Realization Probability and Throughput Sensitivity in a Closed Jackson Network

Article *in* Journal of Applied Probability · September 1989

DOI: 10.2307/3214418

Xi-Ren Cao
The Hong Kong University of Science and Technology

J. Appl. Prob. **26**, 615–624 (1989)
Printed in Israel

REALIZATION PROBABILITY AND THROUGHPUT SENSITIVITY IN A CLOSED JACKSON NETWORK

XI-REN CAO,* *Digital Equipment Corporation*

Abstract

Realization probability is a new concept pertaining to perturbation analysis of closed queuing networks. The sensitivities of throughputs in a closed single-class Jackson network can be expressed in terms of realization probabilities. In this paper, based on a discussion of perturbation analysis for networks with state-dependent service rates, we derive some new formulas for sensitivities of throughputs using realization probability.

PERTURBATION ANALYSIS; CLOSED QUEUING NETWORKS

1. Introduction

A new technique, called perturbation analysis of queuing networks, has been proposed recently (Ho and Cao (1983)). Peturbation analysis takes a dynamic viewpoint and studies how a fictitiously introduced perturbation of an event affects other events as the system evolves. Using perturbation analysis one can obtain the sensitivity of performance based on one sample path of a queuing network. It has been proved that the derivative of the system throughput with respect to the mean service time in a closed Jackson network based on one sample path is a strongly consistent estimate of that of the expected throughput (Cao (1988a)).

The concept of realization probability in a closed single-class Jackson network was first introduced in Ho and Cao (1983) and was rigorously studied in Cao (1987). This concept plays an important role in perturbation analysis. Based on the concept, it has been proved that the perturbation analysis estimate of the elasticity of the system throughput with respect to the mean service time (i.e., the elasticity based on one sample path) converges with probability 1 to that of the steady state throughput. An equation was derived which expresses the sensitivity of the steady state throughput in terms of the realization probabilities and the steady state probabilities. Unfortunately, these results usually do not hold for multiclass customer systems (Cao (1988b)).

The objective of this paper is to address some issues concerning realization probability which are related to perturbation analysis of systems with state-dependent service rates

Received 23 May 1988; revision received 22 September 1988.
* Postal address: Digital Equipment Corporation, MRO1-1/L26, 200 Forest Street, Marlboro, MA 01752, USA.
This work was initiated when the author was with the Division of Applied Sciences, Harvard University, Cambridge, MA 02138.

and to derive some new equations for the sensitivities of throughput. In Section 2, we first give a brief review and an intuitive explanation of realization probability. Then we discuss some issues related to perturbation analysis of state-dependent systems. In Section 3, we give some new formulas for realization probabilities. Especially, a formula which calculates the sensitivity of throughput with respect to the number of customers in a single-class closed Jackson network using realization probabilities is derived.

2. Realization probability and sensitivity

Consider a closed Jackson network consisting of M single-server nodes and N single-class customers. The service discipline of every server is first come, first served. The service time of a customer at server i is exponentially distributed with mean $\bar{s}_i = 1/\mu_i$, $1 \leqq i \leqq M$, and can be determined by the following equation:

$$s = -\bar{s}_i \ln(1-\xi) \tag{1}$$

where ξ is a random number uniformly distributed on [0, 1). The routing probabilities of the system are $q_{i,j}$, $i, j = 1, 2, \cdots, M$.

First let us briefly review perturbation analysis and realization probability. Suppose that the mean service time of server i, $\bar{s}_i$, changes to $s_i + \Delta\bar{s}_i$, where $\Delta\bar{s}_i$ is a small real number. Then the service time of the customer at server i will change to

$$s + \Delta s = -(\bar{s}_i + \Delta\bar{s}_i)\ln(1-\xi). \tag{2}$$

Comparing Equations (1) and (2), we get

$$\Delta s = -\Delta\bar{s}_i \ln(1-\xi) = s\,\frac{\Delta\bar{s}_i}{\bar{s}_i} \approx -s\,\frac{\Delta\mu_i}{\mu_i}. \tag{3}$$

This Δs is called the *perturbation generated* in the service duration of the customer. Because of this change of the service time of the customer, the service completion time of this customer at server i will be delayed by Δs. We say that server i obtains a perturbation Δs at this service completion time.

As the system evolves, a perturbation will be propagated to other servers according to the following propagation rules: (i) A server will keep its perturbation until it meets an idle period; (ii) If server i meets an idle period and after that it receives a customer from server j, then server i will have the same perturbation as server j after this idle period. In this case we say that the perturbation in server j is *propagated* to server i through an idle period. Note that if before an idle period server i has a perturbation Δ and server j has a zero perturbation (i.e., no perturbation), then this zero perturbation at server j will be propagated to server i through the idle period. We say that the perturbation Δ at server i is lost.

The state of the system can be denoted as $\boldsymbol{n} = (n_1, n_2, \cdots, n_M)$, where n_i is the number of customers in server i. Now we study the evolution of a perturbation. A perturbation may be propagated from one server to other servers. Meanwhile, some servers may lose the perturbation which they obtained previously. We use $V \subset \{1, 2, \cdots, M\}$ to denote the set of integers such that server i has the perturbation if and only if $i \in V$. The state of

the system with a perturbation can be denoted by $(\boldsymbol{n}, V)$. Of course, in a perturbed queuing system both $\boldsymbol{n}$ and V depend on time t.

Note that $\Gamma = \{1, 2, \cdots, M\}$ and the null set Φ are two absorbing states of $V(t)$. That is, if at some time t, $V(t) = \Gamma$ (or Φ), then $V(t') = \Gamma$ (or Φ) for all $t' > t$. Suppose that at time $t_0 = 0$, the system is in a perturbed state $(\boldsymbol{n}, V)$. If at some time $t > 0$, $V(t) = \Gamma$, (or Φ), we say that the perturbation in $(\boldsymbol{n}, V)$ is *realized* (or *lost*) by the system. It was proved in Cao (1987) that in a closed Jackson network with irreducible transition matrix $Q = [q_{i,j}]$, a perturbation will, with probability 1, be either realized or lost. Such a network is called an *irreducible network*. The probability that a perturbation in $(\boldsymbol{n}, V)$ is realized in an irreducible network is called the *realization probability* of the perturbation and is denoted as $f(\boldsymbol{n}, V)$. If $V = \{i\}$, we simply write the realization probability as $f(\boldsymbol{n}, i)$. The following properties were proved in Cao (1987):

$$f(\boldsymbol{n}, i) = 0, \quad \text{if } n_i = 0; \tag{4}$$

$$\text{if } V_1 \cap V_2 = \Phi, \text{ and } V_1 \cup V_2 = V_3, \text{ then } f(\boldsymbol{n}, V_1) + f(\boldsymbol{n}, V_2) = f(\boldsymbol{n}, V_3); \tag{5}$$

$$\sum_{i=1}^{M} f(\boldsymbol{n}, i) = 1. \tag{6}$$

The realization probabilities satisfy the following equations:

$$\begin{aligned} \left\{\sum_{i=1}^{M} \varepsilon(n_i)\mu_i\right\} f(\boldsymbol{n}, k) = {} & \sum_{i=1}^{M} \sum_{j=1}^{M} \varepsilon(n_i)\mu_i q_{i,j} f(\boldsymbol{n}_{i,j}, k) \\ & + \sum_{j=1}^{M} \{1 - \varepsilon(n_j)\}\mu_k q_{k,j} f(\boldsymbol{n}_{k,j}, j), \qquad n_k > 0, \quad k \in \Gamma \end{aligned} \tag{7}$$

where $\boldsymbol{n}_{i,j} = (\cdots, n_i - 1, \cdots, n_j + 1, \cdots)$ is a neighboring state of $\boldsymbol{n}$, and $\varepsilon(n) = 1$ if $n > 0$, $\varepsilon(n) = 0$ if $n = 0$. It has also been proved that

$$\frac{\mu_i}{TP(j)} \frac{\partial TP(j)}{\partial \mu_i} = \sum_{\text{all } \boldsymbol{n}} f(\boldsymbol{n}, i)\, p(\boldsymbol{n}), \tag{8}$$

where $p(\boldsymbol{n})$ is the steady state probability of $\boldsymbol{n}$, and $TP(j) = \Sigma_{n_j > 0}\, p(\boldsymbol{n})\mu_j$ is the steady state throughput of server j. Note that the elasticity does not depend on j. In the remainder of the paper we simply write TP for the throughput of a server in similar equations.

To understand the relationship between realization probability and the sensitivity of throughput, it is necessary to discuss perturbation analysis of state-dependent systems. In Equation (8), $p(\boldsymbol{n})$ is proportional to the perturbation generated during the period in which the system is in state $\boldsymbol{n}$, and $f(\boldsymbol{n}, i)$ is the probability that this perturbation will be realized. Intuitively, each term on the right-hand side of Equation (8) contributes to the elasticity of TP a portion that is due to the change in the service time at state $\boldsymbol{n}$. To be more specific, we consider the following situation. Suppose that in a closed single-class Jackson network the mean service rate of server i changes only when the system state is $\boldsymbol{n}$. That is, if the state is $\boldsymbol{n}$ then the mean service rate of server i changes from μ_i to

$\mu_i + \Delta\mu_{i,\boldsymbol{n}}$, otherwise it remains to be μ_i. The derivative with respect to this change is denoted as $\partial TP/\partial\mu_{i,\boldsymbol{n}}$. In this case we can prove an equation similar to Equation (8):

$$\left\{\frac{\mu_{i,\boldsymbol{n}}}{TP}\frac{\partial TP}{\partial\mu_{i,\boldsymbol{n}}}\right\}_{\mu_{i,\boldsymbol{n}}=\mu_i} = p(\boldsymbol{n})f(\boldsymbol{n}, i). \tag{9}$$

The subscript ($\mu_{i,\boldsymbol{n}} = \mu_i$) implies that the value in the braces is taken at the point $\mu_{i,\boldsymbol{n}} = \mu_i$.

A rigorous proof of this equation is essentially the same as that of Equation (8), which was given in Cao (1987). Here we just provide an intuitive explanation. Let T be the length of an observation period. Suppose that in the observation period $[0, T]$, a server serves K customers, then the throughput is $TP = K/T$. If the service rate of server i at state $\boldsymbol{n}$ is perturbed to $\mu_i + \Delta\mu_{i,\boldsymbol{n}}$, then according to an equation similar to Equation (3), there is a total amount $-Tp(\boldsymbol{n})(\Delta\mu_{i,\boldsymbol{n}}/\mu_i)$ of perturbation generated during $[0, T]$, where $Tp(\boldsymbol{n})$ is the length of the period during which the system state is $\boldsymbol{n}$. Among these perturbations, only $-Tp(\boldsymbol{n})(\Delta\mu_{i,\boldsymbol{n}}/\mu_i)f(\boldsymbol{n}, i)$ will be realized by the system. Thus, we can assume that $\Delta T = -Tp(\boldsymbol{n})(\Delta\mu_{i,\boldsymbol{n}}/\mu_i)f(\boldsymbol{n}, i)$ is the perturbation of T. This expression does not take into consideration that some perturbations generated near the end of the period $[0, T]$ may not have been realized or lost at time T. However, the error caused by this effect is negligible if we choose a sufficiently large T. Therefore, the throughput in a system with rate $\mu_i + \Delta\mu_{i,\boldsymbol{n}}$ is $TP + \Delta TP = K/(T + \Delta T)$. Thus we have

$$\Delta TP = \frac{K}{T+\Delta T} - \frac{K}{T} \approx \frac{K}{T}\left(1 - \frac{\Delta T}{T}\right) = -TP\frac{\Delta T}{T}.$$

From this, we have

$$\frac{\mu_i}{TP}\frac{\Delta TP}{\Delta\mu_{i,\boldsymbol{n}}} = -\frac{\mu_i}{T}\frac{\Delta T}{\Delta\mu_{i,\boldsymbol{n}}} = p(\boldsymbol{n})f(\boldsymbol{n}, i).$$

It is important to note that the perturbed system with rate $\mu_{i,\boldsymbol{n}} = \mu_i + \Delta\mu_{i,\boldsymbol{n}}$ is no longer a state-independent Jackson network. The perturbation propagation rules in a system with state-dependent rates are more complicated than that for a state-independent system (Ho and Yang (1986)). As an example, consider the case where at time t_1 the system jumps from state $\boldsymbol{n}_1$ to state $\boldsymbol{n}_2$. Assume that this jump is due to a customer transition from server j to server k. Note that servers j and k may both differ from server i. Suppose that the service completion time t_1 of server j gets a perturbation Δt_1. Although in this case server i does not meet any idle periods at time t_1, the service completion time of the customer in server i will also change, because the service rate of server i in $[t_1, t_1 + \Delta t_1]$ changes from $\mu_{i,\boldsymbol{n}_2}$ to $\mu_{i,\boldsymbol{n}_1}$. This example indicates that it is not straightforward to extend the concept of realization probability $f(\boldsymbol{n}, i)$ and Equations (7), (8), and (9) to systems with state-dependent service rates.

More specifically, let t_s be the service starting time of a customer at server i. The system enters state $\boldsymbol{n}_1$ at t_s. With service rate $\mu_{i,\boldsymbol{n}_1}$, the service completion time of the customer is expected to be

(10) $$t_{c0} = t_s + \frac{r}{\mu_{i,n_1}},$$

where r is the amount of service required by the customer from server i. Now suppose that at t_1 the system jumps to state n_2 because of a customer transition. Denote the amount of the residual service requirement of the customer from server i at t_1 as r_1. Then Equation (10) is equivalent to

(11) $$t_{c0} = t_1 + \frac{r_1}{\mu_{i,n_1}},$$

where t_{c0} is the service completion time of the customer if after t_1 the service rate of server i were still μ_{i,n_1}. However, since the service rate changes to μ_{i,n_2} after t_1, the expected service completion time at t_1 is now

$$t_c = t_1 + \frac{r_1}{\mu_{i,n_2}} = t_1 + \frac{\mu_{i,n_1}}{\mu_{i,n_2}}(t_{c0} - t_1).$$

Let $\theta = \mu_{i,n_1}/\mu_{i,n_2}$, then we have (Ho and Yang (1986)):

$$t_c = (1-\theta)t_1 + \theta t_{c0}.$$

From this, we get

(12) $$\Delta t_c = (1-\theta)\Delta t_1 + (t_{c0} - t_1)\Delta\theta + \theta \Delta t_{c0}.$$

This equation can be written in a form which clearly reveals the perturbation generation and propagation rules in a state-dependent system. Note that Δt_c in Equation (12) is the perturbation of the service completion time of the customer in server i if the service rate is always μ_{i,n_2} after t_1, and Δt_1 is the perturbation of the state transition time t_1. Also,

(13) $$\Delta\theta = \frac{\mu_{i,n_2}\Delta\mu_{i,n_1} - \mu_{i,n_1}\Delta\mu_{i,n_2}}{\mu_{i,n_2}^2},$$

and

(14) $$\Delta t_{c0} = \Delta t_s - \frac{r}{\mu_{i,n_1}^2}\Delta\mu_{i,n_1} = \Delta t_s - \frac{t_{c0} - t_s}{\mu_{i,n_1}}\Delta\mu_{i,n_1}.$$

Substituting Equations (10), (11), (13) and (14) into Equation (12), after some algebraic calculations, we get

(15) $$\Delta t_c = (1-\theta)\Delta t_1 + \theta\left[\Delta t_s - (t_1 - t_s)\frac{\Delta\mu_{i,n_1}}{\mu_{i,n_1}}\right] - \left(\frac{r_1}{\mu_{i,n_2}}\right)\frac{\Delta\mu_{i,n_2}}{\mu_{i,n_2}}.$$

Equation (15) shows that the perturbation of the service completion time Δt_c consists of three parts. The first part is $(1-\theta)\Delta t_1$, which reflects the effect of the change in Δt_1. The second part is $\theta[\Delta t_s - (t_1 - t_s)(\Delta\mu_{i,n_1}/\mu_{i,n_1})]$, where Δt_s is the perturbation of the starting time of the service, and $-(t_1 - t_s)(\Delta\mu_{i,n_1}/\mu_{i,n_1})$ is the perturbation generated in $[t_s, t_1]$. Note that because the service rate of server i changes at time t_1 from μ_{i,n_1} to μ_{i,n_2}, the perturbation accumulated up to t_1 has to be modified by a weighting factor θ. The

third part is $-(r_1/\mu_{i,\boldsymbol{n}_2})(\Delta\mu_{i,\boldsymbol{n}_2}/\mu_{i,\boldsymbol{n}_2})$, which is the perturbation generated after t_1. In the expression, $(r_1/\mu_{i,\boldsymbol{n}_2})$ is the time required to complete the service of the customer after t_1 if the service rate is $\mu_{i,\boldsymbol{n}_2}$ afterwards. If, however, at some time $t_2 \in [t_1, t_c]$ the system jumps to another state, then at t_2 the perturbation has to be modified according to the same rules explained above. Clearly, Equation (15) describes the perturbation analysis rules for systems with state-dependent service rates. The generation rule is the same as the state-independent case. But at each customer transition time the perturbation of each server has to be modified by a weighting factor θ, and an additional perturbation $(1-\theta)\Delta t_1$ has to be added.

Now turn back to realization probability. Since in Equation (9) we assume that the service rate at server i changes only when the system is in state $\boldsymbol{n}_1$, we should apply the general rules shown in Equation (15) to do perturbation propagation. In a state-independent Jackson network $\mu_{i,\boldsymbol{n}_1} = \mu_{i,\boldsymbol{n}_2} = \mu_i$. Thus $\theta = 1$. If we assume that μ_i changes only when the state is $\boldsymbol{n}_1$, then $\Delta\mu_{i,\boldsymbol{n}_2} = 0$, and Equation (15) becomes:

$$\Delta t_c = \Delta t_s - (t_1 - t_s)\,\frac{\Delta\mu_{i,\boldsymbol{n}_1}}{\mu_{i,\boldsymbol{n}_1}}\ .$$

This equation shows that the perturbation of the service completion time is the sum of that of the service starting time and the perturbation generated in the period during which the system is in state $\boldsymbol{n}_1$. This is just the same as the state-dependent case. Modifications of perturbation propagation are not required when we apply perturbation analysis to systems in which service rate changes only if the system is in certain states.

3. Some new formulas

Suppose that in a closed Jackson network the service rate of server i changes only when there are $n_i = j$ customers in server i. The derivative with respect to this change is denoted as $\partial TP/\partial\mu_{i,j}$. By an argument similar to that for Equation (9), we have

$$\left\{\frac{\mu_{i,j}}{TP}\,\frac{\partial TP}{\partial\mu_{i,j}}\right\}_{\mu_{i,j}=\mu_i} = \sum_{n_i=j} p(\boldsymbol{n})f(\boldsymbol{n}, i). \tag{16}$$

Let $\bar{s}_{i,j} = 1/\mu_{i,j}$ and $\bar{s}_{i,j} = \alpha_{i,j}\bar{s}_i$. Then the above equation becomes

$$\left\{\frac{1}{TP}\,\frac{\partial TP}{\partial\alpha_{i,j}}\right\}_{\alpha_{i,j}=1} = \sum_{n_i=j} p(\boldsymbol{n})f(\boldsymbol{n}, i). \tag{17}$$

It was proved in Suri (1983) that

$$\left\{\frac{1}{TP}\,\frac{\partial TP}{\partial\alpha_{i,j}}\right\}_{\alpha_{i,j}=1} = p_{N-1}(n_i \geqq j) - p_N(n_i \geqq j), \tag{18}$$

where $p_N(n_i \geqq j)$ is the probability of $(n_i \geqq j)$ in a closed Jackson network with N total customers in the network. From Equations (17) and (18), we have

$$\sum_{n_i=j} p(\boldsymbol{n})f(\boldsymbol{n}, i) = p_N(n_i \geqq j) - p_{N-1}(n_i \geqq j). \tag{19}$$

Note that $\Sigma_{j=1}^{N} p_N(n_i \geqq j) = Q_N(i)$, where $Q_N(i)$ is the average number of customers in server i in a closed Jackson network with N total customers. From (19), it is easy to prove

$$\sum_{\text{all } \boldsymbol{n}} p(\boldsymbol{n}) f(\boldsymbol{n}, i) = Q_N(i) - Q_{N-1}(i). \tag{20}$$

From this we can further obtain

$$\sum_{i=1}^{M} \sum_{\text{all } \boldsymbol{n}} p(\boldsymbol{n}) f(\boldsymbol{n}, i) = 1. \tag{21}$$

This equation can be verified by using Equation (6). Note that $TP_N(i) = p_N(n_1 \geqq 1)\mu_i$ is the steady state throughput of server i in a closed Jackson network with N total customers. Letting $j = 1$ in (19), we get

$$\Delta TP = TP_N(i) - TP_{N-1}(i) = \mu_i \left\{ \sum_{n_i=1} p(\boldsymbol{n}) f(\boldsymbol{n}, i) \right\} \tag{22}$$

where ΔTP is the increment of the throughput of server i when the number of customers in a Jackson network increases from $N - 1$ to N. An interesting feature of this equation is that it establishes a relationship between realization probability, which measures the sensitivity with respect to infinitesimal changes of a continuous parameter, and the throughput increment because of the change of a discrete parameter.

To further explain the relationship between realization probabilities and the changes in throughput resulting from increment of the population size and to illustrate the application of Equation (22) to a real system, it is necessary to describe the algorithm which implements the right-hand side of Equation (22) based on a trajectory of a queuing system. The algorithm is similar to the one for $\Sigma_{\text{all } \boldsymbol{n}} p(\boldsymbol{n}) f(\boldsymbol{n}, i)$ in Ho and Cao (1983), except that perturbations are generated at server i only when $n_i = 1$. In the algorithm, there is an accumulator A_j associated with server j, $j = 1, 2, \cdots, M$. The value of these A_j's are set to be zero initially.

Algorithm

Step 1. Each time the system reaches a state $\boldsymbol{n}$ with $n_i = 1$, we add the sojourn time of the system in this state to A_i.

Step 2. If server k meets an idle period and after the idle period it receives a customer from server j, then we copy the content of A_j to A_k.

Step 3. At the end of the trajectory, we calculate

$$\sum_{n_i=1} p(\boldsymbol{n}) f(\boldsymbol{n}, i) = \frac{A_i}{T}$$

where T is the length of the trajectory.

The above algorithm offers an intuitive explanation of Equation (22). Note that if there is more than one customer in server i (i.e., $n_i > 1$), then the service status of server i (i.e., busy or idle) will not be affected if the population size changes from N to $N - 1$. Only if $n_i = 1$, then server i may become idle if the system contains one customer less. In Step 1 of the algorithm, the lengths of the sojourn times of all states $\boldsymbol{n}$ with $n_i = 1$ are

accumulated in A_i. This accumulated value represents the decrease in busy time of server 1 if server 1 always contains one customer less. Step 2 reflects the effect of perturbation propagation. Equation (22) simply shows that after propagation the final perturbation due to one customer less at server i is in fact equal to the busy-time difference between the two systems with population size N and $N-1$.

The algorithm can be applied to any trajectory generated by simulation or observed from a real system such as a manufacturing system or computer network. Note that the algorithm requires very little computation and can therefore be used for real-time decision making. In a manufacturing system a customer represents a pallet on which a part to be processed by a set of machines is fixed. In a computer network, a customer represents a program or a batch job. Applying the algorithm to a trajectory obtained by observing the system tells the system manager what will be the effect on system throughput if one pallet is removed or the multi-programming level is decreased by one.

In summary, realization probability satisfies Equations (4)–(7), and the sensitivities of throughputs can be expressed in terms of realization probability (Equations (8), (9), (16), and (22)).

Example 1. Consider a closed Jackson network in which $\mu_i q_{i,j} = \mu_j q_{j,i}$. It was shown in Cao (1987) that the solution to Equations (4)–(7) for this system is

$$f(\boldsymbol{n}, i) = \frac{n_i}{N}.$$

Also,

$$p_N(n_i = 1) = \frac{N(M-1)}{(M+N-1)(M+N-2)}.$$

Thus, according to Equation (22) we have

$$\begin{aligned} \Delta TP &= \mu_i \left\{ \sum_{n_1=1} p(\boldsymbol{n}) f(\boldsymbol{n}, i) \right\} = \mu_i \left\{ \frac{1}{N} \sum_{n_1=1} p(\boldsymbol{n}) \right\} \\ &= \mu_i \left\{ \frac{1}{N}\, p_N(n_i = 1) \right\} \qquad (23) \\ &= \mu_i \left\{ \frac{(M-1)}{(M+N-1)(M+N-2)} \right\}. \end{aligned}$$

On the other hand, by the Jackson formula we have

$$TP_N(i) = \mu_i [1 - p_N(n_i = 0)] = \mu_i \left(\frac{N}{M+N-1} \right).$$

ΔTP in (23) obtained by realization probability is exactly the same as $TP_N(i) - TP_{N-1}(i)$ given by the Jackson formula.

Example 2. In this example we shall consider a system with $M = 3$, $N = 5$, $\bar{s}_1 = 10$, $\bar{s}_2 = 8$, $\bar{s}_3 = 5$, and $q_{1,1} = q_{2,2} = q_{3,3} = 0$, $q_{1,2} = q_{1,3} = 0.5$, $q_{2,1} = 0.8$, $q_{2,3} = 0.2$, $q_{3,1} = 0.3$,

TABLE 1

$\boldsymbol{n}$	$p(\boldsymbol{n})$	$f(\boldsymbol{n}, 1)$	$f(\boldsymbol{n}, 2)$	$f(\boldsymbol{n}, 3)$
(5, 0, 0)	0.19047	1.00000	0.00000	0.00000
(4, 0, 1)	0.06644	0.90584	0.00000	0.09416
(4, 1, 0)	0.15061	0.89385	0.10615	0.00000
(3, 0, 2)	0.02318	0.78826	0.00000	0.21174
(3, 1, 1)	0.05254	0.77060	0.17336	0.05604
(3, 2, 0)	0.11908	0.74279	0.25721	0.00000
(2, 0, 3)	0.00809	0.62556	0.00000	0.37444
(2, 1, 2)	0.01833	0.60901	0.25029	0.14070
(2, 2, 1)	0.04154	0.58286	0.37574	0.04141
(2, 3, 0)	0.09416	0.54528	0.45472	0.00000
(1, 0, 4)	0.00282	0.38089	0.00000	0.61911
(1, 1, 3)	0.00639	0.37327	0.34810	0.27863
(1, 2, 2)	0.01449	0.35728	0.51926	0.12346
(1, 3, 1)	0.03285	0.33315	0.62079	0.04606
(1, 4, 0)	0.07445	0.29754	0.70246	0.00000
(0, 0, 5)	0.00098	0.00000	0.00000	1.00000
(0, 1, 4)	0.00223	0.00000	0.48951	0.51049
(0, 2, 3)	0.00505	0.00000	0.71510	0.28490
(0, 3, 2)	0.01146	0.00000	0.83485	0.16515
(0, 4, 1)	0.02597	0.00000	0.91819	0.08181
(0, 5, 0)	0.05887	0.00000	1.00000	0.00000

and $q_{3,2} = 0.7$. The steady-state probabilities are numerically calculated using the Jackson formula, and the realization probabilities are obtained by solving Equations (4)–(7). The results are listed in Table 1. Using Equation (22) and the data in the table, we get

$$\Delta TP(1) = 0.00417.$$

The throughputs of the system and another system with $N - 1 = 4$ customers are also calculated using the Jackson formula. The values are $TP_5(1) = 0.08954$ and $TP_4(1) = 0.08537$. It is clear that $\Delta TP(1) = TP_5(1) - TP_4(1)$.

A simulation program was written for the system, and the algorithm applied to the trajectory produced by a single run. After 10000 service completions we obtained $\Delta TP(1) = 0.0427$, and after 100000 service completions we obtained $\Delta TP(1) = 0.0419$.

Realization probability for systems with multiclass customers is discussed in Cao (1988b). The main point is that, unfortunately, Equations (8), (9), and (16) do not hold for all multiclass customer systems. We do not discuss the details here.

Our last remark is that using realization probability one does not need to use the Jackson formula for the system with $N - 1$ customers. The practical significance is: the information of ΔTP is completely included in a trajectory of the system with N customers. One can obtain both $p(\boldsymbol{n})$ and $f(\boldsymbol{n}, i)$ by analyzing one trajectory of the system with N customers. This is consistent with the fact that in the algorithm for ΔTP a trajectory for the system with $N - 1$ customers is not needed.

References

CAO, X. R. (1987) Realization probability in closed Jackson queuing networks and its application. *Adv. Appl. Prob.* **19**, 708–738.

CAO, X. R. (1988a) On a sample performance function of closed Jackson networks. *Operat. Res.* **36**, 128–136.

CAO, X. R. (1988b) Realization probability in multi-class closed queuing networks. *European J. Operat. Res.* **36**, 393–401.

HO, Y. C. AND CAO, X. R. (1983) Perturbation analysis and optimization of queueing networks. *J. Optim. Theory Applic.* **40**, 559–582.

HO, Y. C. AND YANG, P. Q. (1986) Equivalent network, load dependent servers and perturbation analysis — an experimental study. In *Teletraffic Analysis and Computer Performance Evaluation*, eds. O. J. Boxma, J. W. Cohen and H. C. Tijms, North-Holland, Amsterdam.

SURI, R. (1983) Robustness of queuing network formulas. *J. Assoc. Comput. Mach.* **30**, 564–594.