# Modeling service distributions in queueing network simulation

**Kimmo E. E. Raatikainen**
*University of Helsinki*, Department of Computer Science
Teollisuuskatu 23, SF-00510 Helsinki, Finland

**Abstract.** *The importance of accurate models for the input processes of simulation is generally recognized. However, the dependencies between the input processes have, at least in queueing network simulation, been paid little attention. In this paper, we consider modeling the joint distribution of the input processes required in queueing network simulation. Particularly, we propose a method for modeling the dependencies between service demands. In addition, we propose a method for dividing the service demand into service times per visit. These methods are important in simulation modeling of a computer system when accounting files provide the identification data. The experiments reported indicate that the proposed methods produce accurate estimates for quantiles of the system response-time distribution.*

**Key words:** discrete-event simulation, modeling input processes, queueing networks, joint distribution of service demands

## Introduction

The importance of accurate models for the input processes of simulation is addressed in many textbooks and several papers including Law and Kelton (1982), Iman (1986), and Kelton (1984). However, the dependencies between the input processes are not so widely considered. In this paper, we consider modeling the distributions required in queueing network simulation. We assume that the primary interest is in the simulated quantiles of the system response-time distribution. The focus is on a practical situation, on constructing a queueing network model for a computer system when accounting files provide the identification data.

In queueing network models, the service demands at different nodes are not usually independent even if the service times per visit are independent. The dependencies arise from the routing behavior. A typical accounting file provides the total cpu-time and counts of io-operations for the tasks executed. In a queueing network model, this information corresponds to the service demand at the cpu-node and to the numbers of visits at the io-nodes.

The fundamental question of modeling the distributions is if we should consider the dependencies between the service demands or if we can assume that the service demands are independent. The assumption of independence (or a specific structure of dependencies) is usual and motivated by simplity. However, it was shown in Mitchell, Paulson and Beswick (1977) that ignoring correlation between service times in a tandem queueing system can result in serious errors in the output. Furthermore, we show that the dependencies between the service demands have a significant effect upon some quantiles of the system response-time distribution. Therefore, the answer to the fundamental question is that we should take the dependencies into account.

Outside the normal and lognormal case, the multidimensional distributions considered in simulation studies have usually been bivariate distributions. Most attention is paid to distributions having exponential and Gamma marginal distributions. Such distributions are considered in Downton (1970), Paulson (1973), Ronning (1977), and Schmeiser and Lal (1982). If the marginal distributions cannot be approximated by Gamma distributions, a technique of generating a bivariate uniform distribution may be used. Since the inverse transform method ($X = F_X^{-1}(U)$) is based on the uniform distribution, any marginal distribution can be used. Bivariate uniform distributions are discussed, for example, in Mardia (1970), Kimeldorf and Sampson (1975a and 1975b), and Johnson and Tenenbein (1981). Some other techniques are considered in Schmeiser and Lal (1980).

The uniform and Gamma distributions in more than two dimensions are impractical. Therefore, we

*Manuscript of paper published in Simulation Vol. 59 # 2 (August 1992): 116–126*

need an alternative way of modeling the distributions required in queueing network simulation. Our approach of modeling the distributions consists of the following three steps:

1. modeling the marginal distributions of the service demands,
2. modeling the dependencies between service demands, and
3. decomposing the service demand into service times.

An important advantage of this approach is the great flexibility. The modeling of the marginal distributions and of the dependencies are separately considered. Particularly, any method of modeling the marginal distributions can be applied.

In this paper, we assume that the analyst has found adequate ways of modeling the marginal distributions of the service demands. In queueing network analysis, the most widely used model for the service distribution is the family of Coxian distributions. Techniques for identifying the Coxian distribution have been proposed, for example, in Bux and Herzog (1977), Lazowska (1977), and Lazowska and Addison (1979). The comparison given in Khoshgoftaar and Perros (1987) indicates that the method of moments should not be used. As an alternative to the Coxian phase-model, we may use an exploratory technique such as proposed in Hora (1983), Avramidis and Wilson (1989), and Raatikainen (1991). The number of visits is usually modeled by a geometric distribution or by a mixture of geometric distributions. Mixtures of geometric distributions are considered in Kobayashi and Reiser (1975).

The second question in modeling the distributions is how we characterize and model the dependencies. For practical reasons, there is a general tendency to describe the dependency between two random variables by a single number, by some kind of correlation coefficient. The most widely used description of dependency is the ordinary, product-moment correlation. There are, however, some problems with this correlation. First of all it is meaningful only in symmetric or slightly skew distributions, but typical distributions of computer resource usage are extremely skew. In Iman and Conover (1982) the rank-correlations are used to characterize the dependencies. The rank-correlations have natural interpretation. The rank-correlation between two random variables is the product-moment correlation between the uniform random-variables used in the inverse transforms. However, the use of rank-correlations becomes difficult outside the bivariate case. On the contrary, the multinormal distribution is the only multidimensional distribution that has straightforward and simple generation in a general multidimensional case. Therefore, we use a multidimensional normal distribution to characterize the dependencies. This characterization is artificial, but its use in generating dependent random variables is simple.

When the marginal distributions of service demands and their dependencies are modeled, the open question is how the service demand can be decomposed into the service times per visits when the number of visits is known. We propose a simple technique based on the beta-distribution.

We start this paper by introducing the queueing network models examined. In Section 3 we show that the dependencies between the service demands have a significant effect on some quantiles of the system response-time distribution. In Section 4 we propose a technique for decomposing the service demand into service times when the demand and the number of visits are known. In Section 5 we introduce a method for modeling the dependencies and report the empirical results.

## Queueing Network Models Examined

In this section we introduce the three queueing network models examined in this study. The first two models are closed models used in other simulation studies. The third model is an open one and it is based on measurement data of a real system.

The first queueing network model (Figure 1) consists of three nodes: **CPU**, **TTY**, and **IO**. The queueing disciplines are PS, IS, and FCFS, respectively. The model has a single parameter, probability of cycle completion ($\pi_0$) that determines the joint distribution of the service demands at the **CPU** and **IO** nodes. The population is 20. The service times at the **TTY** and **IO** nodes are exponentially distributed with means 30 and $0.356\pi_0$, respectively. Two different service-time (per visit) distributions are used at the **CPU**-node: an exponential with mean $1.0333\pi_0$ and a three-stage Coxian (Figure 2). Five different probabilities of cycle completion are examined: 1, 1/2, 1/4, 1/8, and 1/16. The parameters of the Coxian distributions are in Table 1. Observe that the

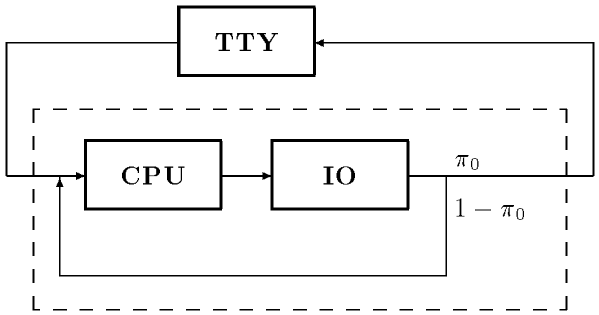



**Figure 1.** Queueing Network Model 1

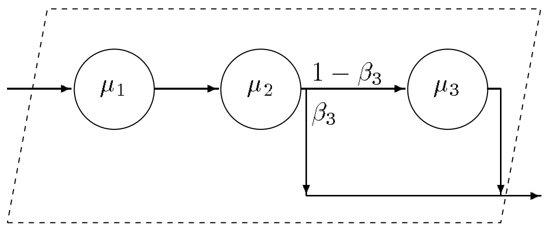



**Figure 2.** Stage Diagram of the Coxian distributions

**Table 1.** Parameters of the Coxian Distributions

| $\pi_0$ | 1 | $^1/_2$ | $^1/_4$ | $^1/_8$ | $^1/_{16}$ |
|---|---|---|---|---|---|
| $\mu_1$ | 2479.6 | 2474.6 | 2464.5 | 2444.1 | 2402.1 |
| $\mu_2$ | 5.3203 | 10.334 | 20.425 | 40.864 | 82.824 |
| $\mu_3$ | .07245 | .07475 | .07594 | .07655 | .07686 |
| $\beta_3$ | .93879 | .96864 | .98413 | .99202 | .99600 |

marginal distributions of the service demands are the same for all $\pi_0$ ($0 < \pi_0 \leq 1$).

The second queueing network model (Figure 3) is due to Heidelberger and Welch (1981). The **TTY** node is an infinite server and all the other nodes (**CPU**, **IO-1**, **IO-2**, **IO-3**, and **IO-4**) are FCFS servers. The population is 25. All the service times are exponentially distributed, and two different sets of service rates are used.

The third queueing network model (Figure 4 and Table 2) is based on actual measurement data of a VAXCluster-system consisting of one VAX 8650 computer and one HSC50 storage controller with two

| | Service Rates | |
|---|---|---|
| **Case** | **A** | **B** |
| TTY | .01 | .01 |
| CPU | 1.0 | 1.0 |
| IO-1,2 | .72 | .18 |
| IO-3,4 | .08 | .04 |

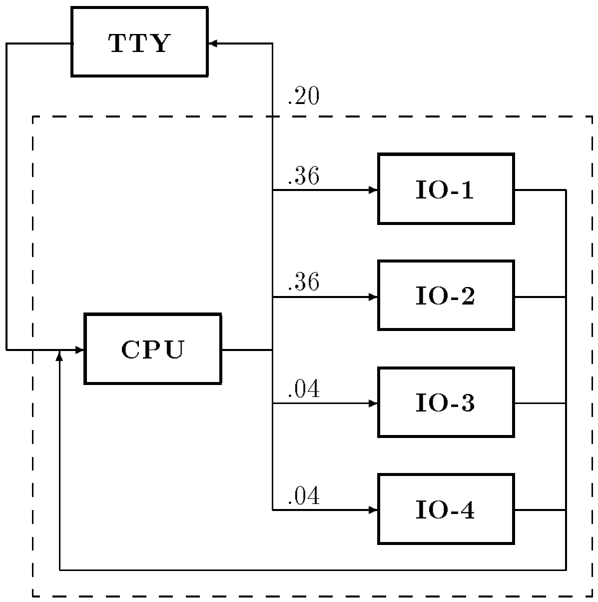



**Figure 3.** Queueing Network Model 2

disk-channels and seven disk-units. The **CPU**-node is a PS-server while the IO-Subsystem is modeled as a single server (**IO**-node) having random order queueing discipline. During one week we measured the interarrival times, total cpu-times, and direct io-operation counts of all the interactive tasks executed daily between 10 a.m. and 3 p.m. The distributions examined are discrete and defined as the measurement data. In other words, each measured value has an equal probability of being sampled.

The service times at the **IO**-node ($S_{io}$) are assumed to have the following distribution:

$$\Pr\{S_{io} = 0\} = 0.2 \,,$$
$$\text{otherwise } S_{io} = 0.0874 + 0.01667U_1 - 0.084\sqrt{U_2} \,,$$

where $U_1$ and $U_2$ are independent uniform random-variables. This distribution is based on the measure-

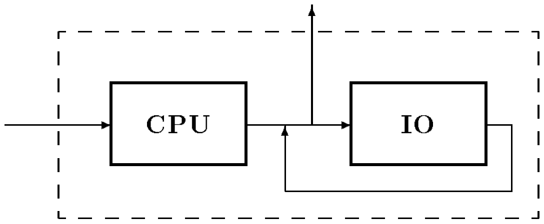



**Figure 4.** Queueing Network Model 3

**Table 2.** Summary of Measurement Data

| | Mon | Tue | Wed | Thu | Fri |
|---|---|---|---|---|---|
| $n$ | 5641 | 6482 | 5663 | 5819 | 5576 |
| $\lambda$ | .3134 | .3601 | .3146 | .3233 | .3098 |
| $D_{cpu}$ | 1.118 | 1.051 | 1.270 | 1.232 | 0.767 |
| $D_{io}$ | 1.336 | 1.158 | 0.726 | 1.620 | 0.782 |
| $V_{io}$ | 45.96 | 39.85 | 24.97 | 55.73 | 26.89 |
| System | | | | | |
| Mean Response | 4.019 | 3.678 | 3.055 | 5.448 | 2.038 |
| Mean Load | 1.26 | 1.32 | 0.96 | 1.76 | 0.63 |
| Utilizations | | | | | |
| CPU | .3504 | .3785 | .3995 | .3983 | .2378 |
| IO | .4187 | .4171 | .2283 | .5237 | .2421 |

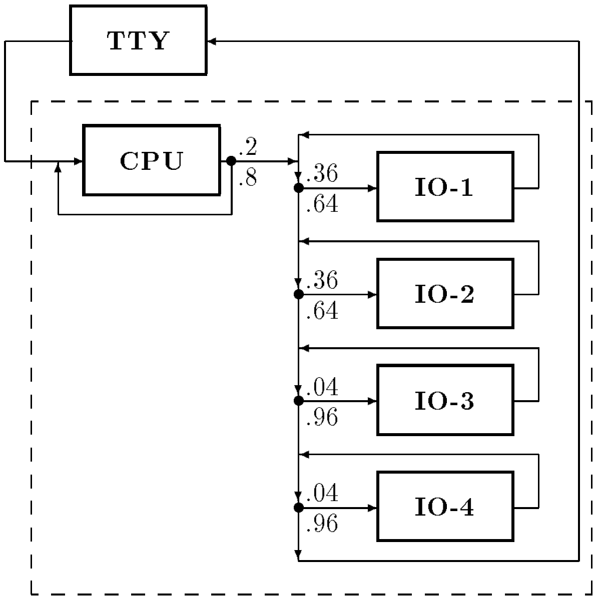



**Figure 5.** Independent Network of Model 2

ment data and on the properties of HSC50 and RA81 Disks.

## Should the Joint Distribution Be Considered?

The fundamental question of modeling the distributions is if the dependencies between the service demands must be considered or if the service demands can be assumed to be independent. In this section we examine how much some quantiles of the system response-time distribution in a network of independent service demands differ from those in a network of dependent service demands. The three queueing network models introduced in the previous section are considered.

In Model 1 the service demands are independent when $\pi_0 = 1$. In Model 2 we must change the routing behavior so that the service demands are independent but have the same marginal distributions as in the original network. This network model is decribed in Figure 5. In Model 3 we obtain the independent distributions by sampling from the measured marginal distributions while the dependent distributions are sampled from the measured joint distributions.

We simulated the original models and the models having independent service demands. In each case we generated 101 sequences of response times, each sequence containing 10 000 observations. The 101 sequences were obtained in a single simulation run. Between the adjacent sequences there was a gap of 1 000 observations. The simulation run was initialized using steady-state probabilities of queue lengths. The first sequence was excluded from the analysis. This experimentation plan ensures that the means of observed quantiles are reliable estimates.

Figure 6 summarizes the results as *comparative boxplots*, in which the ordinary boxplots are divided into two halves. The left-hand side corresponds to the original network (label O) and the right-hand side to the independent network (label I). In the boxplot, the mean is denoted by a bullet and the three quartiles by a box. The ends of the box denote the lower and upper quartiles, and the horizontal line inside the box denotes the median. The whiskers denote the $10^{th}$ and $90^{th}$ percentiles. Separate horizontal lines indicate the locations of the $5^{th}$, $95^{th}$, and $99^{th}$ percentiles.

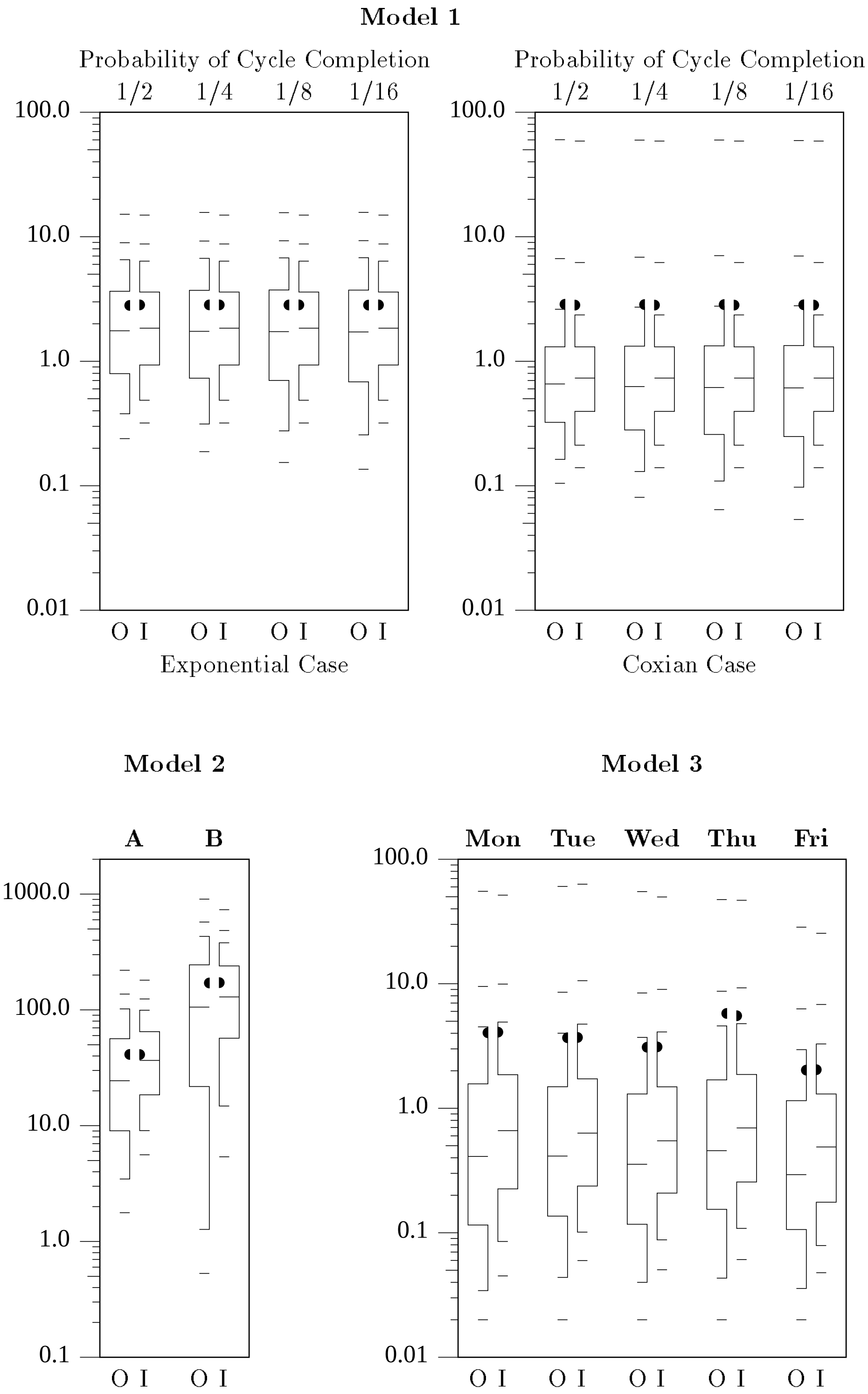



**Figure 6.** Response Times in **O**riginal and **I**ndependent Networks

The figure clearly indicates that the dependencies significantly affect some quantiles. Table 3 contains the estimated values of $\Pr\{|x_{orig} - x_{inde}| \leq 0.1x_{orig}\}$. The probability estimates are based on $t$-approximation:

$$\widehat{\Pr}\{|x_{orig} - x_{inde}| \leq 0.1x_{orig}\} = \int_{-\infty}^{\infty} f(t_1) \int_{\underline{t}(t_1)}^{\overline{t}(t_1)} f(t_2)\, dt_2\, dt_1 \ ,$$

where $f(t)$ is the probability density function of Student's $t$-distribution with 99 degrees of freedom,

$$\begin{aligned} \underline{t}(t_1) &= 0.9\frac{s_{orig}}{s_{inde}}t_1 + \frac{10}{s_{inde}}(0.9\bar{x}_{orig} - \bar{x}_{inde}) \ , \\ \overline{t}(t_1) &= 1.1\frac{s_{orig}}{s_{inde}}t_1 + \frac{10}{s_{inde}}(1.1\bar{x}_{orig} - \bar{x}_{inde}) \ , \end{aligned}$$

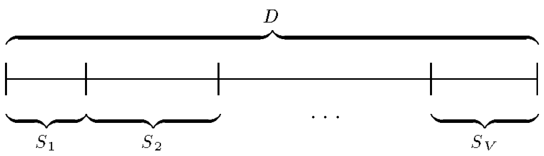
$\bar{x}_{orig}$ and $\bar{x}_{inde}$ are the means of an simulated quantile, $s_{orig}$ and $s_{inde}$ are the standard deviations.

In 74 cases (out of 120) the estimated probability is less than 0.10. Therefore, the conclusion of these simulation experiments is self-evident: *If we are analyzing quantiles of the system response-time distribution, we must take the dependencies between the service demands into account.*

## Decomposing the Service Demand into Service Times

As discussed in the Introduction, a typical accounting file provides the total cpu-time and counts of io-operations. Therefore, in a central server model we obtain the service demand and the number of visits at the cpu-node. However, in a simulation study we need the number of visits and the service times per visit. In this section we introduce a simple technique of decomposing the service demand into service times.

Suppose that we can generate (using the joint distribution) the service demand $D$ and the number of visits $V$. Our task is to divide $D$ into $V$ service times $S_i \ , \ i = 1, \ldots, V$:




A straightforward approach is to generate $V-1$ random points ($U_i \ , \ i = 1, \ldots, V-1$) in interval $(0,1)$ and then to determine

$$S_i = (U_{(i)} - U_{(i-1)})D \ , \ i = 1, \ldots, V \ , \qquad (1)$$

where $U_{(0)} = 0$, $U_{(V)} = 1$, and $U_{(i)}$ is the $i^{\text{th}}$ order statistic of $\{U_1, \ldots, U_{V-1}\}$.

Of practical reasons, however, we want to extract one service time at a time from the remaining service demand, that is

$$S_i = \beta_i(D - \sum_{j=1}^{i-1} S_j) \ , \qquad (2)$$

where $\beta_i$ is a random variable. It is straightforward to show that the two approaches are identical, if $\beta_i \ , \ i = 1, \ldots, V-1$, is the first order statistic of $V-i$ independent uniform random variables and $\beta_V = 1$. Then the distribution of $\beta_i$ is Beta$(1, V-i)$ (see e.g. Shorack and Weller 1986, p. 97).

## Modeling Joint Distribution

In many simulation applications, it is necessary to generate dependent input processes. For practical reasons, there is a great tendency to model the marginal distributions independently and to characterize the dependencies by a matrix of correlation-type coefficients. In this section we propose a technique for modeling the dependencies using a multinormal distribution. The multinormal distribution was selected, since it is easy to use in a general multidimensional case. In addition, we show that this method leads to quite accurate estimates for the quantiles of the system response-time in the networks examined.

### Characterization of Dependencies

The basic problem of modeling input processes which are not independent and have specific marginal distributions is how the dependencies are characterized. Most often the characterization is based on the product moment correlation. The problem of the product moment correlation $\rho$,

$$\rho = \frac{\mathrm{E}[(x - \mathrm{E}(x))(y - \mathrm{E}(y))]}{\sqrt{\mathrm{E}[x - \mathrm{E}(x)]^2 \mathrm{E}[y - \mathrm{E}(y)]^2}} \ ,$$

**Table 3.** Estimated Probabilities of Quantiles in Independent Network being within 10% to Quantiles in Original Network

| | Model 1 | | | | | | | | Model 2 | | Model 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Exponential | | | | Coxian | | | | | | | | | | |
| Quan- | Probability of Cycle Completion | | | | | | | | | | | | | | |
| tile | $^1/_2$ | $^1/_4$ | $^1/_8$ | $^1/_{16}$ | $^1/_2$ | $^1/_4$ | $^1/_8$ | $^1/_{16}$ | **A** | **B** | Mon | Tue | Wed | Thu | Fri |
| $x_{.05}$ | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| $x_{.10}$ | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| $x_{.25}$ | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| $x_{.50}$ | .999 | .999 | .999 | .999 | .045 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| $x_{.75}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .000 | .999 | .000 | .000 | .000 | .362 | .000 |
| $x_{.90}$ | .999 | .999 | .999 | .999 | .038 | .000 | .000 | .000 | .999 | .000 | .838 | .000 | .112 | .999 | .010 |
| $x_{.95}$ | .999 | .999 | .999 | .999 | .937 | .377 | .015 | .050 | .999 | .000 | .999 | .000 | .999 | .988 | .997 |
| $x_{.99}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .000 | .000 | .993 | .999 | .667 | .999 | .142 |

is that it measures symmetric departures. When the marginal distributions are skew, the product moment correlation is not appropriate. Robust coefficients of correlation (see e.g. Devlin, Gnanadesikan and Kettenring 1981) are appropriate if the distributions are moderately skew. In the case of service demands, however, the robust correlations are not appropriate, since the distributions of service demands are typically very skew.

The rank-correlation solves the problem of skewness. In addition, rank-correlations have natural interpretations if the random-number generation is based on inverse transforms: $X_j = F_{X_j}^{-1}(U_j)$. The rank-correlation between $X_j$ and $X_k$ is the product moment correlation between the uniformly distributed $U_j$ and $U_k$. The problem of applying rank-correlations is related to the uniform marginal distributions. Only in the bivariate case we can easily generate the multidimensional $\mathbf{U}$ so that the marginal distributions are uniform and the product moment correlations are those desired. The bivariate uniform distribution is discussed, for example, in Johnson and Tenenbein (1981). The extension to the case of more than two variables is difficult. In Iman and Conover (1982) a method that approximates the desired rank-correlations is proposed. The computer implementation of the method (Iman and Shortencarier 1985) is more than 3 000 lines of Fortran.

We propose that the dependencies are characterized using a multidimensional normal distribution. The approach is motivated by the simplity of generating multidimensional normal distributions. We generate the service demands $X_1, \ldots, X_m$ using a $m$-dimensional normal distribution and the inverses of the marginal cumulative distribution functions of the $X_j$'s:

1. Generate $\mathbf{Z}$ from $N(0, \mathbf{P})$.
2. Set $X_j = F_j^{-1}(\Phi(Z_j))$, $j = 1, \ldots, m$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

The advantage of this method is simplity. The dependencies are characterized by the correlation matrix $\mathbf{P}$. A similar technique is used in Hull (1977) proposing a quite complicated procedure for estimating the correlations. However, we can use the following straightforward algorithm:

1. Construct the corresponding normal scores: If $x_{i,j}$ is the $k^{\text{th}}$ observation in the ordered sample of $x_{\cdot,j}$'s, then $z_{i,j} = \Phi^{-1}(\frac{3k-1}{3n+1})$.
2. Estimate $\mathbf{P}$ as the sample correlation matrix of the $z_{i,j}$'s: $p_{kj} = \frac{1}{n}\sum_{i=1}^{n} z_{i,k} z_{i,j}$.

## Empirical Results

In this section we report our results of applying the proposed methods for modeling the dependencies and

**Table 4.** Measured Gaussian Correlations

**Model 1**

| $\pi_0$ | $D_{cpu}, V_{io}$ Expon. | Coxian |
|---|---|---|
| 1 | .00000 | .00000 |
| $^1/_2$ | .65443 | .60379 |
| $^1/_4$ | .83052 | .74660 |
| $^1/_8$ | .91181 | .83698 |
| $^1/_{16}$ | .95410 | .87806 |

**Model 2**

| | $D_{CPU}$ | $V_{IO-1}$ | $V_{IO-2}$ | $V_{IO-3}$ |
|---|---|---|---|---|
| $V_{IO-1}$ | .75489 | | | |
| $V_{IO-2}$ | .76068 | .61033 | | |
| $V_{IO-3}$ | .35957 | .29512 | .29992 | |
| $V_{IO-4}$ | .37878 | .31713 | .32890 | .15779 |

**Model 3**

| | $A, D_{cpu}$ | $A, V_{io}$ | $D_{cpu}, V_{io}$ |
|---|---|---|---|
| Mon | .16801 | .10353 | .79000 |
| Tue | .14846 | .08796 | .80472 |
| Wed | .12679 | .06513 | .79862 |
| Thu | .11559 | .05382 | .79540 |
| Fri | .11239 | .06184 | .79953 |

decomposing the service demands. The three queueing network models introduced in Section 2 are examined. For Models 1 and 2 we generated "accounting" data (service demand at the **CPU**-node and the numbers of visits at the **IO**-nodes) using the known joint distributions. The generated sets contained 10000 observations. The measured correlations of the underlying normal distributions, which we call the Gaussian correlations, are given in Table 4. The table also contains the Gaussian correlations of the measurement data used in Model 3.

The simulation experiments were conducted in the same way as in Section 3. In each case we generated 101 sequences of response times, each sequence contained 10 000 observations, and the first sequences were excluded. The actual method of generating the service times is described in Appendix.

Figure 7 summarizes the results. The left-hand side of the comparative boxplot corresponds to the original network (label O) and the right-hand side to the modeled network (label M). The visualization indicates that the proposed way of modeling the dependencies and decomposing the demands provides quantiles close to the originals. Table 5 contains the estimated probabilities that the quantile of the modeled network is within 10% to the quantile of the original network. Only in eight cases (out of 120) we have an estimated probability less than 90% that the modeled quantile is not within 10% to the original one.

## Conclusions

We have considered a practical situation of simulating a queueing network model in order to obtain estimates for quantiles of the system response-time distribution. We have shown that the dependencies between the service demands must be considered. We have proposed two practical techniques; one for modeling the dependencies and one for decomposing the service demand into the service times per visit. The simulation experiments clearly indicate that the techniques provide accurate estimates for the quantiles examined.

An important advantage of the proposed techniques is that they can be used with any models of the marginal distributions that provides the quantile function or the inverse of cumulative distribution function. We have not considered the modeling of the marginal distributions discussed in several textbooks and papers.

## References


AVRAMIDIS, A., AND J. R. WILSON. 1989. A Flexible Method for Estimating Inverse Distribution Functions in Simulation Experiments. In *1989 Winter Simulation Conference Proceedings*, pp. 428–426, (December 4–6, 1989, Washington, D.C.), E. A. MacNair, K. J. Musselman and P. Heidelberger (eds.). IEEE, Piscataway, N.J.

BUX, W., AND U. HERZOG. 1977. The Phase Concept: Approximation of Measured Data and Performance Analysis. In *Computer Performance*, pp. 23–38, K. M. Chandy and M. Reiser (eds.). North-Holland, Amsterdam.

DEVLIN, S. J., R. GNANADESIKAN AND J. R. KETTENRING. 1981. Robust Estimation of Dispersion Matrices and Principal Components. *Journal of the American Statistical Association* **76,** 354–362.

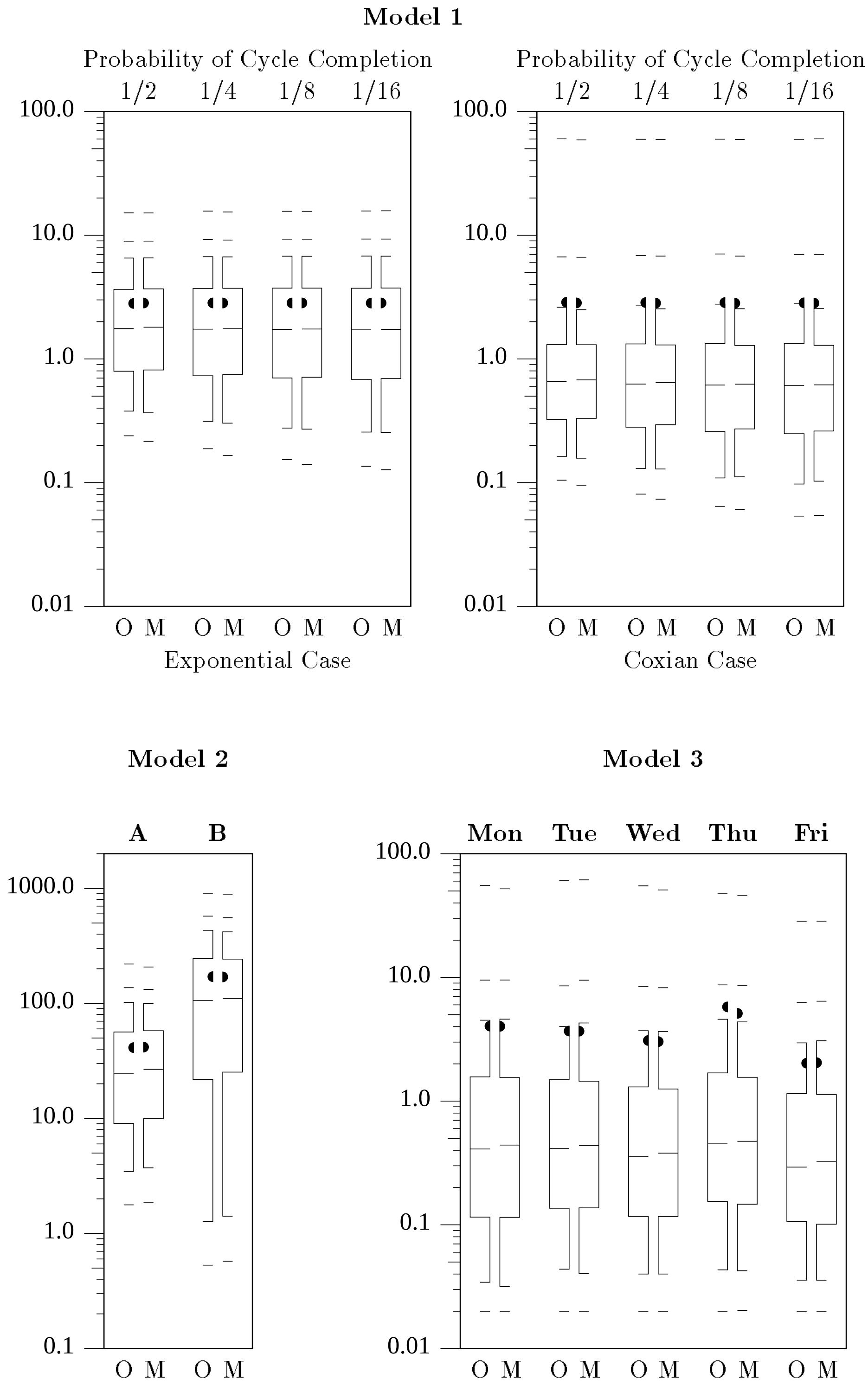



**Figure 7.** Response Times in **O**riginal and **M**odeled Networks

**Table 5.** Estimated Probabilities of Quantiles in Modeled Network being within 10% to Quantiles in Original Network

| Quantile | Model 1 | | | | | | | | Model 2 | | Model 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Exponential | | | | Coxian | | | | | | | | | | |
| | Probability of Cycle Completion | | | | | | | | | | | | | | |
| | $^1/_2$ | $^1/_4$ | $^1/_8$ | $^1/_{16}$ | $^1/_2$ | $^1/_4$ | $^1/_8$ | $^1/_{16}$ | **A** | **B** | Mon | Tue | Wed | Thu | Fri |
| $x_{.05}$ | .714 | .000 | .998 | .999 | .809 | .995 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 |
| $x_{.10}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .019 | .989 | .999 | .999 | .999 | .999 |
| $x_{.25}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .334 | .000 | .999 | .999 | .999 | .999 | .999 |
| $x_{.50}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .983 | .999 | .999 | .999 | .999 | .999 | .001 |
| $x_{.75}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .969 | .999 |
| $x_{.90}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 |
| $x_{.95}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .102 | .999 | .999 | .999 |
| $x_{.99}$ | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .999 | .990 | .999 | .999 |

DOWNTON, F. 1970. Bivariate Exponential Distribution in Reliability Theory. *Journal of the Royal Statistical Society* **B32,** 408–417.

HEIDELBERGER, P., AND P. D. WELCH. 1981. A Spectral Method for Confidence Interval Generation and Run Length Control in Simulation. *Communications of the ACM* **24,** 233–245.

HORA, S. C. 1983. Estimation of the Inverse Function for Random Variate Generation. *Communications of the ACM* **26,** 590–594.

HULL, J. C. 1977. Dealing with Dependence in Risk Simulation. *Operational Research Quarterly* **28,** 201–213.

IMAN, R. L. 1986. Modeling Input Processes. In *1986 Winter Simulation Conference Proceedings*, pp. 34–37, (December 8–10, 1986, Washington, D.C.), J. Wilson, J. Henriksen, and S. Roberts (eds.). IEEE, Piscataway, N. J.

IMAN, R. L., AND W. J. CONOVER. 1982. A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics* **B11,** 311-334.

IMAN, R. L., AND M. J. SHORTENCARIER. 1985. LHS: A Program to Generate Input Samples for Multivariate Simulations. Computer program announcement in *The American Statistican* **39,** 212.

JOHNSON, M. E., AND A. TENENBEIN. 1981. A Bivariate Distribution Family with Specified Marginals. *Journal of the American Statistical Association* **76,** 198–201.

KELTON, W. D. 1984. Input Data Collection and Analysis. In *1984 Winter Simulation Conference Proceedings*, pp. 91–95, (November 28–30, 1984, Dallas, Texas), S. Sheppard, U. W. Pooch and C. D. Pegden (eds.). IEEE, Piscataway, N.J.

KHOSHGOFTAAR, T. M., AND H. G. PERROS. 1987. A Comparison of Three Methods of Estimation for Approximating General Distributions by a Coxian Distribution. In *Modelling Techniques and Performance Evaluation*, pp. 81–89, S. Fdida and G. Pujolle (eds.). North-Holland, Amsterdam.

KIMELDORF, G., AND A. SAMPSON. 1975a. One-Parameter Families of Bivariate Distributions with Fixed Marginals. *Communications in Statistics* **4,** 293–301.

KIMELDORF, G., AND A. SAMPSON. 1975b. Uniform Representations of Bivariate Distributions. *Communications in Statistics* **4,** 617–627.

KOBAYASHI, H., AND M. REISER. 1975. On Generalization of Job Routing Behavior in a Queueing Network Model. IBM Research Report RC-5252. IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.

LAW, A. M., AND W. D. KELTON. 1982. *Simulation Modelling and Analysis.* McGraw-Hill, London, U.K.

Lazowska, E. D. 1977. The Use of Percentiles in Modeling CPU Service Time Distributions. In *Computer Performance*, pp. 53–66, K. M. Chandy and M. Reiser (eds.). North-Holland, Amsterdam.

Lazowska, E. D. and C. A. Addison. 1979. Selecting Parameter Values for Servers of the Phase Type. In *Performance of Computer Systems*, pp. 407–420, M. Arato, A. Butrimenko and E. Gelenbe (eds.). North-Holland, Amsterdam.

Mardia, K. V. 1970. A Translation Family of Bivariate Distributions and Fréchet's Bounds. *Sankhyā* **A32,** 119–122.

Mitchell, C. R., A. S. Paulson and C. A. Beswick. 1977. The Effect of Correlated Exponential Service Times on Single Server Tandem Queues. *Naval Research Logistics Quarterly* **24,** 95–112.

Paulson, A. S. 1973. A Characterization of the Exponential Distribution and a Bivariate Exponential Distribution. *Sankhyā* **A35,** 69–78.

Raatikainen, K. E. E. 1991. Modeling Service-Time Distributions for Queueing Network Simulation. *Communications in Statistics* **B20,** 375-390.

Ronning, G. 1977. A Simple Scheme for Generating Multivariate Gamma Distributions with Non-Negative Covariance Matrix. *Technometrics* **19,** 179–183.

Schmeiser, B. W., and R. Lal. 1980. Multivariate Modeling in Simulation: A Survey. In *ASQC Technical Conference Transactions, Vol 34*, pp. 252–261, (May 20–22, 1980, Atlanta, Ga.). The American Society for Quality Control, Milwaukee, Wis.

Schmeiser, B. W., and R. Lal. 1982. Bivariate Gamma Random Vectors. *Operations Research* **30,** 355–374.

Shorack, G. R., and Wellner, J. A. 1986. *Empirical Processes with Applications to Statistics.* John Wiley & Sons, New York, N.Y.

## Appendix: Detailed Description of Generating the Service Times in the Experiments

**Model 1:** When a customer leaves the **TTY**-node, the service demands are generated in the following manner:

1. Generate $(z_1, z_2)$ from the binormal distribution with means 0, variances 1, and correlation given in Table 4.
2. Set $D_{cpu} = F_{cpu}^{-1}(\Phi(z_1))$. In the exponential case, the distribution is exponential with mean 1.0333. In the Coxian case, the parameters are given in column $\pi_0 = 1$ of Table 1.
3. $V_{cpu} = V_{io} = \lceil \log(1-\Phi(z_2))/\log(1-\pi_0)\rceil$. In the case of $\pi_0 = 1$, $V_{cpu} = V_{io} = 1$.

As the customer enters the **CPU**-node:

1. $V$ is updated to $V - 1$,
2. If $V = 0$, then $\beta = 1$, otherwise $\beta$ is generated from Beta$(1, V)$,
3. the service time is determined as $S = \beta D$, and
4. $D$ is updated to $D - S$.

As the customer enters the **IO**-node, the service time is generated from the exponential distribution with mean $0.356\pi_0$. After the completion of service at the **IO**-node, the customer either leaves the system and enters the **TTY**-node if $V = 0$ or the customer enters the **CPU**-node if $V > 0$. As the customer enters the **TTY**-node, the residence time is generated from the exponential distribution with mean 30.

**Model 2:** When a customer leaves the **TTY**-node, the service demands are generated in the following manner:

1. Generate $(z_0, z_1, z_2, z_3, z_4)$ from the five-dimensional normal distribution with means 0, variances 1, and correlations given in Table 4.
2. Set $D_{cpu} = -5\log(1-\Phi(z_0))$ (exponential distribution with mean 5).
3. Set $V_{io-i} = \lfloor \log(1-\Phi(z_i))/\log(1-\pi_i)\rfloor$, where $\pi_1 = \pi_2 = 5/14$ and $\pi_3 = \pi_4 = 5/6$.
4. Set $V_{cpu} = 1 + V_{io-1} + \cdots + V_{io-4}$.

As the customer enters the **CPU**-node:

1. $V_{cpu}$ is updated to $V_{cpu} - 1$,
2. If $V_{cpu} = 0$, then $\beta = 1$, otherwise $\beta$ is generated from Beta$(1, V_{cpu})$,
3. the service time is determined as $S_{cpu} = \beta D_{cpu}$, and

4. $D_{cpu}$ is updated to $D_{cpu} - S_{cpu}$.

As the customer leaves the **CPU**-node, he enters the **TTY**-node, if $V_{cpu} = 0$, otherwise he enters one of the **IO**-nodes. The **IO**-node the customer enters is selected according to probabilities $V_{io-1}/V_{cpu}, \ldots, V_{io-4}/V_{cpu}$. As the customer enters the **IO-i**-node, $V_{io-i}$ is updated to $V_{io-i} - 1$ and $S_{io-i}$ is generated from the exponential distribution with rate given in Figure 3. As the customer enters the **TTY**-node, the residence time is generated according to the exponential distribution with mean 100.

**Model 3:** When a customer arrives, the service demands and the interarrival time are generated in the following manner:

1. Generate $(z_1, z_2, z_3)$ from the trinormal distribution with means 0, variances 1, and correlations given in Table 4.
2. Set $A = F_{arr}^{-1}(\Phi(z_1))$; i.e. the next customer arrives after $A$ seconds.
3. Set $D_{cpu} = F_{cpu}^{-1}(\Phi(z_2))$.
4. Set $V_{io} = F_{io}^{-1}(\Phi(z_3))$.

The distributions $F_{arr}$, $F_{cpu}$, and $F_{io}$ are discrete and determined by the measurement data: $\Pr\{X = x_i\} = 1/n$, where the $x_i$'s are the measured values.

When the customer arrives, he first requests $D_{cpu}$ seconds of service at the **CPU**-node. Since the **CPU**-node is an PS-server and we examine the steady-state characteristics, the service demand need not to be decomposed into service times. When the customer has received the cpu-service he requested, the customer makes $V_{io}$ requests to the **IO**-node. The service time of the IO-requests is generated from the distribution given in Section 2. If the service time is zero, the customers immediately initiates a new request.