Approximating Response Time Distributions

Kimmo E. E. Raatikainen†
University of Helsinki, Dept. of Computer Science
Teollisuuskatu 23–25, SF-00510 Helsinki, FINLAND
KRAATIKA@FINUH.BITNET

Abstract: *The response time is the most visible performance index to users of computer systems. End-users see individual response times, not the average. Therefore the distribution of response times is important in performance evaluation and capacity planning studies. However, the analytic results cannot be obtained in practical cases. A new method is proposed to approximate the response-time distribution. Unlike the previous methods the proposed one takes into account the service-time distributions and routing behaviour. The reported results indicate that the method provides reasonable approximations in many cases.*

1 Introduction

Queueing network modelling is a popular tool in the performance evaluation of computer systems. It has been successfully used in various applications of modelling computer systems. In most applications only the mean values of performance indices, such as mean device queue-lengths and the mean system response-time, have been considered. The increasing usage of computer systems as production tools has raised the system response-time as the most important single performance indicator. People using the computer system see the response time as the only index of performance. The end-user experiences an individual response time, not the average. Therefore, the distribution of the system response-time, usually summarized by a set of quantiles or fractions, should receive greater interest.

Despite recent advances in analytic queueing network analysis, only for simple cases does an analytic technique give the distribution of the system response-time.

†This work was supported by Emil Aaltonen Foundation

The most general cases are a cyclic queue and a network having a tree topology of exponential servers. The key assumption is the so-called overtaken-free condition [KePo83]. It means that any customer who arrives later than the tagged customer always stays behind him. The overtaken-free condition guarantees that the residence times at different nodes are independent. The residence-time distribution at a single node is available if the node is either a FCFS c-server or an IS server [McKe87].

There are two very simple queueing network models where the overtaken-free condition is not necessary. The first of them is a two-node network consisting of an IS center and a PS center. A detailed analysis is given in [Mitr81]. The second model is a three-node network of FCFS exponential servers, where there are two possible paths: $1 \rightarrow 2 \rightarrow 3$ and $1 \rightarrow 3$. This network is analyzed in [FaIM83].

Some approximate methods have also been proposed in [LaSe79], [SaLa81], and [ReSe88], for example. These methods can be used in slightly more complicated models. All the three methods are based on the same basic idea. The subnetwork the response-time distribution of which is analyzed is replaced by a load-dependent exponential server.

In this paper we introduce an alternative approach to approximate the response-time distribution in multiclass queueing network models having product-form solution. The method is based on the Laplace transform of an approximate response-time distribution. The response time is approximated as a random sum of residence times. The residence-time distribution of an IS-node is the same as its service-time distribution. The residence-time distribution of a PS- or FCFS-node is approximated through analyzing the node in isolation. The distribution of terms in the random sum is represented as the joint probability generating function.

The method includes three phases:

1. The Laplace transforms of the residence-time distribution for the PS- and FCFS-nodes are con-

structed using the queue-length distributions at arrival instants and the theorem of total Laplace transform. In multiclass models we need only the queue-length distribution over all classes. The transforms are constructed for each class the response-time distribution of which is approximated.

2. The joint probability generating function, the z-transform, for the numbers of visits is constructed. In multiclass models we need the z-transforms for the analyzed classes.
3. Approximate quantiles or proportions of the response-time distribution are evaluated using numerical inversion of the Laplace transform.

The construction of the z-transform implies that our method takes into account the dependencies between the service demands at different nodes. Although the service times per visit are independent, the service demands are not mutually independent. The dependencies arise from the routing behavior. The number of visits at one node depends stochastically on the numbers of visits at the other nodes. To the best of our knowledge the effects of the joint distribution have not been considered earlier in queueing network modelling. When the queue-length distributions and means of indices based on them are analyzed the routing behavior, and hence the joint distribution, has no effect. However, when quantiles of the response-time distribution are analyzed, these effects must be considered.

The quantiles of the response-time distribution are evaluated using a Laplace transform. In principle, we can obtain the cumulative distribution function by inverting the transform using partial fractioning. However, in networks of practical interest the partial fractioning requires that roots of a high-order polynomial must be found. The numerical problems of partial fractioning can be avoided, if the transform is inverted numerically. Numerical inversion is not a simple task, but good numerical software is available, for example the Crump's method in NAG Library [NAG87].

The accuracy of the proposed method is examined empirically using two different network models. The models were analyzed with several sets of parameters. In addition we consider a real-life modelling situation. We compare the results obtained by our method to those given by simulation and the method proposed in [LaSe79]. The experimental results indicate that our method provides a reasonable approximation in many situations. Especially, when the service times are not exponentially distributed or when the central server is not the bottleneck of the system our method is superior to the method given in [LaSe79].

2 Description of Method

In this section we describe our approximative method. The description is given in three parts. In the first part we show how the residence-time distributions of PS- and FCFS-nodes can be approximated. The joint distribution of the numbers of visits is discussed in the second part. Finally, we discuss the reasons for using numerical inversion of the Laplace transform.

In product-form networks our method has three primary sources of errors. We assume (incorrectly) that the individual residence times are independent. The residence-time distributions of the PS- and FCFS-nodes are obtained by analyzing the nodes in isolation. The analysis in isolation does not take into account the dependencies between conditional queue-length distributions. The queue-length distributions at a specific arrival instant are independent. However, this does not imply that the queue-length distribution generally seen by a tagged customer is independent of the queue-length distributions he has seen earlier. Secondly, we have not found any reasonable way to consider the dynamic changes in the number of customers at a PS-node. The third source of errors is the numerical method of inverting the Laplace transform.

As pointed out in [SaLa81], the individual residence times are not, in general, independent. We can motivate the assumption of independence as a flow-equivalent approximation, in addition to the mathematical convenience. If we approximate all the non-IS nodes by flow-equivalent IS-servers, we obtain a network consisting only of IS-nodes. In the product-form networks the service times per visit are independent and within each node identically distributed. Therefore the response time in a network of IS-servers is a random sum of independent random variables. Traditionally the flow-equivalence approximation is based on the equal mean rate. Since we are interested in the residence-time distribution, we need a more complete flow-equivalence. The service-time distribution of the flow-equivalent IS-server must approximate the residence-times distribution at the original node.

A specific response time is a sum of specific residence times:

$$R = \sum_{j=1}^{m} \sum_{k=1}^{V_j} R_{jk} ,$$

where m is the number of visited nodes, V_j is the number of visits at node j, and R_{jk} is the residence time of the k^{th} visit at node j. Let $\mathcal{L}_j(s)$ denote the Laplace transform of the residence-time distribution at node j. Let $p(n_1, \ldots, n_m)$ be the probability that the nodes are visited $n_1, \ldots, n_m$ times, respectively. Then, using *the* *theorem of total Laplace transform* and *the independence assumption* we can express the Laplace transform of the response time distribution as

$$\begin{aligned}\mathcal{L}_R(s) &= \\ &\sum_{n_1=0}^{\infty} \cdots \sum_{n_m=0}^{\infty} p(n_1,\ldots,n_m)\mathcal{L}_1(s)^{n_1} \cdots \mathcal{L}_m(s)^{n_m} \\ &= \mathcal{N}\left(\mathcal{L}_1(s),\ldots,\mathcal{L}_m(s)\right) \ , \qquad (1)\end{aligned}$$

where $\mathcal{N}(z_1,\ldots,z_m)$ is the joint probability generating function (z-transform) for the numbers of visits at different nodes.

Equation (1) is the foundation of our approximation. The approximate quantiles of the response-time distribution are obtained by inverting numerically $\mathcal{L}_R(s)/s$, when appropriate approximations of the residence-time distributions are substituted into the formula.

2.1 Approximate Residence Time Distributions

In a **FCFS-node** the residence-time distribution depends on the service rate μ (the same exponentially distributed service time for all the customers is assumed) and on the queue-length distribution at the arrival instants. Let q_n be the probability that an arriving customer sees n customers at the node. Since the exponential distribution is memoryless, the residence time is the sum of $n+1$ independent and identically distributed random variables. Using the theorem of total Laplace transform we get the Laplace transform of the exact residence-time distribution:

$$\mathcal{L}_r(s) = \sum_{n=0} q_n \left(\frac{\mu}{s+\mu}\right)^{n+1} \ . \qquad (2)$$

The case of a c-server is slightly more difficult but using the formulas given in [McKe87] we obtain

$$\begin{aligned}\mathcal{L}_r(s) &= \left(\frac{\mu}{s+\mu}\right) \times \\ &\left[q_0 + \cdots + q_{c-1} + \sum_{n=c} q_n \left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\right] \ .\end{aligned}$$

The case of a **PS-node** is complicated. We cannot construct an exact residence-time distribution, but only an approximation. The service-time distribution is a general (Coxian) distribution and different for each class of customers. Secondly, the proportion of capacity, which a customer receives, depends on the dynamic load condition at the node, and the load condition may change during the residence time. We have not found any reasonable way to take these dynamic changes into account. We assume that the proportion of capacity that a customer receives is constant during the whole residence time. In other words, we assume that the effects of arrivals and departures during each individual residence time balance each other. This implies the assumption that a specific residence time is $n \times$ *Service Time* with probability $\Pr\{\textit{Queue Length} = n | \textit{Queue Length} > 0\}$. In the product-form closed network the conditional probability of queue length n is the probability of queue-length $n-1$ in a network of population $N-1$. In [LaRe80] it is shown that this probability is equal to the probability that an arriving customer sees $n-1$ customers at the node. In open networks arriving customers see the steady-state distribution.

Let $\mathcal{L}(s)$ be the Laplace transform of the service-time distribution for the class being analyzed. The Laplace transform of the approximate residence-time distribution is given by

$$\mathcal{L}_r(s) = \sum_{n=1} q_{n-1}\mathcal{L}(ns) \ . \qquad (3)$$

The multiclass models do not cause any additional difficulties. When we consider a specific class, the other classes affect only the queue-length distribution and the required queue-length distribution is over all classes.

2.2 Joint Distribution of Visits

One of the two major practical problems in our approximate method is that the joint z-transform for the numbers of visits is needed. In product-form queueing networks the paths are generated by a discrete Markov chain. The routing behaviour is completely characterized by the matrix of one-step transition probabilities. Therefore the z-transform can be constructed explicitely (in theory) for any network having constant routing probabilities. In practice, the z-transform of an arbitrary network may be untractable.

For each node in a product-form network we can always evaluate the probability that the node is re-entered. In addition, we can also evaluate the probability that the node is visited at least once. The calculation of these probabilities requires matrix manipulation, but the formulas can be found in fundamental textbooks for Markov chains (see, for example, [Klei75, Ch 2.3]). Let the probabilities be π_{re} and π_{visit}, respectively. Now the number of visits V in that specific node has the following probabilities:

$$\Pr\{V=n\} = \begin{cases} 1-\pi_{visit} & , n=0 \ , \\ \pi_{visit}(1-\pi_{re})\pi_{re}^{n-1} & , n>0 \ . \end{cases}$$

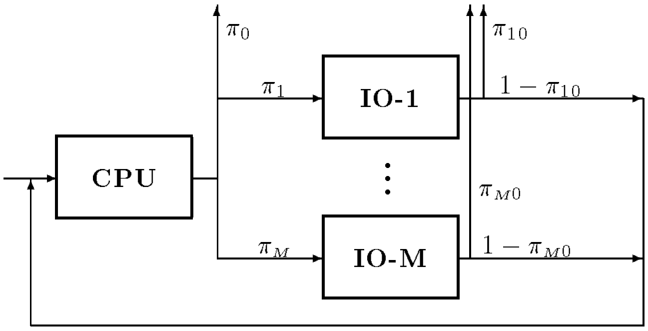

Figure 1: General Central Server Model

Many networks of practical importance include one node so that the conditional joint distribution for the numbers of visits in other nodes can be evaluated, when the number of visit in one node is known. For example, in a central server model the conditional joint distribution is multinomial.

Let us consider a general form of the central server model, given in Figure 1. The probability that the **CPU** node is visited is 1. The probability that the **CPU** node is re-entered is given as

$$\pi_{re} = \sum_{j=1}^{M} \pi_j (1 - \pi_{j0}) .$$

Therefore the number of CPU visits is distributed as

$$\Pr\{V_0 = n\} = (1 - \pi_{re}) {\pi_{re}}^{n-1} , \quad n = 1, 2, \ldots . \tag{4}$$

If the **CPU** is visited n times, the total number of visits in the **IO** nodes is either $n-1$ or n. The corresponding probabilities are $\pi_0/(1-\pi_{re})$ and $1-\pi_0/(1-\pi_{re})$. Each time one of the **IO** nodes is visited, the node is selected independently according to the probabilities $\pi_1/(1-\pi_0), \ldots, \pi_M/(1-\pi_0)$. The conditional distribution of the number of visits is multinomial. The corresponding joint z-transform is of form:

$$\mathcal{N}(z_1, \ldots, z_M \mid n_1 + \cdots + n_M) = \left(\frac{\pi_1 z_1 + \cdots + \pi_M z_M}{1 - \pi_0} \right)^{n_1 + \cdots + n_M} . \tag{5}$$

The joint z-transform for the numbers of visits can now be expressed as

$$\mathcal{N}(z_0, \ldots, z_M) = \sum_{n=1}^{\infty} (1 - \pi_{re}) \pi_{re}^{n-1} z_0^n \times \left\{ \left(1 - \frac{\pi_0}{1 - \pi_{re}} \right) \mathcal{N}(z_1, \ldots, z_M \mid n) + \frac{\pi_0}{1 - \pi_{re}} \mathcal{N}(z_1, \ldots, z_M \mid n-1) \right\} \tag{6}$$

$$= \frac{z_0 \left[\frac{1 - \pi_0 - \pi_{re}}{1 - \pi_0} (\pi_1 z_1 + \cdots + \pi_M z_M) + \pi_0 \right]}{1 - \frac{\pi_{re}}{1 - \pi_0} z_0 [\pi_1 z_1 + \cdots + \pi_M z_M]} .$$

In many cases the expression (6) can be simplified. For example, in the traditional central server model, the π_{i0}'s are zero, so $\pi_{re} = 1 - \pi_0$ and

$$\mathcal{N}(z_0, \ldots, z_M) = \frac{\pi_0 z_0}{1 - z_0 (\pi_1 z_1 + \cdots + \pi_M z_M)} .$$

2.3 Inversion of Laplace Transform

The second major practical difficulty in our approximate method is that we obtain the approximate response-time distribution as the Laplace transform:

$$\mathcal{L}_R(s) = \int_0^\infty \exp(-st) f_R(t) dt , \tag{7}$$

where $f_R(t)$ is the probability density function of the approximate response-time distribution. The cumulative probability function can, in principle, be obtained through the inverse transform:

$$F_R(t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \exp(tc) \frac{\mathcal{L}_R(c + i\omega)}{c + i\omega} \exp(it\omega) d\omega , \tag{8}$$

where c is an arbitrary constant so that $\mathcal{L}_R(s) \neq 0$, when $\mathrm{Re}(s) > c$.

The second possibility of inverting the transform is to use partial fractioning, if the Laplace transforms of all the service-time distributions are rational functions (or have been approximated as rational functions). Therefore, at least in centralized networks the approximate response-time distribution has a rational Laplace transform. However, in practical situations the degrees of the numerator and denominator are high. Only in simple networks with few customers are the degrees less than 20. Numerical difficulties will appear when the partial fractioning is performed.

Usually we are not interested in the analytic expressions of $F_R(t)$, but either in proportions (values of $F_R(t)$), or in quantiles (values of $F_R^{-1}(p)$). Since the functions

$$\exp(-tc) F_R(t) \quad \text{and} \quad \frac{\mathcal{L}_R(c + i\omega)}{c + i\omega}$$

are an integral Fourier-transform pair, the values of $\exp(-ct)F(t)$ can be approximated by Fourier series. The Crump's method [Crum76] of inverting the Laplace transform numerically is based on a Fourier-series approximation.

The numerical inversion has two important advantages. It is not necessary to express the $\mathcal{L}_R(s)$ as a ratio of two polynomials. The construction of the polynomial coefficients is sensitive to rounding errors, since we must substitute the approximations (2) and (3) into (1). Secondly, the roots of the denominator must be solved numerically. The solution of a polynomial equation of a high order is a numerically ill-posed problem (see e.g. [StBu80, Ch. 5.8]). These two factors together result in the numerically solved roots being usually inaccurate, and sometimes completely wrong. For example, we tried to solve the roots of denominator in the first example given in the next section. We used 64-bit floating-point arithmetic and a NAG Library routine, which is of high quality, but still obtained negative rates.

3 Experimental Results

In this section we give experimental results of the accuracy of our approximation. We have analyzed two different queueing network models, with several sets of parameters. In addition to these model we have considered a real-life modelling situation. The results of the proposed method are compared to the simulation results and to the approximations given by the Lazowska-Sevcik method [LaSe79].

3.1 Example Networks

The first queueing network model (see Figure 2) consists of three nodes: **CPU**, **TTY**, and **IO**. The queueing disciplines are PS, IS, and FCFS, respectively. The model has a single parameter, π_0, that determines the joint distribution of the **CPU** and **IO** demand. The z-transform of the joint distribution for the numbers of visits is

$$\mathcal{N}(z_{cpu}, z_{io}) = \frac{\pi_0 z_{cpu} z_{io}}{1-(1-\pi_0) z_{cpu} z_{io}} .$$

The service times in nodes **TTY** and **IO** are exponentially distributed with means 30 and $0.356\pi_0$, respectively. Two different service-time distributions are used in node **CPU**: an exponential with mean $1.0333\pi_0$ and a three-stage Coxian with the Laplace transform:

$$\begin{aligned}\mathcal{L}_S(s) \;=\; & (1+12.9586\, s)/\big(1+(12.9586+1.0333\pi_0)s \\ & +2.60015\pi_0\, s^2 + 0.00104633\pi_0\, s^3\big) \ .\end{aligned}$$

The Coxian distribution approximates the measured CPU-demand distribution in the examined computer system (see Section 3.2) quite closely. Note, that the (marginal) service-demand distributions are the same for all $\pi_0 > 0$.

The simulated and approximated quantiles of the system response-times for five different values of π_0 are visualized in Figure 3 by *comparative boxplots*. It should be noted that the Lazowska-Sevcik method (labelled as L&S) does not take into account the effects due to π_0 or the shape of the service-time distributions. In *the comparative boxplot* the ordinary boxplot (see for example [VeHo81, Ch. 3]) is divided into two halves. The left-hand side corresponds to the simulated results and the right-hand side to our approximate method. The approximations given by the L&S-method are displayed as an ordinary boxplot, since they are the same for all positive π_0's. The "box" corresponds to the interquartile range. The horizontal line inside the box denotes the median. The "whiskers" denote the 10^{th} and 90^{th} percentile. Separate horizontal lines indicate the locations of the 5^{th}, 95^{th}, and 99^{th} percentile. The scale is $\log_{10}$, hence the mismatch of the halves indicates the differences in magnitudes.

With both service-time distributions our method provides approximations, which are close, usually within 10 %, to the simulated quantiles. In the Coxian case the distribution has a tail so long and thin that the numerical inversion of the Laplace transform provides only two leading digits for the 99^{th} percentile. In fact, values in the range of $[56.6, 58.1]$ give 0.9900 as the value of cumulative distribution function.

The L&S-method gives acceptable approximations for all the quantiles only when the service time is exponential and π_0 is small. This is not surprising, since the L&S-method is derived using the assumption that π_0 is small.

The second queueing network model (see Figure 4) is due to Heidelberger and Welch [HeWe81]. All the service times are exponentially distributed. The **TTY** node is an infinite server and all the other nodes are FCFS servers. Two different sets of service demands were used. In **Case A** the central server, **CPU**, is the bottleneck of the system. The **IO-1** and **IO-2** are the bottlenecks in **Case B**. The simulated and approximated quantiles of the system response-times are visualized by ordinary boxplots in Figure 5. Note that the 99^{th} percentile is not given, because the simulation results are taken from our earlier study.

In **Case A** both our method and the L&S-method approximate the upper quantiles with good accuracy. The relative differences to the simulation results in the lower quantiles are high. Our method is slightly better. In the median the relative difference is 18% in our method and 31% in the L&S-method. In **Case B**, the L&S-method

Population	20
Service Demands	
CPU	1.033
IO	.356
TTY	30
System	
Throughput	.630
Mean Response	2.829
Mean Load	1.724
Utilizations	
CPU	.630
IO	.217

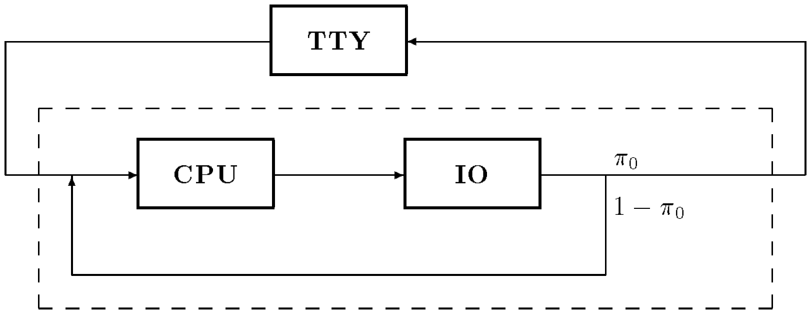

Figure 2: Queueing Network Model 1

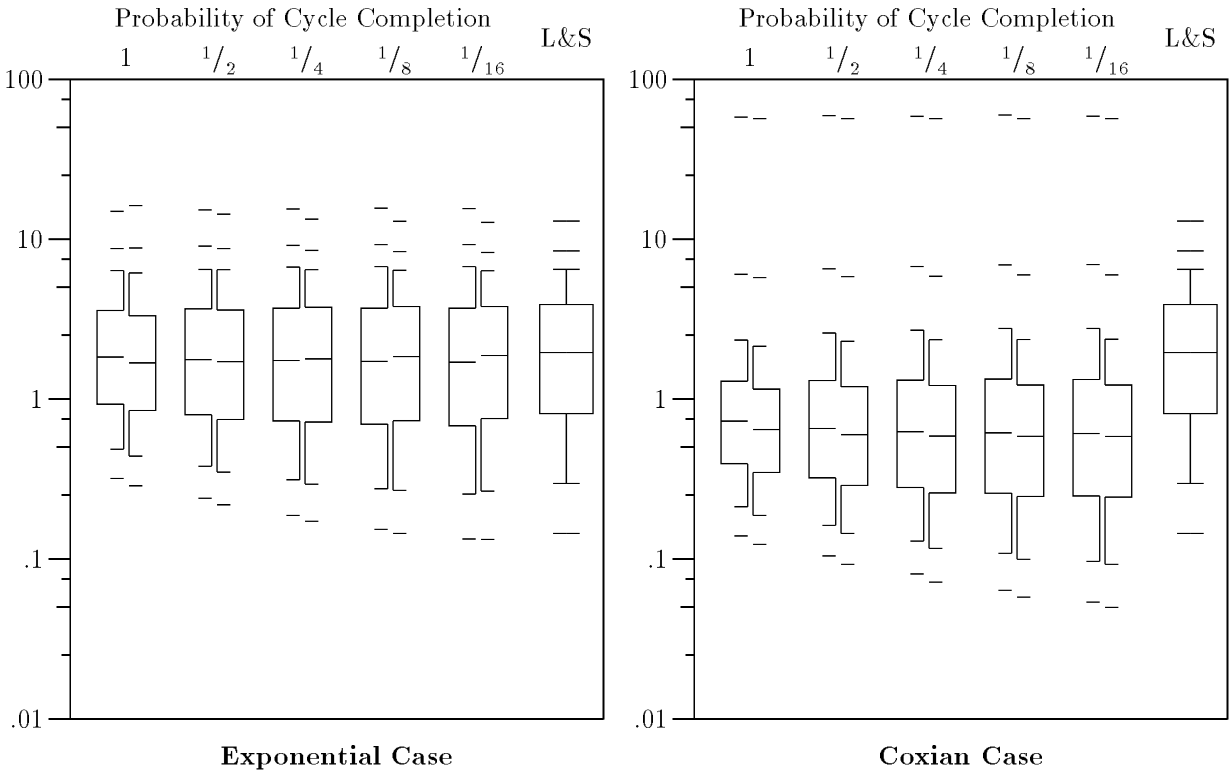

Figure 3: Comparative Boxplots of Response Times – Example 1

cannot capture the thin left-hand tail of the response-time distributions. On the other hand, our method predicts correctly that the distribution has a thin left-hand tail.

The overall conclusion from these experiments is that our method approximates correctly at least the magnitudes of a wide range of quantiles. It takes into account both the shape of the service demand distributions and their joint distribution. In addition the computational costs of our method are low. In the examples the cpu-time to evaluate $F_R(x)$ for a given x varied from .01 to .05 seconds using a VAX 8800 and from .1 to .3 seconds using a μVAX-II.

3.2 Real-Life Modelling Situation

We have also examined the proposed method in a "real-life" modelling situation. The examined system is a VAX 8650 with a HSC50 disk-server used in academic research. The reported experiences are based on the initial phase of the modelling study. The analyzed VAX/VMS accounting files contained the data for 44198 interactive tasks executed in prime-time during one week.

In the VAX/VMS-system the IO activity is measured in counts of *direct* and *buffered io-operations*. We used benchmark results to convert the counts to service times. The empirical Laplace transforms of the measured io-delay times and the empirical count distributions were used to construct the empirical Laplace transforms of the service-demand distributions:

$$\begin{aligned}
\mathcal{L}_{xio}(s) &= \sum_c \Pr\{xio\ count = c\} \times \left[\tilde{\mathcal{L}}_{xio}(s)\right]^c , \\
\tilde{\mathcal{L}}_{xio}(s) &= \frac{1}{N}\sum_k \exp(-sX_k) ,
\end{aligned}$$

where N is the number of measured io-requests and the X_k's are the observed response times.

The service-demand distributions were approximated by Coxian distributions. The approximate distributions were constructed by fitting a minimal degree rational approximation [Robe77]. The obtained Laplace transforms were

$$\begin{aligned}
\mathcal{L}_{CPU}(s) &= (1+12.9586\,s)/(1+13.9919\,s \\
&\quad +2.60015\,s^2+.00104633\,s^3) , \\
\mathcal{L}_{BIO}(s) &= (1+95.6177\,s+14.4486\,s^2)/ \\
&\quad (1+105.206\,s+199.980\,s^2) , \text{ and} \\
\mathcal{L}_{DIO}(s) &= (1+16.2309\,s+1.29018\,s^2)/ \\
&\quad (1+17.3094\,s+5.99810\,s^2) . \qquad (9)
\end{aligned}$$

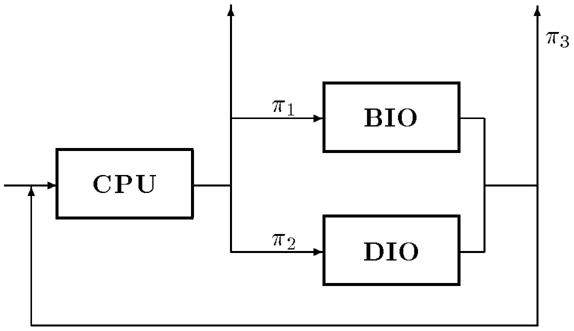

Figure 6: Examined Queueing Network Model

We modelled the system as an open queueing network (see Figure 6) containing one CPU node and two IO nodes: DIO and BIO. The CPU was modelled as a PS center and the IO nodes as IS centers. The arrival queue-length probabilities at the CPU were obtained through trace-driven simulation with actual arrivals and CPU service times: $q_0 = .79645$, $q_1 = .15738$, $q_2 = .03761$, $q_3 = .00793$, $q_4 = .00063$.

The joint z-transform for the numbers of visits is

$$\mathcal{N}(z_0, z_1, z_2) = \frac{z_0\,[\pi_3(\pi_1 z_1 + \pi_2 z_2) + 1 - \pi_1 - \pi_2]}{1-(1-\pi_3)z_0(\pi_1 z_1 + \pi_2 z_2)} . \qquad (10)$$

The service-demand distributions (9) lead to the following service-time (per visit) distributions:

$$\begin{aligned}
\mathcal{L}_0(s) &= \mathcal{L}_{CPU}(s)/ \\
&\quad \{1+(\pi_1+\pi_2)(1-\pi_3)[\mathcal{L}_{CPU}(s)-1]\} , \\
\mathcal{L}_1(s) &= \{[1-(1-\pi_3)\pi_2][\mathcal{L}_{BIO}(s)-1]+\pi_1\}/ \\
&\quad \{\pi_1[\pi_3+(1-\pi_3)\mathcal{L}_{BIO}(s)\} \text{ , and} \\
\mathcal{L}_2(s) &= \{[1-(1-\pi_3)\pi_1][\mathcal{L}_{DIO}(s)-1]+\pi_2\}/ \\
&\quad \{\pi_2[\pi_3+(1-\pi_3)\mathcal{L}_{DIO}(s)\} . \qquad (11)
\end{aligned}$$

The branching probabilities $\pi_1 = 0.698$, $\pi_2 = 0.283$ and $\pi_3 = 0.12$ minimize the Kolmogorov distance between the measured and modelled (Equation 10) joint distribution of the *buffered* and *direct io-count*.

The approximate quantiles of the response-time distribution are shown in Figure 7. As the reference we give the results of a trace-driven simulation, where the measured joint-distribution of interarrival time, cpu-demand and io-counts were used. The comparison indicates that our method produces reasonable approximations of upper quantiles, with the exception of the 95^{th} percentile. It should be noted that in this case

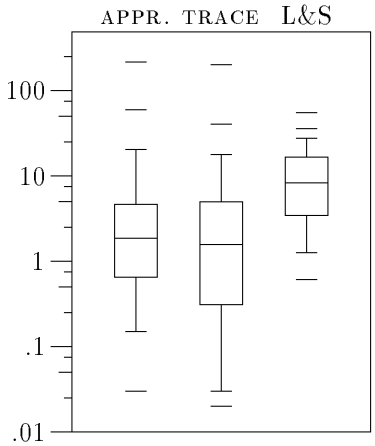

Figure 7: Summary of Response Times in Case Study

the approximations have two additional sources of error. Firstly, the actual distributions of service demands are approximated by Coxian distributions. Secondly, the actual joint distribution of service demands were approximated as

$$\mathcal{L}_D(s_0, s_1, s_2) = \frac{\mathcal{L}_0(s_0)\,[\pi_3(\pi_1\mathcal{L}_1(s_1) + \pi_2\mathcal{L}_2(s_2)) + 1 - \pi_1 - \pi_2]}{1 - (1-\pi_3)\mathcal{L}_0(s_0)\,[\pi_1\mathcal{L}_1(s_1) + \pi_2\mathcal{L}_2(s_2)]} .$$

4 Non-Product Form Models

We have introduced a new method to approximate the response-time distribution in queueing network models. The method was initially derived for network models having the product-form solution. In this section we discuss how our method can be used in situations where the network does not have the product-form solution. We restrict the discussion to two topics: *routing behavior* and *non-exponential service-time distribution at a FCFS-node*.

The joint z-transform for the numbers of visits in different nodes was constructed from the Markovian one-step transition probabilities. In many practical situations, a single Markovian matrix does not capture the actual routing behavior accurately enough. In [KoRe75] it was shown that many discrete distributions can be approximated quite accurately as mixtures of geometric distributions.

Using a similar approach we can capture an essentially wider range of routing behaviors: at the arrival instant to the system a customer selects one of the one-step Markovian transition matrices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_C$ according to probabilities $\beta_1, \ldots, \beta_C$. Now, the overall joint z-transform is given by

$$\mathcal{N}(z_1, \ldots, z_m) = \sum_{c=1}^{C} \beta_c \mathcal{N}_c(z_1, \ldots, z_m) ,$$

where $\mathcal{N}_c$ is the joint z-transform corresponding the Markovian matrix $\boldsymbol{P}_c$.

The approach can be interpreted as if we divide a class into subclasses so that all the subclasses have the same service-time distributions but different branching probabilities. The construction of approximate residence times do not become more difficult, since we need the queue-length distributions only over all the classes.

A common situation in real-life modelling is that the service-time distribution over all classes is not exponential in a FCFS-server. The reason can be that all the classes do not have the same exponential distribution and/or at least one class has not an exponential service-time distribution. In many cases, the overall service-time distribution can be approximated quite accurately by a Coxian distribution, i.e. as a mixture of exponential distributions. Many approximative methods can be used to approximate the arrival queue-length distribution in such situation. Reviews of available methods can be found, for example, in [LaSa83] and [HeLa84].

Let $\mathcal{L}(s)$ and $\mathcal{L}_c(s)$ denote the Laplace transform of the overall service-time distribution and the service time distribution of the analyzed class. Now the residence-time distribution is approximated as

$$\mathcal{L}_r(s) \approx \mathcal{L}_c(s) \sum_{k=0} q_k \mathcal{L}(s)^k .$$

Summary

We have shown that the proposed new method approximates the response-time distribution satisfactorily in many cases. Unlike the previous methods the proposed one takes into account the service-time distributions and routing behaviour. This is the strength of the method but detailed workload specification is also required. Especially the service-demand distributions should be characterized quite accurately.

The numerical inversion of the Laplace transform is the "working horse" of the method. In general, the numerical inversion is an ill-posed problem but most of the numerical problems are avoided, since the cumulative distribution function is usually smooth and bounded. We have met numerical problems only to evaluate the probability that the response time is less a very small value, say 0.01.

The need of the joint probability generating function for the numbers of visits in different nodes limits the topology of network models that can be solved. The z-transform has a simple form only in serial, tree and central-server models. However, using the decomposition technique most network models of practical interest can be expressed as a hierarchical model involving only them.

Acknowledgements

The ideas and suggestions that arose in discussions with Dr. Timo Alanko during the reasearch have been very valuable and stimulating. The author is also thankful to Prof. Domenico Ferrari, Dr. Matt Mutka and an unknown referee for their comments.

References

[Crum76] Crump, K. S. 1976. "Numerical Inversion of Laplace Transforms Using a Fourier Series Approximation." *Journal of the ACM* **22**: 89–96.

[FaIM83] Fayolle, G.; R. Iasnogorodski and I. Mitrani. 1983. "The Distribution of Sojourn Times in a Queueing Network with Overtaking: Reduction to a Boundary Problem". In *Performance '83*. Eds. A. K. Agrawala and S. K. Tripathi. North-Holland, Amsterdam: 477–486.

[HeLa84] Heidelberger, P. and S. S. Lavenberg. 1984. "Computer Performance Evaluation Methodology." *IEEE Transactions on Computers* **C-33**: 1195–1220.

[HeWe81] 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.

[KePo83] Kelly, F. P. and P. K. Pollet. 1983. "Sojourn Times in Closed Queueing Networks". *Advances in Applied Probability* **15**: 638–656.

[Klei75] Kleinrock, L. 1975. *Queueing Systems — Volume I: Theory*. Wiley, New York, N.Y.

[KoRe75] Kobayashi, H. and M. Reiser. 1975. "On Generalization of Job Routing Behaviour in a Queueing Network Model." IBM Research Report RC-5252. IBM T. J. Watson Research Center, Yorktown Heights, N.Y. (Feb.).

[LaRe80] Lavenberg, S. S. and M. Reiser. 1980. "Stationary State Probabilities at Arrival Instants for Closed Queueing Networks with Multiple Type of Customers". *Journal of Applied Probability* **17**: 1048–1061.

[LaSa83] Lavenberg, S. S. and C. H. Sauer. 1983. "Approximate Analysis of Queueing Networks." In *Computer Performance Modeling Handbook*. Ed. S. S. Lavenberg. Academic Press, New York, N.Y.: 173–221.

[LaSe79] Lazowska, E. D. and K. C. Sevcik. 1979. "Exploiting Decomposability to Approximate Quantiles of Response Times in Queueing Networks". In *Operating Systems: Theory and Practice*. Ed. D. Lanciaux. North-Holland, Amsterdam: 149–166.

[McKe87] McKenna, J. 1987. "Asymptotic Expansion of the Sojourn Time Distribution Functions of Jobs in Closed, Prodect-Form Queueing Networks". *Journal of the ACM* **34**: 985–1003.

[Mitr81] Mitra, A. B. 1981. "Waiting Time Distributions from Closed Queueing Network Models of Shared Processor Systems". In *Performance '81*. Ed. F. J. Kylstra. North-Holland, Amsterdam: 113–131.

[NAG87] The Numerical Algorithms Group Limited. 1987. *NAG Fortran Library Manual — Mark 12, Volume 1*. Numerical Algorithms Group, Oxford, U.K.

[ReSe88] Rege, K. M. and B. Sengupta. 1988. "Response Time Distribution in a Multiprogrammed Computer with Terminal Traffic." *Performance Evaluation* **8**: 41–50.

[Robe77] Roberts, F. D. K. 1977. "Minimal Degree Rational Approximation". *Rocky Mountain Journal of Mathematics* **7**: 659–666.

[SaLa81] Salza, S, and S. S. Lavenberg. 1981. "Approximating Response Time Distributions in Closed Queueing Network Models of Computer Performance". In *Performance '81*. Ed. F. J. Kylstra. North-Holland, Amsterdam: 133–145.

[StBu80] Stoer, J. and R. Bulirsch. 1980. *Introduction to Numerical Analysis*. Springer-Verlag, New York, N.Y.

[VeHo81] Velleman, P. F. and D. C. Hoaglin. 1981. *Applicattions, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press, Boston, Mass.