Higher-order distributional properties in closed queueing networks

Alexandre Brandwajn
Baskin School of Engineering
University of California Santa Cruz
USA
+1 831 459 4023
alexb@soe.ucsc.edu

Thomas Begin
Laboratoire LIP6 - CNRS
Université Pierre et Marie Curie
France
+33 1 44 27 88 38
thomas.begin@lip6.fr

ABSTRACT

In many real-life computer and networking applications, the distributions of service times, or times between arrivals of requests, or both, can deviate significantly from the memoryless negative exponential distribution that underpins the product-form solution for queueing networks. Frequently, the coefficient of variation of the distributions encountered is well in excess of one, which would be its value for the exponential. For closed queueing networks with non-exponential servers there is no known general exact solution, and most, if not all, approximation methods attempt to account for the general service time distributions through their first two moments.

We consider two simple closed queueing networks which we solve exactly using semi-numerical methods. These networks depart from the structure leading to a product-form solution only to the extent that the service time at a single node is non-exponential. We show that not only the coefficients of variation but also higher-order distributional properties can have an important effect on such customary steady-state performance measures as the mean number of customers at a resource or the resource utilization level in a closed network.

Additionally, we examine the state that a request finds upon its arrival at a server, which is directly tied to the resulting quality of service. Although the well-known Arrival Theorem holds exactly only for product-form networks of queues, some approximation methods assume that it can be applied to a reasonable degree also in other closed queueing networks. We investigate the validity of this assumption in the two closed queueing models considered. Our results show that, even in the case when there is a single non-exponential server in the network, the state found upon arrival may be highly sensitive to higher-order properties of the service time distribution, beyond its mean and coefficient of variation.

This dependence of mean numbers of customers at a server on higher-order distributional properties is in stark contrast with the situation in the familiar open *M/G/1* queue. Thus, our results put into question virtually all traditional approximate solutions, which concentrate on the first two moments of service time distributions.

Keywords

Closed queueing networks; Non-exponential service time; High variability; Higher-order moments; Arrival Theorem; Approximation methods

1. INTRODUCTION

The objective of this paper is to show that there can be big discrepancies between exact results and traditional approximations due to the influence of distributional properties of inter-arrival and service times on the performance of queueing networks. Here, we consider two very simple closed queueing networks which deviate from the product form only in that a single node is non-exponential. We examine customary steady-state performance metrics (mean number of requests at a server, server utilization), as well as the degree of departure from the Arrival Theorem.

Since in many real-life situations the service and/or inter-arrival times tend to exhibit high variability (e.g. due to the use of caching, or intrinsic nature of certain types of Internet traffic [7]), we focus on the case where the coefficient of variation of the service time exceeds one. Using a recently-developed semi-numerical solution method [8] and its generalization [9], we show that the state found upon arrival and customary steady-state performance metrics may exhibit important dependence on higher-order properties of the service time distribution. Such dependence casts a doubt over the value of approximations traditionally limited to the first two moments of the distribution.

In a large number of real-life computer and networking applications, the state of a resource that a request finds upon its arrival at the resource greatly impacts the resulting quality of service. To some extent, what an arriving request "sees" may be viewed as more important than the customary steady-state performance metrics such as the mean number of requests or server utilization. As an example, from the standpoint of an I/O request generated by the host the probability that the requests find a free I/O path is a more critical performance measure than the overall path utilization.

The Arrival Theorem [29, 20, 30] for closed product-from networks states that the state found upon arrival is the same as the steady-state of the network without the arriving request. This theorem is at the heart of the Mean Value Analysis of queueing networks [24, 26, 27, 31]. The elegant simplicity of the Arrival Theorem makes it an attractive basis for approximations even when the network does not posses a product-form solution [5, 25, 1, 18, 2, 12, 32, 10, 15, 17, 36].

To the best of our knowledge, there is a limited number of studies attempting to quantify the degree of applicability of the Arrival Theorem in networks with non-exponential service times [6, 4, 15, 11]. Possibly inspired by the distributional dependence factor in the Pollaczek-Khintchine formula [19, 3] for the *M/G/1* queue, most existing studies seem to concentrate on the influence of the coefficient of variation of the service time distribution [6, 4, 15]. This appears to be the case as well in some attempts to improve the approximation given by the "raw" Arrival Theorem by introducing corrective terms related to the first two moments of the service time distribution [25, 1, 17, 36]. The influence of properties of order higher than two (such as skewness and kurtosis) of the service time distribution seems to have attracted little attention [37, 35, 6].

Our contribution is threefold. First, we show that, even for a very simple closed network with just a single non-exponential server, the performance of the system may depend in an important way on higher-order properties, beyond the first two moments, of the service time distribution. This provides evidence that many traditional approximations for non-exponential closed queueing networks (e.g. [22, 23, 28, 33, 38, 17]) need to be re-evaluated. Second, we examine the degree of applicability of the Arrival Theorem as a function of both the distribution of the service times and the number of users in the system. Our results provide some indication when the theorem can be expected to be a reasonable approximation, and when the deviation from it can be almost arbitrarily large. Third, we show that the influence of higher order properties is not limited to the skewness of the service time distribution but includes properties of even higher order.

This paper is organized as follows. In Section 2 we describe the first queueing network considered and we present our numerical results for this system. Section 3 is devoted to the second simple network and its numerical results. Section 4 concludes this paper.

2. SIMPLE TWO-NODE NETWORK

We first consider the two-node closed queueing network represented in Figure 1. This network consists of a multi-server queue with s servers and a single-server queue, referred to as Nodes 1 and 2, respectively.

We start by examining the effect of a single non-exponential server in a two-node network where the other server has exponentially distributed service times. In such a simple network one might think that the presence of an exponential server would make it close to an *M/G/1* queue where only the first two moments of the non-exponential server matter for the computation of the mean number of users. This case is studied in Section 2.1. Given the current tendency to use multiple processors in many real-life applications, it seems important to examine our network with multiple non-exponential servers at one of its nodes. Section 2.2 is devoted to such a case. In Section 2.3 we consider a similar network but with multiple exponential servers at one of its nodes and a single non-exponential server at the other. This allows us to see how much the performance of a memoryless multi-server is affected by high-variability of service at the queue feeding the former.

We denote by N the total number of users (tokens, customers, requests) in this network. The service time at the multi-server node is represented by a two-stage Coxian distribution [13], with mean $1/\mu_1$. We denote by μ_{1a} and μ_{1b} the respective service rates of the stages of this distribution, and by $\hat{q}_{1a} = 1 - q_{1a}$ the probability of moving from stage a to stage b. To represent an exponential distribution, it sufficies to set $\hat{q}_{1a} = 0$. The service time at the single-server node is also represented by a two-stage Coxian distribution with mean $1/\mu_2$. For the latter distribution, we denote by μ_{2a} and μ_{2b} the respective service rates of this distribution, and by q_{2a} the probability to complete the service process following the first stage. $\hat{q}_{2a} = 1 - q_{2a}$ denotes the probability that the customer arrival process proceeds to the second stage upon completion of the first stage. We denote by cv_1 and cv_2 the coefficients of variation at Nodes 1 and 2, respectively.

We denote by $\overline{n}_1(N)$ the steady-state mean number of requests at Node 1 in a network with a total of N users, and by $\overline{n}_1^A(N)$ the mean number of users found at Node 1 by a request arriving from Node 2. If the Arrival Theorem was to apply in our network, the state of Node 1 found by a request leaving Node 2 would correspond to the steady state of Node 1 in a network with the same service time distributions but $N-1$ users. Therefore, we measure the deviation from the Arrival Theorem by the quantity $\Delta(N) = \left|1 - \overline{n}_1(N-1)/\overline{n}_1^A(N)\right|$, expressed in percent.

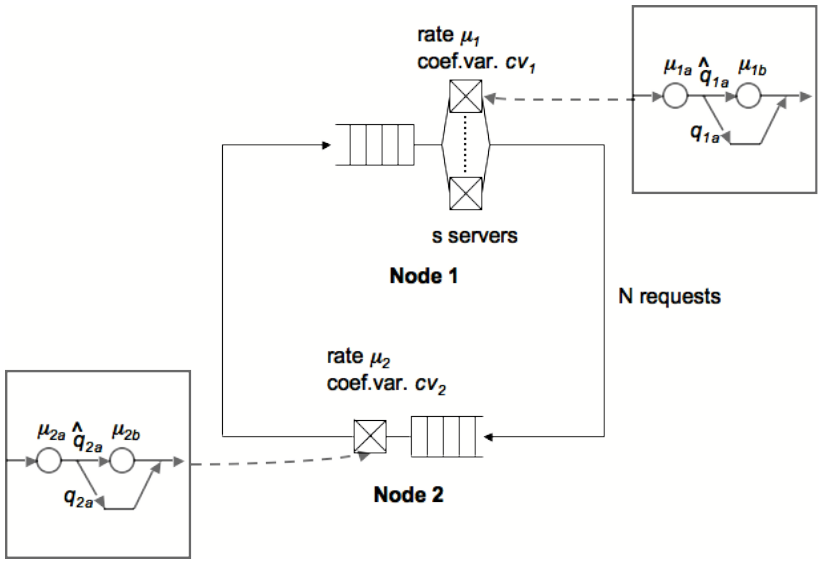

Figure 1. Simple two-node network

Additional quantities studied in this paper include the server utilization level for Node 1 defined as $U_1(N) = \overline{m}_1(N)/s$, where $\overline{m}_1(N)$ is the steady-state expected number of busy servers at Node 1 in a network with a total of N customers.

We use a generalization to $C_k/C_2/c$-type queues [9] of a recently published semi-numerical solution method for $M/C_k/1$-type queues [8] to obtain the above quantities. This method yields the steady-state probability $p(j,l_2,n)$ that the system is in the state described by (j,l_2,n) where n is the current total number of customers at Node 1, l_2 is the number of Node 1 customers in the second stage of their Coxian service time, and j is the current service stage at Node 2. The solution methods used rely on a Markovian model with standard balance equations. For single-server queues the method [8] requires no iteration and is thus exact. The solution for multi-server [9] queues requires a fixed-point iteration, and we used the convergence criterion of relative difference of less than 10^{-9} between consecutive iterates.

Let $P_N^A(n)$ be the steady-state probability that a request arriving from Node 2 finds n users at Node 1 in a network with a total of N customers. As we show in the Appendix, we have in general

$$P_N^A(n) = \frac{\sum_{j=1}^{k} \mu_{2j} q_{2j} \sum_{l_2=0}^{\min(c,n)} p(j,l_2,n)}{\sum_{l<N} \sum_{j=1}^{k} \mu_{2j} q_{2j} \sum_{l_2=0}^{\min(c,l)} p(j,l_2,l)}$$

for $n = 0,1...,N-1$.

$\overline{n}_1^A(N)$ is then expressed as $\sum_{n=0}^{N-1} n P_N^A(n)$. In the particular case when the service time at Node 2 is exponentially distributed, we get for the probability upon arrival

$$P_N^A(n) = \frac{\sum_{l_2=0}^{\min(c,n)} p(l_2,n)}{\sum_{l<N} \sum_{l_2=0}^{\min(c,l)} p(l_2,l)} \quad \text{for } n = 0,1...,N-1.$$

The set of Cox-2 distributions used throughout our paper is described in Table 1.

2.1 Single non-exponential server (Cox-2) at Node 1 and exponential server at Node 2

We start our study by assuming that the service time at Node 2 is exponentially distributed and that there is only one non-exponential server at Node 1 ($s = 1$). Thus, in a sense, Node 1 may be viewed as an $M/G/1$-like queue in a closed network.

Figure 2a shows the deviation (expressed in percent) from the Arrival Theorem in a network with $N = 10$ users, for varying server utilization levels and several Cox-2 distributions at Node 1 with a mean of 1 and coefficient of variation of 2, 4, 6, 8 and 10, respectively. The corresponding actual values of $\overline{n}_1^A(N)$ are represented in Figure 2b. The parameter values for the Cox-2 distributions used in our

examples are given in Table 1. The distributions are identified in our graphs by their index in Table 1 and by their coefficient of variation denoted by cv in our figures.

Table 1.

Distribution Index	Mean value	cv_i	Skewness	Kurtosis	μ_{ia}	μ_{ib}	q_{ia}
D1	1.0	2.0	19.26	608.91	1.11	6.25E-02	9.938E-01
D2	1.0	2.0	3.07	12.77	1000.0	4.00E-01	6.010E-01
D3	1.0	4.0	54.10	4107.3	1.11	1.32E-02	9.987E-01
D4	1.0	4.0	6.01	48.28	1000.0	1.18E-01	8.830E-01
D5	1.0	6.0	86.00	10087.28	1.11	5.68E-03	9.994E-01
D6	1.0	6.0	9.01	108.30	1000.0	5.40E-02	9.460E-01
D7	1.0	8.0	116.99	18480.19	1.11	3.17E-03	9.997E-01
D8	1.0	8.0	12.01	192.43	1000.0	3.10E-02	9.690E-01
D9	1.0	10.0	147.58	29276.89	1.11	2.02E-03	9.998E-01
D10	1.0	10.0	15.02	300.63	1000.0	1.98E-02	9.802E-01
D11	0.67	6.0	86.04	10097.03	1.67	8.52E-03	9.994E-01
D12	0.67	6.0	9.01	111.30	1500.0	8.10E-02	9.461E-01

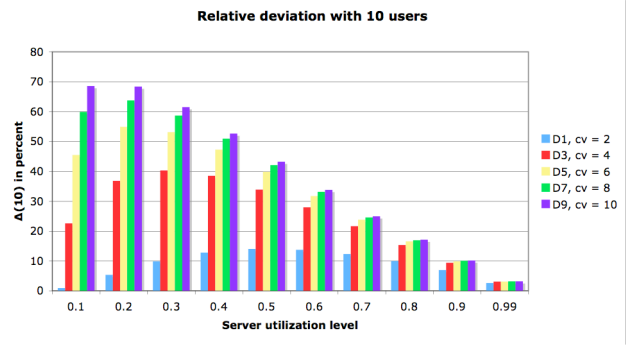

Figure 2a. Relative deviation from the Arrival Theorem for a subset of distributions from Table 1 used for the service time at Node 1 with $N = 10$

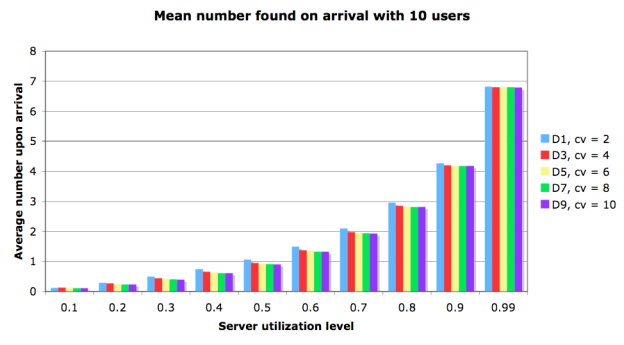

Figure 2b. Mean number found on arrival for the same subset of distributions from Table 1 used for the service time at Node 1 with $N = 10$

We observe that, for the distributional parameters considered, the deviation from the Arrival Theorem seems to depend on both the server utilization level and the coefficient of variation of the service time distribution at Node 1. In this particular case, the deviation ranges from some 10% to around 70% and tends to peak for relatively small server utilization levels. These observations seem to confirm the conclusions of previous research [15]. We also note that, for this particular example, the expected number of users found by an arrival, $\bar{n}_1^A(N)$, does not depend much on cv_1, the coefficient of variation of the service time distribution. To properly interpret these results it is important to note that the value of the service rate at Node 2 has been adjusted for each value of cv_1 so as to maintain the specified utilization levels.

In reality, things appear more complicated than implied by previous research. In Figures 3a and 3b, we have represented analogous results using a different set of Cox-2 distributions with the same mean and coefficients of variation as in Figures 2a and 2b but different higher order properties. The parameter values for the Cox-2 distributions used in this example correspond to another subset of distributions given in Table 1.

Quite unlike what we saw before, the deviation from the Arrival Theorem in this example can exceed 800% and appears to increase as the server utilization level increases. Additionally, Figure 3b shows that, in this particular case, the expected number of users found by an arrival $\bar{n}_1^A(N)$ varies significantly as cv_1, the coefficient of variation of the service time distribution at Node 1, changes. Notice that here large deviations from the Arrival Theorem occur for relatively large values of the mean number of users found upon arrival. Thus such deviations cannot be viewed as large relative errors limited to small mean numbers of users. Again, note that the value of the service rate at Node 2 has been adjusted for each value of cv_1 so as to maintain the specified utilization levels.

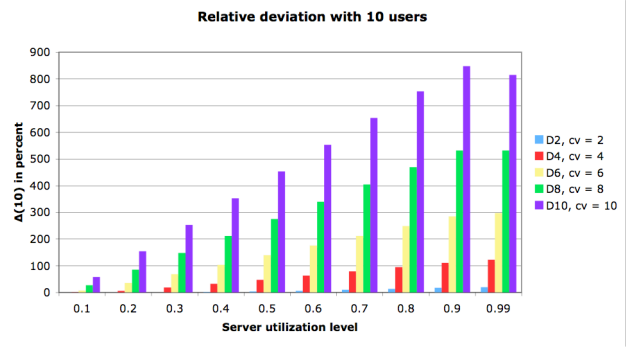

Figure 3a. Relative deviation from the Arrival Theorem for another subset of distributions from Table 1 used for the service time at Node 1 with $N = 10$

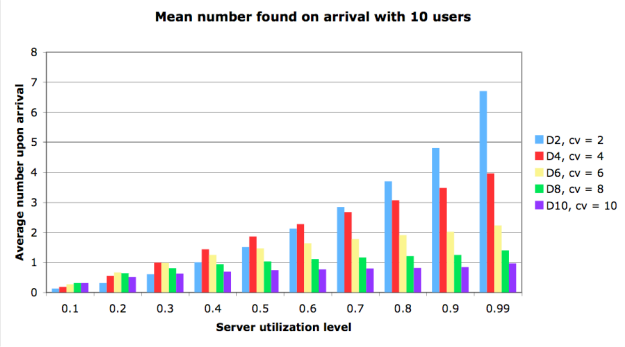

Figure 3b. Mean number found on arrival for the same subset of distributions from Table 1 used for the service time at Node 1 with $N = 10$

Comparing the results of these two examples (Figures 2b and 3b), it is clear that properties of order higher than two (i.e., beyond the mean and the coefficient of variation) of the service time distribution may have a dramatic effect on the mean number of users found upon arrival. As it turns out, higher-order distributional properties influence not only the expected state found upon arrival but also such customary steady-state performance measures as the mean number of customers at a node or the node utilization level.

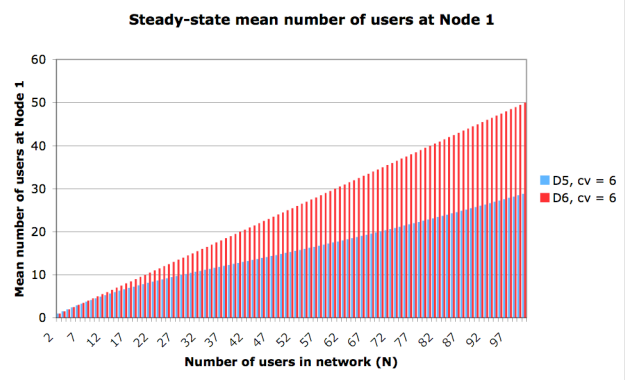

Figure 4a. Influence of higher-order moments of the service time distribution at Node 1 on the mean number of users $\overline{n}_1(N)$

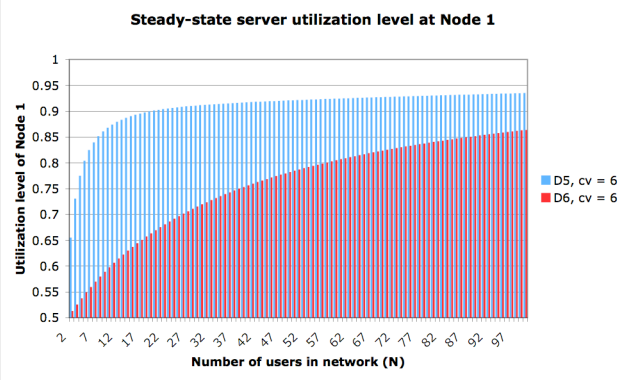

Figure 4b. Influence of higher-order moments of the service time distribution at Node 1 on the server utilization level $U_1(N)$

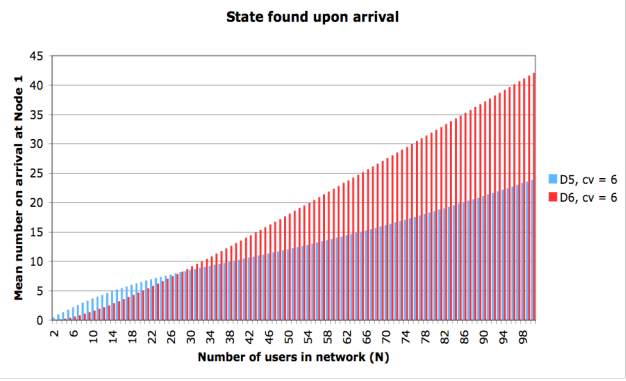

Figure 4c. Influence of higher-order moments of the service time distribution at Node 1 on the mean number of users found by an arriving request $\overline{n}_1^A(N)$

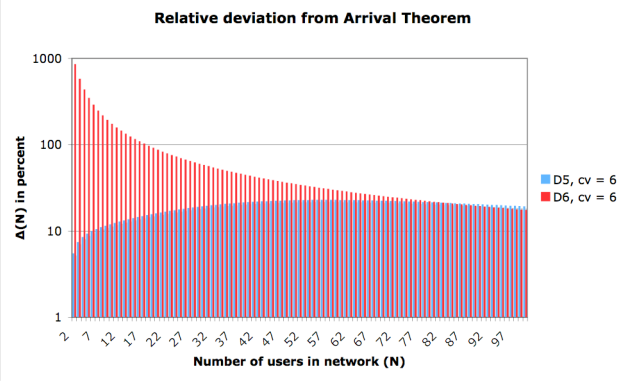

Figure 4d. Relative deviation from the Arrival Theorem $\Delta(N)$ for two distributions of the service time at Node 1 with the same first two moments (on logarithmic scale)

Thus, Figure 4a displays the mean number of users at Node 1, $\overline{n}_1(N)$, as a function of the total number of users in the network, N, for the two sets of parameters given in Table 1 for $cv_1 = 6$. Figure 4b shows the corresponding server utilization levels $U_1(N)$. Figures 4c and 4d display the values of $\overline{n}_1^A(N)$ and of $\Delta(N)$, respectively, to illustrate how these quantities vary with N. Recall that both Cox-2

distributions labeled D5 and D6 have the same mean and coefficient of variation but different higher-order properties. The mean service time at Node 2 is $1/\mu_2 = 1$.

We notice in Figure 4a the important effect higher-order properties have on $\bar{n}_1(N)$, the mean number of customers at Node 1. This is quite unlike what one would expect in an open *M/G/1* queue where only the first two moments of the service time distribution would matter.

Similar effects of the higher-order properties of the service distribution can be observed for the server utilization levels and the mean number of users found by an arriving request in Figures 4b and 4c, respectively. As illustrated in Figure 4d, the values of $\Delta(N)$, the deviation from the Arrival Theorem, also differ significantly for the two distribution types considered. In general, although the deviation from the Arrival Theorem decreases as N increases, it remains non-negligible even for higher numbers of users in the network (note the logarithmic y-axis scale in Figure 4d). The amplitude of the deviation from the Arrival Theorem varies with network parameters, and, as an example, is close to 50% for $N = 100$ when the mean service time at Node 2 is $1/\mu_2 = 0.5$.

With respect to the state "seen" by an arriving request, so far we have considered the mean number of users found upon arrival at Node 1, and we found that it may depend on higher-order properties of the service time distribution. Figure 5 illustrates the actual probability distribution of the number of users found by an arriving request at Node 1, for the two Cox-2 distributions given in Table 1 for $cv_1 = 4$. As before, service at Node 2 is exponentially distributed. The value of the mean service time for Node 2 is $1/\mu_2 = 1/3$ in this example. Recall that we denote by $p_N^A(n)$ the probability that an arrival finds n requests at Node 1, where N is the total number of users in the network, and we have $n = 0,1,..,N-1$. The results in Figure 5 have been obtained for a total of ten requests in the network ($N = 10$).

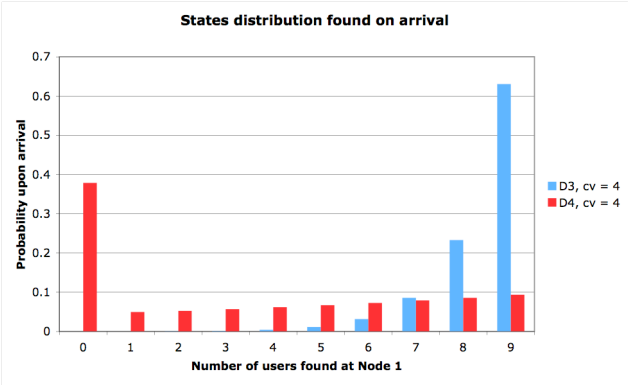

Figure 5. Influence of higher-order moments of the service time distribution at Node 1 on $p_N^A(N)$

We observe how strikingly different the distributions $p_N^A(n)$ can be for the same mean and coefficient of variation of the service time at Node 1. In particular, if one considers $p_N^A(0)$, the probability that the arriving request does not have to wait before service, it varies from close to zero in one case to almost 40% in the other.

2.2 Multiple non-exponential (Cox-2) servers at Node 1 and exponential server at Node 2

We now turn our attention to the case where there are several servers at Node 1. As before, the service time at Node 2 is exponentially distributed. Figure 6a illustrates the deviation from the Arrival Theorem as a function of the number of users in the network for 2, 4, and 8 servers. We use the set of parameter values corresponding to distribution labeled D4 in Table 1 ($cv_1 = 4$) for the service time distribution at Node 1. The mean service time at Node 2 is $1/\mu_2 = 1$. Here we observe that $\Delta(N)$ peaks for lower values of N and then decreases as the number of users increases. In our example, with two servers at Node 1 ($s = 2$) the deviation reaches almost 140% while the decrease with N tends to be slow, so that the deviation exceeds 30% with 30 users in the network. Perhaps not surprisingly, the deviation decreases as the number of servers increases, but, even with 8 servers, it can exceed 40%.

In Figure 6b we examine the deviation from the Arrival Theorem for a fixed population level ($N = 16$) as a function of the server utilization level for different values of the number of servers s. The same Cox-2 distribution labeled D4 in Table 1 is used as the service time distribution for the non-exponential servers at Node 1. Note that the value of the service rate at Node 2 has been adjusted for each value of the number of servers s so as to maintain the specified server utilization levels.

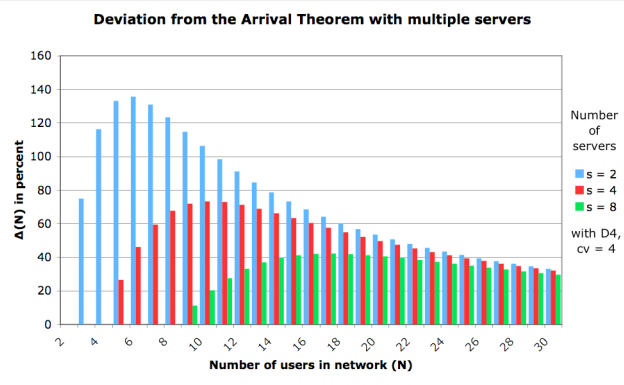

Figure 6a. Relative deviation from the Arrival Theorem $\Delta(N)$ as a function of the number of users in the network N for varying number of servers s at Node 1

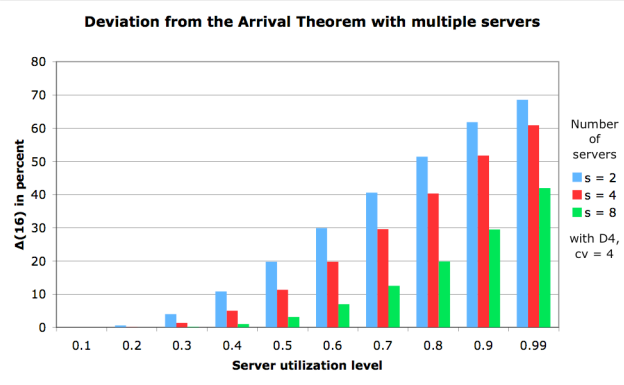

Figure 6b. Relative deviation from the Arrival Theorem $\Delta(16)$ as a function of the server utilization level for varying number of servers s at Node 1

From the results shown in Figures 6a and 6b, it is apparent that for queues with multiple non-exponential servers, just like in the case of a single server, one has to approach with caution approximations based on the Arrival Theorem.

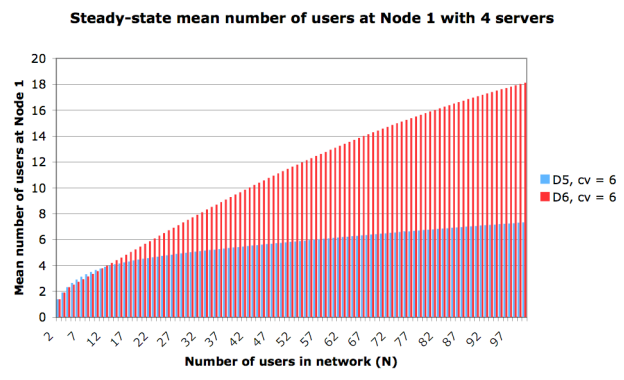

Figure 7a. Influence of higher-order moments of the service time distribution at Node 1 on the mean number of users $\overline{n}_1(N)$

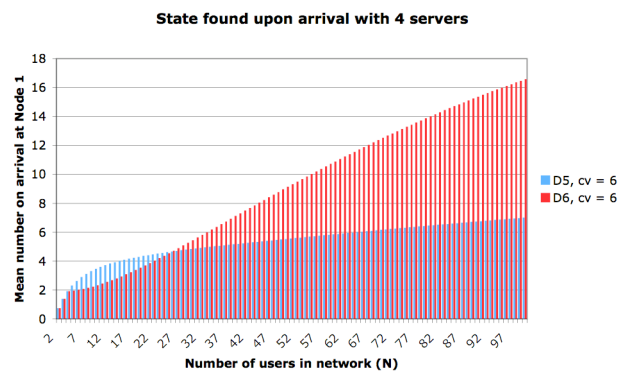

Figure 7b. Influence of higher-order moments of the service time distribution at Node 1 on the mean number of users found by an arriving request $\overline{n}_1^A(N)$

The dependence on higher-order properties of the service time distribution in the open *M/G/c* queue has been suspected and later shown empirically by some authors, e.g. [37, 34, 21, 35, 16]. Hence, it may not be surprising that this type of dependence is also present in our network. Figure 7a shows the steady-state mean number of users at Node 1 with 4 servers ($s = 4$) for the two distributional parameters given in Table 1 in the case when the coefficient of variation of the service time is 6 ($cv_1 = 6$), labeled D5 and D6. The exponential service time distribution at Node 2 has a mean of $1/\mu_2 = 0.35$. We observe that the values for $\overline{n}_1(N)$ differ by close to 100% (or 50%, depending on how you look at it) as N exceeds 50 users in the system. The corresponding values of $\overline{n}_1^A(N)$, the mean number found upon arrival, are shown in Figure 7b. It is interesting to note that, depending on the number of users in the network, one or the other of the distribution types can lead to a larger value of $\overline{n}_1^A(N)$. Again, we observe the important influence of higher-order properties of the service time distribution. We also observe that this influence varies with the number of users in the network and persists as the latter increases. This persistence is consistent with distributional dependencies in *M/G/c* queues [16, 34]. Results not reported in this paper indicate that, for a given server utilization level, distributional effects tend to decrease for larger numbers of servers.

2.3 Multiple exponential servers at Node 1 and single non-exponential (Cox-2) server at Node 2

In Sections 2.1 and 2.2 we looked at the effect of higher-order properties of the service time on the non-exponential server. It is interesting to examine how a single non-exponential server with high service time variability affects the performance of a multi-server node with memoryless service. Hence, we consider here the case where the service time at Node 1 is exponentially distributed and Node 2 has a general service time distribution. As illustrated in Figure 8, here again we observe significant dependence on higher-order properties of the non-exponential service time distribution. Figure 8 shows the values of $\overline{n}_1^A(N)$ for $s = 2$ servers at Node 1, mean service time at this node $1/\mu_1 = 1$, and values $1/\mu_2 = 0.67$ and $cv_2 = 6$ for the Cox-2 distribution of service time at Node 2. Distributions labeled D11 and D12 in this figure refer to parameter values from Table 1. We note that the influence of higher-order properties persists as the number of users in the network increases. This is consistent with similar distributional dependencies in *G/M/c* queues [34].

Not surprisingly, additional results, not shown in this paper, indicate the importance of higher-order distributional properties for performance metrics not displayed in our figure, as well as in the case where both Nodes 1 and 2 are non-exponential.

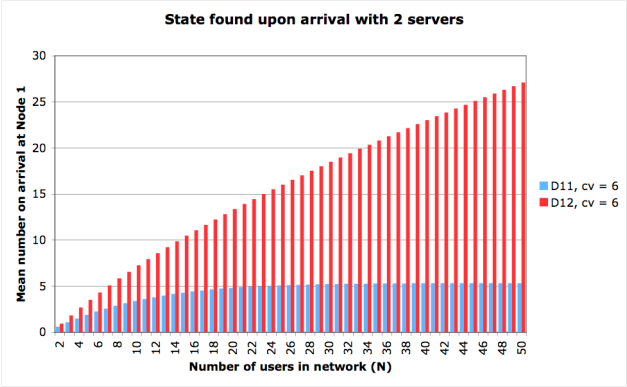

Figure 8. Influence of higher-order moments of the service time distribution at Node 2 on the mean number of users found by an arriving request at Node 1 $\bar{n}_1^A(N)$

3. MACHINE REPAIR MODEL

The second model considered in this study is the machine repairmen model shown in Figure 9. Here we have a total of N request sources or users, the time spent at a source ("machine up time") is exponentially distributed, and there are s servers ("repairmen") at Node 1. As before, the service time at Node 1 ("machine repair time") has a Cox-2 distribution with the same notations as in Figure 1. We denote by $1/\lambda$ the mean time a request remains at a source ("mean machine up time").

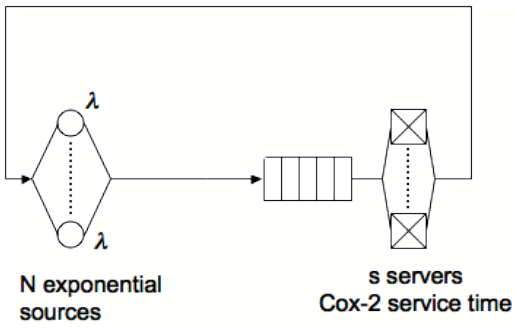

Figure 9. Machine repair model with multiple servers

Such a model corresponds in particular to a set of users with exponentially distributed idle times and a multiple server resource. We consider this system in the case when the service time distribution at the shared resource has a coefficient of variation greater than one. As noted in the introduction, higher coefficients of variation may be encountered in many systems, including in the presence of caching in I/O subsystems or Web servers. On the surface of things, one might think that the memoryless sources might act as "buffers" to dampen distributional effects at the shared multi-server resource.

We study the behavior of this model with $s = 4$ servers at Node 1. The mean service time $1/\mu_1$ is set to 1 and the number of sources in the network at $N = 20$.

In Figures 10a and 10b we show the deviation from the arrival theorem for the same two subsets of distributions of Tables 1 used in Figures 2a and 3a, respectively. Note that in these figures, analogously to what we did for the network considered in Section 2.1, we adjust the value of $1/\lambda$ ("the mean machine up time") for each distribution so as to maintain the specified server utilization levels.

While the results shown in Figure 10a may give the impression that the Arrival Theorem works very well for this network, Figure 10b shows that for a different set of distributions with the same first two moments deviations may exceed 30%. It is thus clear from the results in Figures 10a and 10b that the deviations from the Arrival Theorem may be more or less significant depending on the specific distribution of the service time at the non-exponential servers and on the server utilization level. With respect to Figure 10a, it is interesting to note that the ratio $\bar{n}_1(N-1)/\bar{n}_1^A(N)$ used to measure the deviation from the Arrival Theorem starts out greater than one and then becomes less than one as the server utilization level increases. This accounts for the dip in the amplitude of the deviation seen in Figure 10a. We conclude that higher-order properties of the service time distribution, beyond its mean and coefficient of variation, may significantly influence this deviation. When interpreting the results of Figures 10a and 10b one should keep in mind that the rate λ of the exponential sources has been adjusted for each distribution so as to maintain the specified server utilization level.

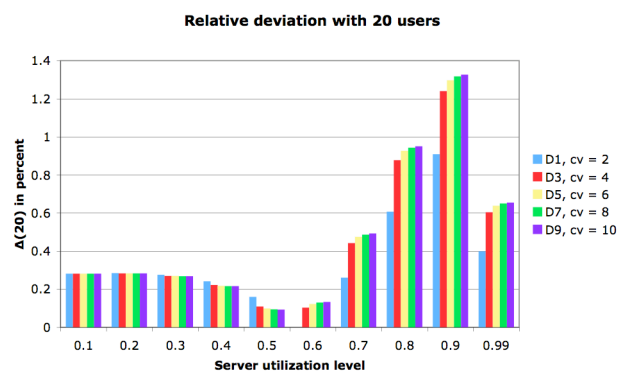

Figure 10a. Relative deviation from the Arrival Theorem for a subset of distributions from Table 1 used for the service time at Node 1 with $N = 20$

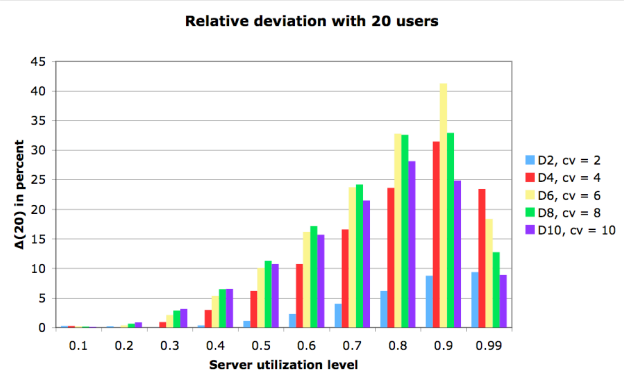

Figure 10b. Relative deviation from the Arrival Theorem for another subset of distributions from Table 1 used for the service time at Node 1 with $N = 20$

As a final example, we now fix the value of the "mean machine up time" at $1/\lambda = 0.2$ and we examine the influence of higher-order distributional properties as the number of users increases. We show in Figures 11a and 11b the values of $\Delta(N)$ and $\overline{n}_1^A(N)$, respectively, for the two distribution types of Table 1 with $cv_1 = 6$, labeled D5 and D6.

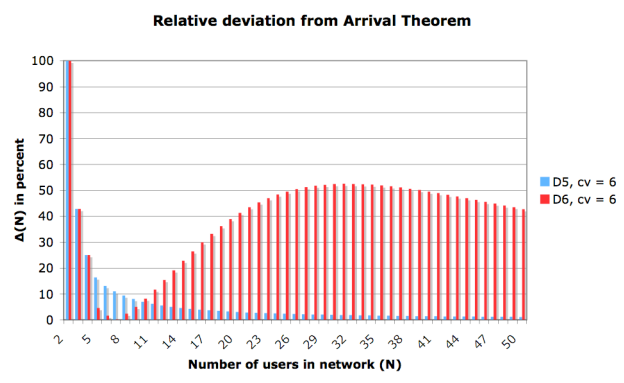

Figure 11a. Relative deviation from the Arrival Theorem $\Delta(N)$ for two distributions of the service time at Node 1 with the same first two moments in a model with 4 repairmen

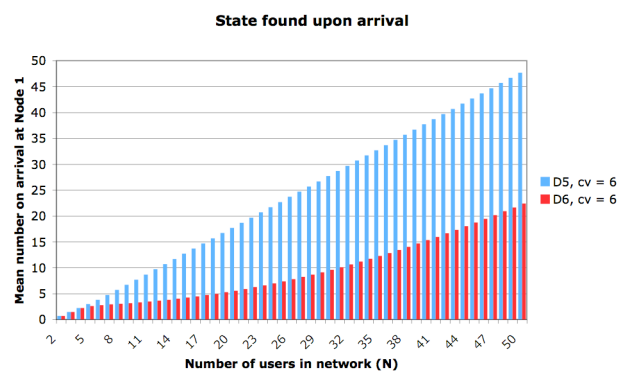

Figure 11b. Influence of higher-order moments of the service time distribution at Node 1 on the mean number found on arrival $\overline{n}_1^A(N)$ with 4 repairmen

It has been our experience that in general the amplitude of the departure from the Arrival Theorem, $\Delta(N)$, tends to increase as the variability of the service time increases, however, the form of its evolution with N clearly depends on higher-order properties of the service time distribution. Similarly, higher-order properties have an important effect on the number of users found upon arrival (Figure 11b), as well as the customary mean number of users at Node 1. Note that the effect of higher-order distributional properties remains significant (especially for $\overline{n}_1^A(N)$) as the number of users in the system N increases.

Overall, it is clear just from the two simple models considered in this paper that the performance of closed queueing networks with non-exponential servers is sensitive not only to the first two moments of the service time distributions but also, and in many cases to a large degree, to higher-order properties of those distributions. In general, the larger the service time variability the more important higher-order properties appear to be. The results presented so far, given the particular form of the distribution used, viz. Cox-2, do not allow us to determine the individual importance of the skewness and kurtosis. We briefly address this question in the Appendix (6.3).

4. CONCLUSION

In this paper we have examined the influence of the service time distributions on the mean number of users at each node and the server utilization levels, as well as the degree of departure from the Arrival Theorem for two simple closed queueing networks. These networks depart only minimally from the product-form structure since only one node has non-exponential service times. Our results indicate that

higher-order properties of the service time distribution, beyond the first two moments, may have an important effect on both steady-state properties and the state found upon arrival.

The observed dependence on higher-order properties clearly shows that approximations limited to the first two moments of the service time distribution may be highly inaccurate. Our results suggest that the effect of higher-order properties of the service distribution tends to increase as the service time variability increases. The amplitude of the departure from the Arrival Theorem also tends to increase under similar conditions. As the number of users in the system increases, the deviation from the Arrival Theorem tends to decrease although it can remain significant even for larger numbers of users. The influence of higher-order distributional properties on the state found upon arrival (and on the mean number of users) persists for larger values of the number of users in the system.

Results not shown in this paper confirm that the Arrival Theorem may be a good approximation when the service time distributions have a coefficient of variation of less than one [1, 15]. However, in closed networks of queues with high-variability service time distributions, the Arrival Theorem may be totally wrong. The degree of departure from the theorem varies with the level of server utilization but there appears to be no obvious and simple relationship one can establish between the two. For some distributions the discrepancy is most obvious in the middle range of server utilization, while for others it may increase as the server utilization increases. Moreover, the ratio $\overline{n}_1(N-1)/\overline{n}_1^A(N)$ may be greater or smaller than one depending on the distribution and the server utilization or the number of users, so that no systematic bias or correction can be easily established. It appears that the relative deviation from the Arrival Theorem tends to be more important for distributions with lower skewness values.

Our results show that, unlike what happens for the open *M/G/1* queue, in a closed queueing network, the dependence on higher-order properties may be important even for the mean queue length at a node with a single server (the fact that higher order properties matter in the open *M/G/c* queue is generally known). Mean queue lengths, server utilization, and the state found upon arrival may be radically different for two service time distributions with the same mean and coefficient of variation but different higher-order moments. Since the state upon arrival has a direct influence on the quality of service, our results have clear implications for performance studies in this domain.

One could argue that in real life it is difficult to know the coefficients of variation of many service time distributions, much less higher-order properties, so that the fact that various performance metrics may exhibit strong dependence on properties of higher order can be safely overlooked. We believe, however, that it is important to realize that, given the influence of higher-order properties, traditional approximations for non-exponential queueing networks produce results that cannot be viewed as trustworthy since it is impossible to say which one of the infinitely many distributions with the given first two moments they might well correspond to. Results not reported in this paper seem to indicate that traditional approximations fare better in cases when the coefficient of variation is below unity.

An interesting unanswered question is which higher order properties (skewness, kurtosis or perhaps other properties) of the distribution are most important. This determination and a search for improved approximations is the subject of future research.

Acknowlegement

The authors wish to thank the anonymous referees for their constructive remarks on an earlier version of this paper.

5. REFERENCES

[1] Akyildiz, I. F. 1987. Mean value analysis of closed queueing networks with Erlang service time distributions. Computing 39, 3 (Dec. 1987), 219-232.

[2] Akyildiz, I. F. 1988. Mean Value Analysis for Blocking Queueing Networks. IEEE Transactions on Software Engineering. 14, 4 (Apr. 1988), 418-428.

[3] Allen, A. O., Probability, Statistics, and Queueing Theory with Computer Science Applications. Academic Press, 2nd edition, 1990.

[4] Balsamo, S. 2000. Closed queueing networks with finite capacity queues: approximate analysis. In Proceedings of the 14th European Simulation Multiconference on Simulation and Modelling: Enablers For A Better Quality of Life (May 2000), Ed. SCS Europe, 593-600.

[5] Bard, Y. 1979. Some Extensions to Multiclass Queueing Network Analysis. In Proceedings of the 3rd International Symposium on Modelling and Performance Evaluation of Computer Systems (Feb. 1979). PERFORMANCE 1979. Eds. North-Holland Publishing Co., Amsterdam, The Netherlands, 51-62.

[6] Bondi, A. B. and Whitt, W. 1986. The influence of service-time variability in a closed network of queues. Performance Evaluation. 6, 3 (Sep. 1986), 219-234.

[7] Borgnat, P., Dewaele, G., Fukuda, K., Abry, P. and Cho, K. 2009. Seven Years and One Day: Sketching the Evolution of Internet Traffic. To appear in Proceedings of INFOCOM 2009.

[8] Brandwajn, A., and Wang, H. 2008. A Conditional Probability Approach to M/G/1-like Queues. Performance Evaluation. 65, 5 (2008), 366-381.

[9] Brandwajn, A., and Begin, T. 2008. Preliminary Results on a Simple Approach to G/G/c-like Queues. Submitted for publication, Available upon request.

[10] Buitenhek, R., van Houtum, G. J., and Zijm, W. H. M. 2000. AMVA based solution procedures for open queueing networks with a population constraint. Annals of Operations Research, 93 (Mar. 2000), 15-40.

[11] Casale, G., Mi, N., and Smirni, E. 2008. Bound analysis of closed queueing networks with workload burstiness. In Proceedings of the 2008 ACM SIGMETRICS international Conference on Measurement and Modeling of Computer Systems (Annapolis, MD, USA, Jun. 2008). SIGMETRICS '08. ACM. New York, NY, 13-24.

[12] Clò, C. 1998. MVA for product-form cyclic queueing networks with blocking. Operations Research. 79, (1998), 83-96.

[13] Cox, D. R., and Smith, W. L., Queues. John Wiley & Sons, New York, 1961.

[14] Cooper, R., Introduction to Queueing Theory. Second Edition, North Holland, New York, 1984.

[15] Eager, D. L., Sorin, D. J., and Vernon, M. K. 2000. AMVA techniques for high service time variability. In Proceedings of the 2000 ACM SIGMETRICS international Conference on Measurement and Modeling of Computer Systems (Santa Clara, CA, USA, Jun. 2000). SIGMETRICS '00. ACM, New York, NY, 217-228.

[16] Gupta, V., Harchol-Balter, M., Dai, J., and Zwart, B. 2007. The Effect of Higher Moments of Job Size Distribution on the Performance of an M/G/s Queueing System. Perfomance Evaluation Review, 35, 2 (Sep. 2007), 12-14.

[17] Halachmi, I., Adan, I. J. B. F., van der Wal, J., Heesterbeek, J. A. P., and van Beek, P. 2000. The design of robotic dairy barns using closed queueing networks. European Journal of Operational Research, Elsevier. 124, 3 (Aug. 2000), 437-446.

[18] Kaufman, J. S., and Wong, W. S. 1994. Approximate analysis of a Gordon-Newell like non-product-form queueing network. Operations Research. 48, 3 (Jun. 1994), 249-271.

[19] Kleinrock, L., Queueing Systems, Volume 1: Theory. John Wiley & Sons, New York, 1975.

[20] Lavenberg, S. S., and Reiser, M. 1980. Stationary State Probabilities at Arrival Instants for Closed Queueing Networks with Multiple Types of Customers. Journal of Applied Probability. 17, 4 (Dec. 1980), 1048-1061.

[21] Ma, B., N. W., and Mark, J. W. 1995. Approximation of the Mean Queue Length of an M/G/c Queueing System. Operations Research, 43, 1, Special Issue on Telecommunications Systems: Modeling, Analysis and Design. (Jan. - Feb. 1995), 158-165.

[22] Marie, R.A. 1979. An Approximate Analytical Method for General Queueing Networks. IEEE Transactions on Software Engineering. 5, 5 (Sept. 1979), 530-538.

[23] Onvural, R. O. 1990. Survey of closed queueing networks with blocking. ACM Computing Surveys. 22, 2 (Jun. 1990), 83-121.

[24] Reiser, M. 1979. Mean Value Analysis for Queueing Networks - A New Look at an Old Problem. Proceedings of the 3rd International Symposium on Modelling and Performance Evaluation of Computer Systems (Feb. 1979). PERFORMANCE 1979. Eds. North-Holland Publishing Co., Amsterdam, The Netherlands, 63-77.

[25] Reiser, M. 1979. A Queueing Network Analysis of Computer Communication Networks with Window Flow Control. IEEE Transactions on Communications. 27, 8, (Aug. 1979), 1199-1209.

[26] Reiser, M., and Lavenberg, S. S. 1980. Mean value analysis of closed multi-chain queueing networks. J. Ass. Comp. Mach. 27 (Apr. 1980), 313-322.

[27] Reiser, M. 1981. Mean-Value Analysis and Convolution Method for Queue-Dependent Servers in Closed Queueing Networks. Performance Evaluation. 1, 1 (1981), 7-18.

[28] Sadre, R., Haverkort, B. R., and Reinelt, P. 2007. A Fixed-Point Algorithm for Closed Queueing Networks. In Proceedings of the 4th European Performance Engineering Workshop (Berlin, Germany, Sep. 2007). EPEW 2007. 154-170.

[29] Sevcik, K. C., and Mitrani, I. 1979. The distribution of queueing network states at input and output instants. Proceedings of the 4th Internat. Symposium on Modeling and Performance Evaluation of Computer Systems. North Holland, 1979.

[30] Sevcik, K. C., and Mitrani, I. 1981. The distribution of queuing network states at input and output instants. Journal of the ACM, 28, 2 (Apr. 1981), 358-371.

[31] Schweitzer, P. 1979. Approximate Analysis of Multiclass Closed Networks of Queues. International Conf. Stochastic Control and Optimization, 1979.

[32] Varki, E. 1999. Mean value technique for closed fork-join networks. In Proceedings of the 1999 ACM SIGMETRICS international Conference on Measurement and Modeling of Computer Systems (Atlanta, GA, USA, May 1999). SIGMETRICS '99. ACM, New York, NY, 103-112.

[33] Walstra, R. J. 1985. Nonexponential networks of queues: a maximum entropy analysis. In Proceedings of the 1999 ACM SIGMETRICS international Conference on Measurement and Modeling of Computer Systems (Austin, TX, USA, May 1985). SIGMETRICS '85. ACM, New York, NY, 27-37.

[34] Whitt, W. 1980. The Effect of Variability in the GI/G/s Queue. Journal of Applied Probability. 17, 4 (Dec. 1980), 1062-1071.

[35] Whitt, W. 2004. A Diffusion Approximation for the G/GI/n/m Queue. Operations Research, 52, 6 (Nov. - Dec. 2004), 922–941.

[36] Winands, E. M., Adan, I. J., and Van Houtum, G. J. 2006. Mean value analysis for polling systems. Queueing Syst. Theory Appl. 54, 1 (Sep. 2006), 35-44.

[37] Wolff, R. W. 1977. The Effect of Service Time Regularity On System Performance. Computer Performance, North Holland, (1977), 297-304.

[38] Wu, J. 1992. Maximum Entropy Analysis of Open Queueing Networks with Group Arrivals. The Journal of the Operational Research Society. 43, 11, (Nov. 1992), 1063-1078.

6. APPENDIX

6.1 Steady-state probability upon arrival at Node 1

To derive the steady-state probabilities "seen" by a request arriving from Node 2 to Node 1 in the network of Figure 1, we follow a reasoning similar to the one presented by Cooper [14]. The state of the system is described by (j,l_2,n) where n is the current total number of customers at Node 1, l_2 is the number of Node 1 customers in the second stage of their Coxian service time, and j is the current service stage at Node 2. The rate of request arrivals from Node 2 to Node 1 when this state is in effect corresponds to the rate of departures from the server at Node 2, i.e. $\mu_{2j}q_{2j}$. Hence, the rate of arrivals to Node 1 when there are n requests at this node can be expressed as

$$\sum_{j=1}^{k}\mu_{2j}q_{2j}\sum_{l_2=0}^{\min(c,n)}p(j,l_2,n).$$

The overall rate of departures from Node 2, corresponding to all possible system states, can similarly be expressed as

$$\sum_{l<N}\sum_{j=1}^{k}\mu_{2j}q_{2j}\sum_{l_2=0}^{\min(c,l)}p(j,l_2,l).$$

Thus, the probability that a request leaving Node 2 finds n requests at Node 1 is seen to be

$$P_N^A(n)=\frac{\sum_{j=1}^{k}\mu_{2j}q_{2j}\sum_{l_2=0}^{\min(c,n)}p(j,l_2,n)}{\sum_{l<N}\sum_{j=1}^{k}\mu_{2j}q_{2j}\sum_{l_2=0}^{\min(c,l)}p(j,l_2,l)}.$$

6.2 Selection of the parameters of a Cox-2 distribution given values for mean, coefficient of variation and skewness

Denote by $(m,cv,skew)$ the vector of the desired values of mean, coefficient of variation and skewness for a Coxian distribution. We describe a simple method to select the parameters of a matching Cox-2 distribution of the type represented in Figure 12a. Such a Cox-2 distribution has three parameters, i.e. μ_1, μ_2 and q_1.

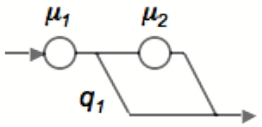

Figure 12a. A Coxian distribution with two stages

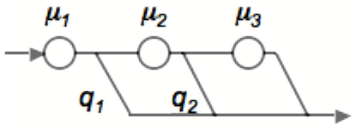

Figure 12b. A Coxian distribution with three stages

Let γ be a real-valued parameter between 0 and 1. For a given mean m and coefficient of variation cv, the parameters μ_1, μ_2 and q_1 of the Cox-2 distribution can be set as follows:

$$\mu_1 = 1/\gamma m$$

$$q_1 = 1-2(1-\gamma)^2/(cv^2+(1-\gamma)^2-\gamma^2)$$

$$\mu_2 = p_2/m(1-\gamma)\ .$$

For different feasible values of γ, the resulting Cox-2 distribution will have a different skewness value. More precisely, as γ increases, the skewness $skew$ increases as well. Thus, within a certain range, a simple bisection technique allows us to select the value of γ so that the resulting Cox-2 distribution has the desired skewness. As an example, the distributions labeled D1 and D2 in Table 1 correspond to values of γ of 0.9 and 0.001, respectively.

However, while it is possible to find a Cox-2 distribution whose mean and coefficient of variation match any given couple of values m and cv (provided the latter is greater than $1/\sqrt{2}$), the range of attainable values for the skewness is limited and depends on the value of m and cv. For this reason, having obtained a value for γ, one has to ensure that the specified value for skewness is feasible. For this, it suffices to check that the resulting value for q_1 is indeed between 0 and 1.

For the case of two Cox-3 distributions with the same first three moments considered in Section 4, our procedure is simply to choose a set of parameters (μ_1, μ_2, μ_3, q_1, q_2) (see Figure 12b), then to compute the values for its mean, coefficient of variation, skewness and its kurtosis, and finally to select a Cox-2 distribution with the same first three moments as per the previously described scheme.

6.3 Single non-exponential (Cox-3) server at Node 1 and exponential server at Node 2

The results presented in Sections 2 and 3 were obtained for Cox-2 distributions, and thus do not allow us to determine the individual importance of properties such as skewness and kurtosis. The Cox-2 distribution has three degrees of freedom, and it is therefore impossible to have two Cox-2 distributions with different kurtosis but the same first three moments. Since in the open *M/G/1* queue only the first two moments matter in the determination of the mean number in the system, a natural question is whether perhaps only the first three moments matter for a closed network or if other quantities come into the picture. In order to provide an element of answer to this question, we studied a network akin to the one considered in Section 2.1 but with a Cox-3 distribution at Node 1. As before, the service time at Node 2 is exponentially distributed. Thus we were able to create two distributions with the same mean, coefficient variation and skewness but different kurtosis (see Table 2). This network was solved using a semi-numeric recurrence method [8].

Table 2.

	First Cox-3	Second Cox-3	Relative differences
Mean	1.0		
Coeff. Var.	6.40		
Skewness	2331.54		
Kurtosis	1.44E07	7.43E06	
$\overline{n}_1(15)$	6.34	7.44	17.4%
$\overline{n}_1^A(15)$	5.28	6.93	31.1%

The values presented in Table 2 indicate that the dependence on higher-order properties is not limited to the first three moments since distinctly different results are obtained for distributions with different kurtosis and properties of yet higher order.