Renewal Process. II. Superposition Arrival Processes to Queues. *Operations Research*, **32(5)**(1984), 1133-1162.

[3] Baccelli, F., Massey, W.A. and Towsley, D., Acyclic Fork/Join Queuing Networks. *Journal of ACM*, **36(1)**(1989), 615-642.

[4] Bhat, U.N., Finite Capacity Assembly Like Queues. *Queuing Systems*, **1**(1986), 85-101.

[5] Buzacott, J.A. and Shanthikumar, J.G., *Stochastic Models of Manufacturing Systems*, Prentice Hall, New Jersey, 1993.

[6] Di Mascolo, M., Frein, Y. and Dallery, Y., An Analytical Method for Performance Evaluation of Kanban Controlled Production Systems. *Operations Research*, **44(1)**(1996), 50-64.

[7] Harisson,J.M., Assembly-like Queues. *Journal of Applied Probability*, **10(2)**(1973), 354-367.

[8] Hopp, W. and Simon, J.T., Bounds And Heuristics For Assembly-Like Queues. *Queuing Systems*, **4**(1989), 137-156.

[9] Heidelberger, P. and Trivedi, K. S., Queuing Network Models for Parallel Processing with Asynchronous Tasks. *IEEE Transactions on Computers*, **C-31(11)**(1983), 1099-1109.

[10] Kamath, M., Suri, R. and Sanders, J. L., Analytical Performance Models for Closed Loop Flexible Assembly Systems. *The International Journal of Flexible Manufacturing Systems*, **1**(1988), 51-84.

[11] Krishnamurthy, A., Suri, R. and Vernon, M., Analytical Performance Analysis of Kanban Systems Using a New Approximation for Fork/Join Stations. *Proceedings of the Industrial Engineering Research Conference* Dallas, Texas, May 2001.

[12] Krishnamurthy, A., Suri, R. and Vernon, M., Analysis of a Fork/Join Synchronization Station with Coxian Inputs From a Finite Population. *Submitted to Annals of Operations Research*, (2000).

[13] Krishnamurthy, A., Analytical Performance Models for Material Control Strategies in Manufacturing Systems. *PhD. Thesis, Department of Industrial Engineering, University of Wisconsin-Madison*, (2002).

[14] Krishnamurthy, A., Suri, R. and Vernon, M., Two-Moment Approximations for the Variability Parameter of the Departure Process from a Fork/Join Station with General Inputs. *Technical Report, Department of Industrial Engineering, University of Wisconsin-Madison*, (2002).

[15] Kuehn, P.J., Approximate Analysis of General Queuing Networks by Decomposition. *IEEE Transactions on Communications*, **COM-27**(1979), 113-126.

[16] Kimura, T., A Two-Moment Approximation for the Mean Waiting Time in the GI/G/s Queue. *Management Science*, **32(6)**(1986), 751-763.

[17] Latouche, G., Queues With Paired Customers. *Journal of Applied Probability*, **18**(1981), 684-696.

[18] Liberopoulous, G. and Dallery, Y., A Unified Framework for Pull Control Mechanisms in Multistage Manufacturing Systems. *Annals of Operations Research*, **93**(2000), 325-355.

[19] Lipper, E.H. and Sengupta, B., Assembly Like Queues With Finite Capacity: Bounds, Asymptotic And Approximations. *Queuing Systems*, **1**(1986), 67-83.

[20] Marshall, K.T., Some Inequalities in Queuing. *Operations Research*, **16**(1968), 661-665.

[21] Prabhakar, B., Bambos, N. and Mountford, T.S., The Synchronization of Poisson Processes and Queuing Networks with Service and Synchronization Nodes. *Advances in Applied Probability*, **32**(2000), 824-843.

[22] *www.promodel.com.*

[23] Rao, P.C. and Suri, R., Approximate Queuing Network Models of Fabrication/Assembly Systems: Part I-Single Level Systems. *Production and Operations Management*, **3(4)**(1994), 244-275.

[24] Rao, P.C. and Suri, R., Performance Analysis Of An Assembly Station With Input From Multiple Fabrication Lines. *Production and Operations Management*, **9(3)**(2000), 283-302.

[25] Som, P., Wilhelm, P.E. and Disney, R.L., Kitting Process in a Stochastic Assembly System. *Queuing Systems*, **17**(1994), 471-490.

[26] Suri, R., Sanders, J.L. and Kamath, M., Performance Evaluation of Production Networks. In: *Handbooks of OR/MS (vol. 4)*, S. Graves, A. Rinooy Kan and P. Zipkin (eds.), North Holland, 1993.

[27] Suri, R., Diehl, G.W., deTreville, S. and Tomsicek, M., From CAN-Q to MPX: Evolution of Queuing Software for Manufacturing. *Interfaces*, **25(5)**(1995), 128-150.

[28] Takahashi, M., Osawa, H. and Fujisawa, T., A Stochastic Assembly System With Resume Levels. *Asia-Pacific Journal of Operations Research*, **15**(1996), 127-146.

[29] Takahashi, M., Osawa, H. and Fujisawa, T., On A Synchronization Queue With Two Finite Buffers. *Queuing Systems*, **36**(2000), 107-123.

[30] Takahashi, M. and Takahashi, Y., Synchronization Queue With Two MAP Inputs And Finite Buffers. *Proceedings of the Third International Conference on Matrix Analytical Methods in Stochastic Models*, Leuven, Belgium, 2000.

[31] WHITT, W., Approximating a Point Process by a Renewal Process, I: Two Basic Methods. *Operations Research*, **30(1)**(1982), 125-147.

[32] WHITT, W., The Queuing Network Analyzer. *Bell Systems Technical Journal*, **62**(1983), 2779-2815.

[33] WHITT, W., Approximations to Departure Processes and Queues in Series. *Naval Research Logistics Quarterly*, **31**(1984), 499-521.

[34] WHITT, W., The Best Order of Queues in Series. *Management Science*, **31**(1985), 475-487.

[35] WHITT, W., Approximations for the GI/G/m Queue. *Production and Operations Management*, **2(2)**(1993), 114-161.

[36] WHITT, W., Variability Functions for Parametric Decomposition Approximations of Queuing Networks. *Management Science*, **41**(1995), 1704-1715.

[37] VARKI, E., Mean Value Technique For Closed Fork-Join Networks. *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, Atlanta, Georgia, 1999.