M. Reiser
H. Kobayashi

Queuing Networks with Multiple Closed Chains: Theory and Computational Algorithms

Abstract: In this paper a recent result of Baskett, Chandy, Muntz, and Palacios is generalized to the case in which customer transitions are characterized by more than one closed Markov chain. Generating functions are used to derive closed-form solutions to stability, normalization constant, and marginal distributions. For such a system with N servers and L chains the solutions are considerably more complicated than those for systems with one subchain only. It is shown how open and closed subchains interact with each other in such systems. Efficient algorithms are then derived from our generating function representation.

Introduction

For some ten years, the most general class of queuing networks for which an analytical solution was known is that treated by Jackson [1]. However, applications of such queuing networks to modeling of multiprogrammed/multiprocessor computer systems have been drawing increasing attention in the past few years [2, 3]. Noteworthy progress in extending the class of analytically solvable queuing networks has been reported recently by Baskett, Chandy, Muntz, and Palacios [4, 5]. These authors have succeeded in casting into a unified theory previously known but unconnected results such as queue-size distributions for M/M/1 with FCFS, M/G/1 with processors sharing, M/G/∞ queues, preemptive-resume LCFS discipline, and queuing systems with various classes of customers.

The constribution of the present paper is threefold, namely:

1. To generalize the results represented in [5] to the case of several closed subchains and to give a constructive derivation of the product-form solution.
2. To exploit the probability generating function method as a device to obtain a concise representation of such expressions as normalization constants and distributions in the complex network model treated in this paper.
3. To present computationally efficient algorithms for the general class of networks.

Solutions for a queuing network with classes of customers, multiple subchains, and generalized servers

In this section we define the class of queuing network models and present its solution. We generalize the result of [5] to a queuing network in which customer routing transitions are characterized by a Markov chain decomposable into multiple subchains. We also take a more constructive approach than the previous work so that the reader may follow more easily the derivation of the product-form solution. Networks with closed subchains are introduced as a limiting case of a suitably chosen open network. This technique leads to a unified presentation of the final results. In the last section, several aggregate states and their marginal distributions are introduced.

Definition of the queuing network

The queuing system $\mathcal{N}$ is defined in terms of the following parameters: 1) system configuration, 2) routing probabilities, 3) arrival processes, and 4) service rate and work demand distribution and queue discipline at the individual service centers. In more detail:

1. There are N service centers, R classes of customers, and L disjoint routing chains (or subchains). Throughout the rest of the paper, indices n, r, and l refer to service center, class, and subchain, respectively, and

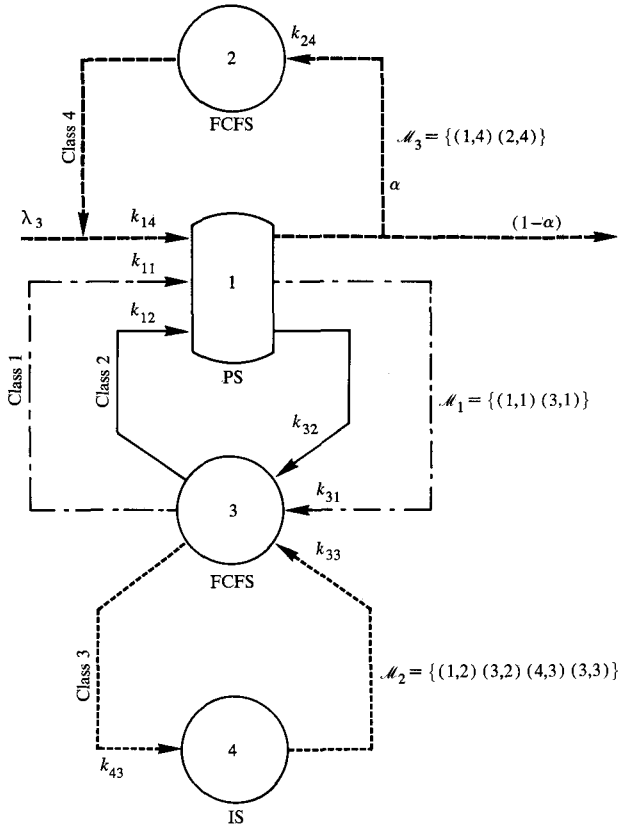


Figure 1 Example of a mixed network with two closed subchains and one open subchain. The routing transition probabilities are $p_{3,(14)} = 1$, $p_{(14),(24)} = \alpha$, $p_{(24),(14)} = 1$, $p_{(14),3} = (1-\alpha)$; $p_{(11),(31)} = 1$, $p_{(31),(11)} = 1$; $p_{(12),(32)} = 1$, $p_{(32),(43)} = 1$, $p_{(43),(33)} = 1$, $p_{(33),(12)} = 1$. All other probabilities are zero. Due to the class change at service center 3, customers proceed through $\mathcal{M}_2$ in a figure eight pattern.

their values range over $n = 1, 2, \cdots, N$, $r = 1, 2, \cdots, R$, and $l = 1, 2, \cdots, L$. We assume $R \geq L \geq 1$ without loss of generality.

2. Jobs (customers) proceed through the network $\mathcal{N}$ according to a first-order Markov chain $\mathcal{M}$. The transition matrix is $NR \times NR$ with elements $p_{(nr),(n'r')}$, which are the probabilities of state transitions $(nr) \rightarrow (n'r')$ in $\mathcal{M}$, namely, the probability that a job of class r completing service at center n will next go to center n' and change its class membership to r'. The Markov chain $\mathcal{M}$ is in general decomposable into L subchains $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_L$, which are all irreducible.
3. The subchains $\mathcal{M}_l$ are either open or closed. *Open subchains* are driven by independent Poisson arrival streams whose rates, λ_l, may be functions of the population size K_l of $\mathcal{M}_l$, i.e., $\lambda_l = \lambda_l(K_l)$; K_l is the number of jobs in $\mathcal{M}_l$ at a given system state $\mathcal{S}$. A newly arriving class r customer of stream l will first join service center n with probability $p_{l,(nr)}$. Similarly, a class r' customer completing service at center n' departs the network with probability $p_{(n'r'),l}$. In a *closed subchain* $\mathcal{M}_l$, the number of jobs is held constant at $K_l{}^*$. Furthermore, $p_{l,(nr)} = p_{(nr),l} = 0$ for all $(nr) \in \mathcal{M}_l$.
4. Service center n is described by a continuous-time Markov chain whose state is given by the ordered set of customers, their associated class membership, and possibly an additional integer (denoting the stage in a cascade of exponential servers at which the customer is). The parameters of this Markov chain may be state-dependent. The following queues have a suitable Markov chain representation:

First-come, first-served (FCFS)
The server has a queue-dependent rate $\mu_n(k_n)$. Each customer has an associated work demand which is assumed to be drawn from an exponential distribution. Note: All classes must have the same distribution. The mean work demand is w_n.

Processor sharing (PS) [6] or preemptive-resume LCFS
The server has a queue-dependent rate $\mu_n(k_n)$. The work demand of customers is assumed to be drawn from a general distribution. Note: The distribution may be a different one for different classes. The mean work demand of a class r job is w_{nr}.

Infinite Server (IS) [7]
Jobs have service (or delay) times which may have a general distribution with mean w_{nr}. Formally we set $\mu_n = 1$.

Note that in our definition the term service rate refers to the actual speed of the server in work units/s. Each job has associated with it a work demand in work units. A job stays at a service center until all work is done. Successive work demands are independent and identically distributed random variables. If we assume constant rates and an FCFS discipline, then the definition of service times is simply work demand divided by rate. It is convenient to express queue-dependent service rates in the form

$$\mu_n(j) = \mu_n{}^0 b_n(j), \tag{1}$$

where $\mu_n{}^0$ is a suitably chosen normalization value and $b_n(j)$ is a dimensionless scaling function. Similarly, for the arrival rate we put

$$\lambda_l(j) = \lambda_l{}^0 a(j). \tag{2}$$

An example of a network in the above class is given in Fig. 1.

• *Relative workload intensity*
The following set of quantities e_{nr} plays an important role in the solution of the network $\mathcal{N}$. The quantities are

defined by L sets of simultaneous linear equations, one set for each subchain $\mathcal{M}_l$:

$$e_{nr} = p_{l,(nr)} + \sum_{(n'r') \in \mathcal{M}_l} e_{nr'} p_{(n'r'),(nr)}, \quad (nr)$$

$$\text{for } n = 1, 2, \cdots, N; \; r = 1, 2, \cdots, R. \qquad (3)$$

Note that $e_{nr} = 0$ if $(nr) \ni \mathcal{M}_l$. A physical interpretation of Eq. (3) is that e_{nr} is the average number of times that the service center n is visited with class membership r by a job which belongs to a subchain $\mathcal{M}_l$. The solution of (3) is uniquely given only if subchain $\mathcal{M}_l$ is open, i.e., $p_{l(nr)} \neq 0$ for at least one $(nr) \in \mathcal{M}_l$, and similarly $p_{(n'r'),l} \neq 0$ for at least one $(n'r') \in \mathcal{M}_l$. If subchain $\mathcal{M}_l$ is closed, then the solution of (3) is determined only up to a constant factor π_l.

We define the *relative workload intensity* by

$$\rho_{nr} = e_{nr} \lambda_l w_{nr} / \mu_n^0 \qquad (4)$$

where μ_n^0 and λ_l^0 are the normalization values of (1) and (2) and l is such that $(nr) \in \mathcal{M}_l$.

• *Outline of the solution*

It is known [8] that if the Laplace transform of a given service time distribution is a rational function, then the distribution can be represented by a set of exponential servers (or stages) combined in serial and parallel manner. The system behavior can be treated as a *birth-and-death process* by introducing an appropriate state space. The equilibrium probability distribution is then determined by a system of linear equations, also known as *balance equations*. These equations are difference equations which relate the steady state probability of a given state with the probabilities of the adjacent states. The size of the state space is such that a numerical solution of the balance equations is impossible for all but the most simple examples.

Solutions of a general nature, therefore, depend on the existence of a so-called *product-form solution*. Such solutions are known to exist only for a restricted class of networks. The notion of *individual balance* or *local balance* [9] is useful in the search for product-form solutions of more and more generality. A brief review of this method is given in the Appendix.

In the "method of stages" representation of general service-time distributions, only one stage can accommodate a job at a given time. Therefore, if the service discipline is FCFS, a customer waiting at the head of the line is not allowed to enter the first stage until the job currently in service completes its last stage and departs from this service center. This is equivalent to saying that the entrance stage is *blocked* as long as a job exists in some stage. Only without blocking, however, does the steady state distribution always take a simple form, i.e., it is given in a product form. Once blocking is introduced, the solution is rather complicated even for the simplest queuing system [10]. This is why we were forced to assume that the service center is a queue-dependent exponential server if that center is under FCFS discipline.

If the queue discipline is either PS or IS, then the problem of blocking in the fictitious exponential servers disappears. In an infinite-server queue, there are always more servers available than jobs and no waiting line develops; thus blocking is nonexistent. A single server with processor sharing is, in effect, an infinite server queue in which service rate is lowered according to the number of jobs in the center, i.e., the service rate of the fictitious exponential server in the individual stage is divided by k_n, which is the number of jobs in this center at a given time. Blocking is not an issue in a PS center either, since no queue exists, just as in an IS center.

A service center under preemptive-resume LCFS can be viewed as consisting of sufficiently many parallel servers, each of which is described as stages of exponential servers. Each time a new job arrives at this station, it immediately enters the first stage of the server provided to it. The job that entered the system just prior to it and has been served by its own server is then *frozen* on the spot. Any job which has been frozen at some stage resumes receiving service when it becomes the youngest among those remaining in the system. Since the product-form solution exists, any newly arrived job enters the service center without being blocked.

Under the IS, PS, and LCFS queue disciplines discussed above, the system state can be completely described by specifying service *stages* of the jobs present in that center.

We now discuss in some detail the solution of a network $\mathcal{N}$ with N FCFS service centers and L open subchains. The state $\mathcal{S}$ of such a network is described by an array of N FCFS stacks, viz.,

$$\mathcal{S} = [\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_N], \qquad (5)$$

with $\mathbf{S}_n = [r_n(1), r_n(2), \cdots, r_n(k_n)]$, where $r_n(j)$ is the class membership of the jth job queuing for service at center n. Let $\mathcal{S}([nr]^-)$ denote a state which is the same as $\mathcal{S}$ except that the last entry of the stack $\mathbf{S}_n$ is missing. Thus a transition $\mathcal{S}([nr]^-) \to \mathcal{S}$ takes place upon arrival of a class r job at service center n.

By applying the principle of local balance, we now equate

rate of transitions $\mathcal{S}([nr]^-) \to \mathcal{S}$ due to arrivals = rate of transitions $\mathcal{S} \to$ other states due to departures of customers

and obtain the simple recurrence equation

$$\lambda(K_l) e_{nr} P\{\mathcal{S}([nr]^-)\} = \mu_n(k_n) P\{\mathcal{S}\}, \qquad (6)$$

where $P\{\mathcal{S}\}$ is the equilibrium probability for state $\mathcal{S}$;

$P\{\mathcal{S}\}$ may now be obtained by applying (6) repeatedly to a sequence of transitions leading from $\mathcal{S}$ down to the empty system $\mathcal{S}_0 = [\mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}]$. This procedure yields

$$P\{\mathcal{S}\} = C\left[\prod_{n=1}^{N} B_n(k_n)\right] \prod_{l=1}^{L} A_l(k_l) \prod_{(nr)\in\mathcal{M}_l} \rho_{nr}{}^{k_{nr}}, \tag{7}$$

where $B_n(j) = \prod_{i=1}^{j} b_n{}^{-1}(i)$, $A_l(j) = \prod_{i=0}^{j-1} a_l(i)$, and $C = P\{\mathcal{S}_0\}$ is determined by normalization.

• *Solution for closed subchains*

If $\mathcal{N}$ is closed with respect to the subchain $\mathcal{M}_l$, then the recurrence equation (6) is not directly applicable since a direct transition $\mathcal{S}([nr]^-) \rightarrow \mathcal{S}$ is not possible. We can, however, view $\mathcal{N}$ as the limiting case of a suitably chosen open network $\mathcal{N}^0$. The solution of $\mathcal{N}$, then, is obtained directly from Eq. (7).

The treatment of $\mathcal{N}$ is further complicated by the fact that for closed subchains $\mathcal{M}_l$, the quantities e_{nr} are defined only up to an arbitrary factor π_l which is reflected by the nonuniqueness of $\mathcal{N}^0$ with respect to $\mathcal{N}$. However, Eq. (7) reveals that π_l appears as $\pi_l{}^{K_l}$ in the solution. Therefore, since $K_l = K_l{}^*$ is a fixed value in $\mathcal{N}^0$, $\pi_l{}^{K^*_l}$ is a constant factor which can be absorbed in the normalization constant.

In summary, we find that the solution of a network with closed subchains is formally the same as that given by Eq. (7) if we set $\lambda_l = 1$ and $A_n(K_l) = \delta(K_l, K_l{}^*) = 1$ if $K_l = K_l{}^*$, 0 otherwise.

• *Aggregate states and marginal distributions*

In many practical cases, we may not necessarily be interested in $P\{\mathcal{S}\}$, the distribution of system state $\mathcal{S}$. Instead, we may want to obtain a marginal distribution, such as the total queue-size distribution. Marginal distributions are, by definition, the probability distribution of *aggregate states*, an aggregate state being a subset of the state space $\mathcal{F}$. The following is a list of important aggregate states:

1. $\mathcal{K} = [\mathbf{k}_1, \mathbf{k}_2, \cdots, \mathbf{k}_N]$ with $\mathbf{k}_n = [k_{n1}, k_{n2}, \cdots, k_{nr}]$ where k_{nr} is the number of class r jobs at center n and the specific orderings of the individual FCFS stacks are ignored.
2. $\mathbf{k} = [k_1, k_2, \cdots, k_N]$ where k_n represents the total number of jobs at center n, i.e., $k_n = \Sigma_r k_{nr}$.
3. k_n; we are interested in the queue-size distribution at a specific center n only.
4. $\mathbf{K} = [K_1, K_2, \cdots, K_L]$ where K_l is the total number of jobs in subchain $\mathcal{M}_l$, i.e.,

$$K_l = \sum_{(nr)\in\mathcal{M}_l} k_{nr}.$$

Since $P\{\mathcal{S}\}$ is invariant to permutations of the elements in the FCFS stacks, we obtain $P\{\mathcal{K}\}$ easily as

$$P\{\mathcal{K}\} = \left[\prod_{n=1}^{N} k_n!\left(\prod_{r=1}^{R} k_{nr}!\right)^{-1}\right] P\{\mathcal{S}\}, \tag{8}$$

where the factorial term represents the total number of distinct permutations and is a product of multinomial coefficients.

• *Summary of the general results for the queuing networks $\mathcal{N}$*

It is an interesting result that at the level of detail described by the aggregate state $\mathcal{K}$, only mean work demands enter into the solution.

The distribution $P\{\mathcal{K}\}$ for the queuing network $\mathcal{N}$ can be expressed in a unified way as

$$P\{\mathcal{K}\} = C \cdot A(\mathbf{K}) \prod_{n=1}^{N} g_n(\mathbf{k}_n, \boldsymbol{\rho}_n), \tag{9}$$

with

$$\boldsymbol{\rho}_n = [\rho_{n1}, \rho_{n2}, \cdots, \rho_{nR}] \text{ and}$$

$$A(\mathbf{K}) = \prod_{l=1}^{L} A_l(K_l), \tag{10}$$

$$g_n(\mathbf{k}_n, \boldsymbol{\rho}_n) = B_n(k_n)k_n! \prod_{r=1}^{R} \frac{\rho_{nr}{}^{k_{nr}}}{k_{nr}!}, \tag{11}$$

$$B_n(j) = \begin{cases} \prod_{i=1}^{j} b_n{}^{-1}(i) & \text{for FCFS, PS, and LCFS;} \\ \frac{1}{j!}\prod b_n{}^{-1}(i) & \text{for IS;} \end{cases} \tag{12}$$

$$A_l(j) = \begin{cases} \prod_{i=0}^{j-1} a_l(i) & \text{for open } \mathcal{M}_l; \\ \delta(j, K_l{}^*) & \text{for closed } \mathcal{M}_l \text{ with } K_l{}^* \text{ customers.} \end{cases} \tag{13}$$

The quantities ρ_{nr} are defined by (4), where we substitute unity for λ_e if $\mathcal{M}_e$ is closed. As defined earlier, k_{nr} is the number of class r jobs at center n and $\mathbf{k}_n = [k_{n1}, k_{n2}, \cdots, k_{nr}]$; k_n is the total number of jobs at center n, K_l is the total number of jobs in subchain $\mathcal{M}_l$, and $K_l{}^*$ is the fixed number of jobs in a closed subchain $\mathcal{M}_l$.

Generating functions

In this section we introduce the probability generating function (p.g.f.) for the queue-size distribution. The p.g.f. method allows a simple evaluation of the normalization constant and of marginal distributions. It also provides important theoretical results, e.g., stability criteria and asymptotic behavior.

The first step is the derivation of the p.g.f. for an open system with constant arrival rates. We obtain simple explicit expressions for the normalization constant and for the marginal distributions. We find that, as far as the queue size distribution is concerned, we can treat the servers as though they were mutually independent and behaved exactly like separate single servers, subject to an equivalent traffic intensity.

Mixed networks, i.e., networks with open and closed subchains, are treated next. An important connection with the p.g.f. for the open system is found via the marginal distribution of subchain populations. Closed subchains interact in a complicated way and we can no longer treat the servers as being mutually independent. Consequently, there exists no simple closed-form expressions for the p.g.f. Nevertheless, we obtain results concerning how open and closed subchains interact with each other and also how to calculate marginal distributions.

In the final part of this section we discuss the stability problem for mixed networks, and we show that the stability is unaffected by the presence of closed chains.

• *Open networks with constant arrivals*

Let $\mathcal{N}$ be open with respect to all subchains $\mathcal{M}_l$ which are driven with separate Poisson streams of constant arrival rate λ_l. We define the p.g.f. for $P\{\mathcal{K}\}$ by

$$G^*(\mathcal{Z}) = \mathrm{E}\left[\prod_{k=1}^{N}\prod_{r=1}^{R} z_{nr}^{\,k_{nr}}\right] = C \cdot G(\mathcal{Z}), \tag{14}$$

where $\mathcal{Z}$ is the array of transformation variables $[\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N]$ with $\mathbf{z}_n = [z_{n1}, z_{n2}, \cdots, z_{nR}]$, $1 \le n \le N$. The improper generating function $G(\mathcal{Z})$ is defined by a sum over the product-form terms, viz.,

$$\begin{aligned} G(\mathcal{Z}) &= \sum_{\mathcal{K}\ge 0}\prod_{n=1}^{N} B_n(k_n)k_n!\prod_{r=1}^{R}\frac{(\rho_{nr}z_{nr})^{k_{nr}}}{k_{nr}!} \\ &= \sum_{\mathcal{K}\ge 0}\prod_{n=1}^{N} g_n(\mathbf{k}_n, [\rho_{nr}z_{nr}]), \end{aligned} \tag{15}$$

where $[\rho_{nr}z_{nr}] = [\rho_{n1}z_{n1}, \rho_{n2}z_{n2}, \cdots, \rho_{n2}z_{nr}]$. The structure of (15) allows us to switch sum and product, yielding $G(\mathcal{Z})$ as a product of N terms $G_n(\mathbf{z}_n, \boldsymbol{\rho}_n)$, namely,

$$G(\mathcal{Z}) = \prod_{n=1}^{N}\sum_{\mathbf{k}_n\ge 0} g_n(\mathbf{k}_n, [\rho_{nr}z_{nr}]) = \prod_{n=1}^{N} G_n(\mathbf{z}_n, \boldsymbol{\rho}_n), \tag{16}$$

where $\boldsymbol{\rho}_n = [\rho_{n1}, \rho_{n2}, \cdots, \rho_{nR}]$. Here $G_n(\mathbf{z}_n, \boldsymbol{\rho}_n)$ is explicity given by

$$G_n(\mathbf{z}_n, \boldsymbol{\rho}_n) = \sum_{\mathbf{k}_n\ge 0} B_n(k_n)k_n!\prod_{r=1}^{R}[(\rho_{nr}z_{nr})^{k_{nr}}/k_{nr}!], \tag{17}$$

which we recognize as a power series in R dimensions. Therefore, we obtain the simple result

$$G_n(\mathbf{z}_n, \boldsymbol{\rho}_n) = \sum_{i=0}^{\infty} B_n(i)(\boldsymbol{\rho}_n \cdot \mathbf{z}_n)^i = \Phi_n(\boldsymbol{\rho}_n \cdot \mathbf{z}_n), \tag{18}$$

where $\boldsymbol{\rho}_n \cdot \mathbf{z}_n$ is the ordinary inner product of two vectors and $\Phi_n(\zeta)$ is an analytic function defined by the power series $\sum_{i=0}^{\infty} B_n(i)\zeta^i$. Since $G^*(\mathbf{1}) = 1$ from the definition of p.g.f., we obtain the normalization constant as $C = 1/G(\mathbf{1})$, where $\mathbf{1}$ is an $N \times R$ array of all entries one. Since C itself may be written as a product on N terms, viz., $C = \prod_{n=1}^{N} G_n^{-1}(\mathbf{1}, \boldsymbol{\rho}_n)$, we find finally for the p.g.f.

$$G^*(\mathcal{Z}) = \prod_{n=1}^{N} \Phi_n(\boldsymbol{\rho}_n \cdot \mathbf{z}_n)/\Phi(\rho_n), \tag{19}$$

where $\rho_n = \Sigma_r \rho_{nr}$ is the total traffic intensity at center n.

For non-queue-dependent service centers, the function $\Phi_n(\zeta)$ is of the following form:

$$\Phi_n(\zeta) = \begin{cases} 1/(1-\zeta) \text{ for PS, LCFS, and FCFS,} \\ \exp(\zeta) \text{ for IS.} \end{cases} \tag{20}$$

For practical applications, the case of a *limited queue-dependent* service center is of special interest. By limited queue-dependence we mean that the scaling function $b(j)$ is such that $b(j) = \beta = \text{const.}$ for $j \ge r$; in other words, the server has constant rate if the queue size exceeds r. Parallel servers of multiplicity r fall into this class with $b(j) = \min\{j, r\}$. It is not difficult to see that for a limited queue-dependent service center, $\Phi_n(\zeta)$ can always be written as

$$\Phi_n(\zeta) = \phi_n(\zeta)/(1-\zeta), \tag{21}$$

where $\phi_n(\zeta)$ is a polynomial of degree $r-1$.

We obtain the p.g.f. of *marginal distributions* by equating certain transform variables and setting others to unity. For example:

1. The p.g.f. for $P\{\mathbf{k}\}$ is obtained by setting in (19) $z_{n1} = z_{n2} = \cdots = z_{nR} = z_n$ for all $n = 1, 2, \cdots, N$.
2. The p.g.f. for $P\{k_n\}$ is obtained by setting in (19) $z_{n1} = z_{n2} = \cdots = z_{nR} = z_n$ and all other z variables to unity.
3. The p.g.f. for $P\{\mathbf{K}\}$, i.e., for the number of customers in each subchain $\mathcal{M}_l$, is obtained by substituting z_l for all z_{nr} such that $(nr) \in \mathcal{M}_l$ and for all $l = 1, 2, \cdots, L$.

The solution is particularly simple for the case of $P\{k_n\}$, the marginal queue-size distribution. Because $G^*(\mathcal{Z})$ is a product of independent terms for each server, we find that the marginal queue-size distribution at center n is identical to that of a single server with workload intensity $\rho_n = \Sigma_r \rho_{nr}$. For service centers with constant rates, the queue size distribution is given by

$$P\{k_n\} = (1-\rho_n)\rho_n^{\,k_n}, \tag{22}$$

which is the familiar expression for an M/M/1 system. Similarly, for an IS service center,

$$P\{k_n\} = (\rho_n/k_n!)\exp(-\rho_n), \tag{23}$$

the result for an M/G/∞ system. The joint distribution $P\{\mathbf{k}\}$ is simply the product of the marginal distributions $P\{k_n\}$, i.e.,

$$P\{\mathbf{k}\} = \prod_{n=1}^{N} P\{k_n\}. \tag{24}$$

• *Mixed networks with closed chains and constant arrivals*

Let us assume that $\mathcal{N}$ has L_c closed subchains which, for notational convenience, are labeled $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_{L_c}$. The remaining $L - L_c$ open subchains are driven by Poisson streams with constant arrival rates λ_l. The state space of such networks is the set

$$\mathcal{F} = \{\mathcal{K};\ k_{nr} \geq 0 \text{ and } K_l = K_l^*\}. \tag{25}$$

The p.g.f. $G^*(\mathcal{Z}) = \sum P\{\mathcal{K}\} z_{nr}^{k_{nr}}$ is defined by a sum over the state space $\mathcal{F}$. This sum is more complicated than the one in (16) and does not split into a simple product of N terms, each dependent on one server only.

We can derive the p.g.f. $G^*(\mathcal{Z})$ from the simple function $G(\mathcal{Z})$ of Eq. (17) by following steps similar to those applied in obtaining the marginal distribution of subchain populations. We introduce new variables $\Theta = (\theta_1, \theta_2, \cdots, \theta_{L_c})$ which we associate with closed subchains. We next define $G(\mathcal{Z}, \Theta)$ as the function which is obtained by setting $\lambda = 1$ and substituting $\theta_l z_{nr}$ for z_{nr} in the original expression $G(z)$ for all $(nr) \in \mathcal{M}_l$ and $1 \leq l \leq L_c$. Let us assume for convenience that the classes in closed subchains are numbered as $1, 2, \cdots, R_c$ with $R_c \geq L_c$. Then we may write $G(\mathcal{Z}, \Theta)$ as

$$G(\mathcal{Z}, \Theta) = \prod_{n=1}^{N} \Phi_n\Big(\sum_{r=R_c+1}^{R} \rho_{nr} z_{nr} + \sum_{l=1}^{L_c} \theta_l \sum_{r=1}^{R_c} \rho_{nr} z_{nr} \Big). \tag{26}$$

For the sample network of Fig. 1 we have $\mathcal{Z} = [z_{11}, z_{12}, z_{14}, z_{24}, z_{31}, z_{32}, z_{33}, z_{43}]$, $\Theta = [\theta_1, \theta_2]$, $L = 3$, $L_c = 2$, $R = 4$, $R_c = 3$, and

$$G(\mathcal{Z}, \Theta) = \frac{1}{1 - [\rho_{14} z_{14} + \theta_1 \rho_{11} z_{11} + \theta_2 \rho_{12} z_{12}]} \times \frac{1}{1 - \rho_2 z_{24}} \times \frac{1}{1 - \rho_3 [\theta_1 z_{31} + \theta_2 (z_{32} + z_{33})]} \exp(\rho_4 \theta_2 z_{43}). \tag{27}$$

The p.g.f. is then found as the coefficient of $\theta_1^{K_1} \theta_2^{K_2} \cdots \theta_{L_c}^{K_{L_c}}$ in a multivariate power series expansion of $G(\mathcal{Z}, \Theta)$ in θ, symbolically written as

$$G^*(\mathcal{Z}) = C \cdot \partial_\Theta(K_1^*, K_2^*, \cdots, K_{L_c}^*) \cdot G(\mathcal{Z}, \Theta). \tag{28}$$

This power series expansion of $G(\mathcal{Z}, \Theta)$, however, does not lead to simple analytical expressions for $G^*(\mathcal{Z})$. Its numerical computation is discussed in the next section.

• *Normalization constant and related quantities*

The normalization constant, marginal distributions, and average queue sizes can be obtained as described in the subsection on open networks with constant arrivals.

The normalization constant is given by

$$C = [\partial_\Theta(K_1^*, K_2^*, \cdots K_{L_c}^*) \cdot G(\mathbf{1}, \Theta)]^{-1} \tag{29}$$

with

$$G(\mathbf{1}, \Theta) = \prod_{n=1}^{N} \Phi_n(\rho_n^{\,o} + \Theta \cdot \boldsymbol{\rho}_n^{\,c}), \tag{30}$$

where $\rho_n^{\,o} = \sum_{r=R_c+1}^{R} \rho_{nr}$ is the total workload intensity due to the open subchains and $\boldsymbol{\rho}_n^{\,c} = [\rho_{n1}^{\,c}, \rho_{n2}^{\,c}, \cdots, \rho_{nL_c}^{\,c}]$ is an L_c-vector whose element

$$\rho_{nl}^{\,c} = \sum_{(nr) \in \mathcal{M}_l} \rho_{nr} \tag{31}$$

is the total workload intensity in the closed subchain $\mathcal{M}_l$. To separate the effects of the open and closed chains, we expand $\Phi(\rho_n^{\,o} + \boldsymbol{\rho}_n^{\,o} \cdot \Theta)$ into a Taylor series around $\rho_n^{\,o}$, viz.,

$$\Phi_n(\rho_n^{\,o} + \boldsymbol{\rho}_n^{\,c} \cdot \Theta) = \Psi_n(\boldsymbol{\rho}_n^{\,c} \cdot \Theta) \tag{32}$$

with

$$\Psi_n(\zeta) = \begin{cases} \sum_{i=0}^{\infty} \frac{1}{i!} \Phi_n^{(i)}(\rho_n^{\,o}) \zeta^i \text{ (in general)} \\ \frac{1}{1 - \rho_n^{\,o}} \frac{1}{1 - \frac{\zeta}{1 - \rho_n^{\,o}}} \text{ (fixed rates)} \\ \exp(\rho_n^{\,o}) \exp(\zeta) \text{ (IS server)}. \end{cases} \tag{33}$$

From (29) and (33) it follows that

$$C = \Big[\partial_\Theta(\mathbf{K}^*) \cdot \prod_{n=1}^{N} \Psi_n(\boldsymbol{\rho}_n \cdot \Theta) \Big]^{-1}. \tag{34}$$

Equations (29) to (33) find an interesting interpretation.

1. For fixed-rate or IS service centers, an "open-system term" $\rho_n^{\,o}/(1 - \rho_n^{\,o})$ resp. $\exp(\rho_n^{\,o})$ can be factored out of the expression for the normalization constant.
2. A fixed-rate service center which is in open and closed subchains contributes to the closed subchain term in the same manner as a similar service center with adjusted traffic rates $\omega_{nl} = \rho_{nl}^{\,c}(1 - \rho_n^{\,o})^{-1}$. In other words, the effect of an open subchain with traffic rate $\rho_n^{\,o}$ on the closed subchains is to increase the traffic intensity by a factor $(1 - \rho_n^{\,o})^{-1}$.
3. Open subchains and closed subchains do not interact at an IS service center.

The p.g.f. for the marginal distribution $P\{k_n\}$ of the (total) queue size of service center m is

$$G(z_m) = C\partial_\Theta(\mathbf{K}^*) \cdot \Big(\Phi_m(z_m[\rho_m^{\,o} + \boldsymbol{\rho}_m^{\,c} \cdot \Theta]) \prod_{\substack{n=1 \\ n \neq m}}^{N} \Phi_n(\rho_n^{\,o} + \boldsymbol{\rho}_n^{\,c} \cdot \Theta) \Big). \tag{35}$$

The mean queue size $E[k_m]$ is found in the usual way, i.e., by $\partial G / \partial z_m$ at $z_m = 1$. This procedure yields

$$E[k_m] = C\partial_{\Theta}(\mathbf{K}) \cdot \left[\rho_m^{o} + \boldsymbol{\rho}_m^{c} \cdot \boldsymbol{\Theta}\right]\Phi^{1}(\rho_m^{o} + \boldsymbol{\rho}_m^{c} \cdot \boldsymbol{\Theta}) \prod_{\substack{n=1 \\ n\neq m}}^{N} \Psi_n(\boldsymbol{\rho}_n^{c} \cdot \boldsymbol{\Theta})\Big]. \quad (36)$$

For fixed-rate service centers, we find the special form

$$E[k_m] = \frac{\rho_m^{o}}{1-\rho_m^{o}} + \frac{C}{1-\rho_m^{o}}\partial_{\Theta}(\mathbf{K}^*) \times \left[(\boldsymbol{\rho}_m^{c} \cdot \boldsymbol{\Theta})\Psi_m(\boldsymbol{\rho}_m^{c} \cdot \boldsymbol{\Theta}) \prod_{n=1}^{N} \Psi_n(\boldsymbol{\rho}_n^{c} \cdot \boldsymbol{\Theta})\right], \quad (37)$$

and similarly for IS service centers we have

$$E[k_m] = \rho_m^{o} + C \exp(\boldsymbol{\rho}_m^{o})\partial_{\Theta}(\mathbf{K}^*) \times \left[(\boldsymbol{\rho}_m^{c} \cdot \boldsymbol{\Theta}) \prod_{n=1}^{N} \Psi_n(\boldsymbol{\rho}_n^{c} \cdot \boldsymbol{\Theta})\right]. \quad (38)$$

The first term in Eqs. (34) and (38) is the mean queue size of service center n, subject to "open" traffic with workload intensity ρ_m^{o}. Note that in the last two examples, the product $\Pi\Psi_n$ is the same as found in the expression for the normalization constant.

• *Stability of mixed networks*

A network is said to be stable if its servicing capacity can handle the arriving traffic and therefore all queues remain finite. More precisely, $\mathcal{N}$ is stable if the balance equations have a nonzero solution. In terms of the product-form solution, the sums defining $\partial_{\Theta}(\mathbf{K}^*) \cdot G(\mathbf{1}, \boldsymbol{\Theta})$ converge. We investigate this convergence by a new generating function which we obtain from $G(\mathscr{Z}, \boldsymbol{\Theta})$ by substituting ζ for all z_{nr} with (nr) in an open chain and $z_{nr} = 1$ for all (nr) in closed subchains, viz.,

$$G(\zeta, \boldsymbol{\Theta}) = \prod_{n=1}^{N} \Phi_n(\zeta\rho_n^{o} + \boldsymbol{\rho}_n^{c} \cdot \boldsymbol{\Theta}). \quad (39)$$

The transform variable ζ is associated with the total population in all open subchains. Now $\mathcal{N}$ is stable if

$$G(\zeta) = \partial_{\Theta}(\mathbf{K}^*) \cdot G(\zeta, \boldsymbol{\Theta}) \quad (40)$$

is analytic inside the unit circle in the complex ζ plane. By a Taylor series expansion of (39) we obtain

$$G(\zeta, \boldsymbol{\Theta}) = \prod_{n=1}^{N} \sum_{i=0}^{\infty} \frac{1}{i!} \Phi_n^{(i)}(\zeta_n^{o})[\boldsymbol{\rho}_n^{c} \cdot \beta]^{i} \quad (41)$$

within the region of convergence. Apparently $G(\zeta)$ is a linear combination of all $\Phi_n^{(i)}(\zeta\rho_n^{o})$ for $n = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, \Sigma K_l^*$, and therefore $G(\zeta)$ is analytic inside $|\zeta| \leq 1$ only if all the Φ_n and their derivatives are analytic. Since Φ_n is defined by the power series (19), all the derivatives are analytic whenever Φ_n is analytic. Therefore, for stability of the center n it suffices to require that $\Phi_n(\zeta\rho_n^{o})$ be analytic for $|\zeta| \leq 1$ for all $n = 1, 2, \cdots, N$. This is the stability condition of service center n alone. For example, in the case of fixed-rate FCFS, PS, or LCFS we have $\Phi_n(\zeta\rho_n^{o}) = [1 - \zeta\rho_n^{o}]^{-1}$, which is analytic inside $|\zeta| \leq 1$ if and only if $\rho_n^{o} \leq 1$. Similarly for IS service centers $\Phi_n(\zeta\rho_n^{o}) = \exp(\zeta\rho_n^{o})$ which is always analytic on the entire ζ-plane.

We conclude by recapitulating that in order for a mixed network to be stable, it is only required that each service center n be stable with respect to the open chain workload intensity ρ_n^{o}. Closed subchains do not contribute to instability.

Numerical solution of mixed networks

To evaluate numerically the normalization constant, marginal distributions, and moments of the queue-size distribution, we have to sum the product-form terms over appropriate regions in the state space. A simple term-by-term summation, however, is out of question for all but the most simple models, since the number of terms in these sums grows combinatorially with the size of the problem (i.e., with N and $\mathbf{K}^*$). Hence, more efficient algorithms are required for practical use of our network model. Such algorithms were previously reported for networks with only one closed subchain [11-14]. In the present section we further extend the previous work and discuss computational algorithms applicable to the general class of queuing networks.

The generating function method not only sets the earlier algorithms into perspective but also leads quite naturally to the general result discussed below. The problem, we recall, is a power series expansion of a multivariate function $Q(\boldsymbol{\Theta}) = G(\mathbf{1}, \boldsymbol{\Theta})$ which is a product of N terms $Q_n(\boldsymbol{\Theta}) = \Psi_n(\boldsymbol{\rho}_n^{c} \cdot \boldsymbol{\Theta})$. Two methods are available, namely:

1. Partial fractions if $Q(\boldsymbol{\Theta})$ is a rational function of only one variable θ (i.e., only one closed subchain).
2. Basic coefficient multiplication of power series.

The first approach was applied to a limited class of queuing network models by Moore [11]. The second method is essentially equivalent to the recursive algorithms discussed by Buzen [12] and by Reiser and Kobayashi [13, 14]. Below we generalize the partial fraction method to the case of mixed networks with one closed subchain and also remove several constraints forced in [11]. We then give a general algorithm which is based on multiplication of power series and which, in the light of generating functions, becomes remarkably simple.

• *Partial fraction method for a mixed network with only one closed subchain*

Let $\mathcal{N}$ be a network with only one closed chain and

let us further assume that there is no IS service center and that all queue-dependent FCFS service centers are of the limited queue-dependent type. For this network, the function $Q(\Theta)$ for which we seek a power series expansion is of the form

$$Q(\Theta) = G(1, \Theta) = \prod_{n=1}^{N} \frac{\psi_n(\omega_n\theta)}{1-\omega_n\theta}, \tag{42}$$

where $\omega_n = \rho_n^{\circ}(1-\rho_n^{\circ})^{-1}$ is the adjusted traffic intensity and $\psi_n(\zeta) = \phi_n(\rho_n^{\circ}+\zeta)$ is the polynomial of Eq. (21). Note that $\psi_n = 1$ for fixed rate service centers. It is easy to write (42) in partial fractions, viz.,

$$Q(\Theta) = \psi_q(\theta) + \sum_{n=1}^{N} \alpha_n/(1-\omega_n\theta). \tag{43}$$

where $\psi_q(\theta)$ is the quotient of the polynomial division $[\Pi\psi_n(\omega_n\theta)]/\Pi(1-\omega_n\theta)$ and the α_n are residues. If all ω_n are distinct, these residues are

$$\alpha_n = \psi_r(\omega_n^{-1}) \prod_{i=1,\ i\neq n}^{N} (1-\omega_i/\omega_n)^{-1}, \tag{44}$$

with ψ_r being the remainder of the above division. From (43) follows immediately the desired power series expansion of $Q(\Theta)$; in particular, we find

$$\partial_{\Theta}(K^*)Q(\Theta) = q(K^*) + \sum_{n=1}^{N} (\psi_r\omega_n^{-1}\omega_n^{K^*})\left[\prod_{i=1,\ i\neq n}^{N}(1-\omega_i/\omega_n)\right]^{-1},$$

where $q(K^*)$ is the coefficient of θ^{K^*} in the polynomial $\psi_q(\theta)$. Equation (45) gives rise to the following algorithm for computing $\partial_{\Theta}(\mathbf{K}^*)\cdot Q(\Theta)$:

Step 1 Compute quotient and remainder of the polynomial equation in θ,

$$\left[\prod_{n=1}^{N}\psi_n(\omega_n\theta)\right]\Big/\prod_{n=1}^{N}(1-\omega_n\theta).$$

Step 2 Compute residues α_n for $n = 1, 2, \cdots, N$ by (44).

Step 3 For given K^* compute $\partial_{\Theta}(K^*)\cdot Q = q(K^*) + \sum_{n=1}^{N}\alpha_n\omega_n^{K^*}$.

The polynomial division and the required polynomial evaluations are conveniently done using Horner's rule. Once step 2 is completed, step 3 can be repeated for various values of K^* at little extra cost. Most computational effort is spend at step 2 which has an operation count of $\mathcal{O}(N^2)$. It is not difficult to remove the restriction on ω_n.

The partial fraction method, when applicable, is usually the more efficient algorithm. It has, however, two drawbacks, namely:

1. The sum in (45) has alternating signs and may be subject to round-off errors in some cases.
2. The method cannot be generalized to the case with more than one closed chain.

• *Multiplication of multinomial power series*

We start with mathematical formulas pertinent to the general algorithm discussed in the next section. Let $Q_1(\Theta)$ and $Q_2(\Theta)$ be functions in m variables $\Theta = [\theta_1, \theta_2, \cdots, \theta_m]$ which are defined by power series with coefficients $q_1(\mathbf{i}) = q_1(i_1, i_2, \cdots, i_m)$ and $q_2(\mathbf{i})$, respectively, viz.,

$$Q_1(\Theta) = \sum_{\mathbf{i}>\mathbf{0}} q_1(\mathbf{i})\Theta_1^{i_1}\Theta_2^{i_2}\cdots\Theta_m^{i_m}, \tag{46}$$

and similarly for $Q_2(\Theta)$. The product $Q = Q_1Q_2$ has coefficents $q(\mathbf{i})$ which can be obtained from those of Q_1 and Q_2 by

$$q(\mathbf{K}) = \sum_{i_1=0}^{K_1}\sum_{i_2=0}^{K_2}\cdots\sum_{i_m=0}^{K_m} q_1(\mathbf{i})q_2(\mathbf{K}-\mathbf{i}), \tag{47}$$

where $\mathbf{K} = [K_1, K_2, \cdots, K_m]$ and $\mathbf{i} = [i_1, i_2, \cdots, i_m]$ are index vectors. Equation (48) is an m-dimensional convolution. Note that in order to evaluate $q(\mathbf{K})$ at a point $\mathbf{K}$ in the index space, only coefficients $q_1(\mathbf{i})$ and $q_2(\mathbf{i})$ at points $\mathbf{i}$ closer to the origin are required, i.e., $\mathbf{0} \le \mathbf{i} \le \mathbf{K}$ or $0 \le i_1 \le K_1, 0 \le i_2 \le K_2, \cdots, 0 \le i_m \le K_m$. We can interpret (47) as a *multidimensional linear filtering* [15] with $q_1(\mathbf{i})$ as input, $q_2(\mathbf{i})$ as filter coefficients, and $q(\mathbf{i})$ as output. Then the filter we discuss here is a *causal* one [16] in the sense that the output is dependent only on the past input, i.e., those $q_1(\mathbf{i})$ with $\mathbf{0} \le \mathbf{i} \le \mathbf{K}$. We observe that the filter Q_2 acts on the entire past history of $q_1(\mathbf{i})$; hence $\mathcal{O}[(K_1, K_2, \cdots, K_m)^2]$ operations are required to compute the output $q(\mathbf{i})$ at all points $\mathbf{0} \le \mathbf{i} \le \mathbf{K}$.

A significant reduction in the operation count is possible if the filter function $Q_2(\Theta)$ is a rational function in Θ, viz.,

$$Q_2(\Theta) = \frac{Q_F(\Theta)}{1-Q_B(\Theta)}, \tag{48}$$

where $Q_F(\Theta)$ is a polynomial of degree d_F with coefficients $q_F(\mathbf{i})$ defined on the index set $\mathcal{T}_F = \{\mathbf{i} > \mathbf{0}$ and $(i_1 + i_2 + \cdots + i_m) \le d_F\}$ and similarly $Q_B(\Theta)$ is a polynomial of degree d_B with coefficients $q_B(\mathbf{i})$ for $\mathbf{i}$ in the index set $\mathcal{T}_B = \{i; \mathbf{i} > \mathbf{0}$ and $(i_1 + i_2 + \cdots + i_m) \le d_B\}$. Note that Q_B has no constant term. It is a well known fact that the filter function (48) can be realized by a feed-forward/feed-back filter as shown in Fig. 2. The input-output relation is therefore given by

$$q(\mathbf{K}) = \sum_{\mathbf{i}\in\mathcal{T}_F} q_F(\mathbf{i})q_1(\mathbf{K}-\mathbf{i}) + \sum_{\mathbf{i}\in\mathcal{T}_B} q_B(\mathbf{i})q(\mathbf{K}-\mathbf{i}). \tag{49}$$

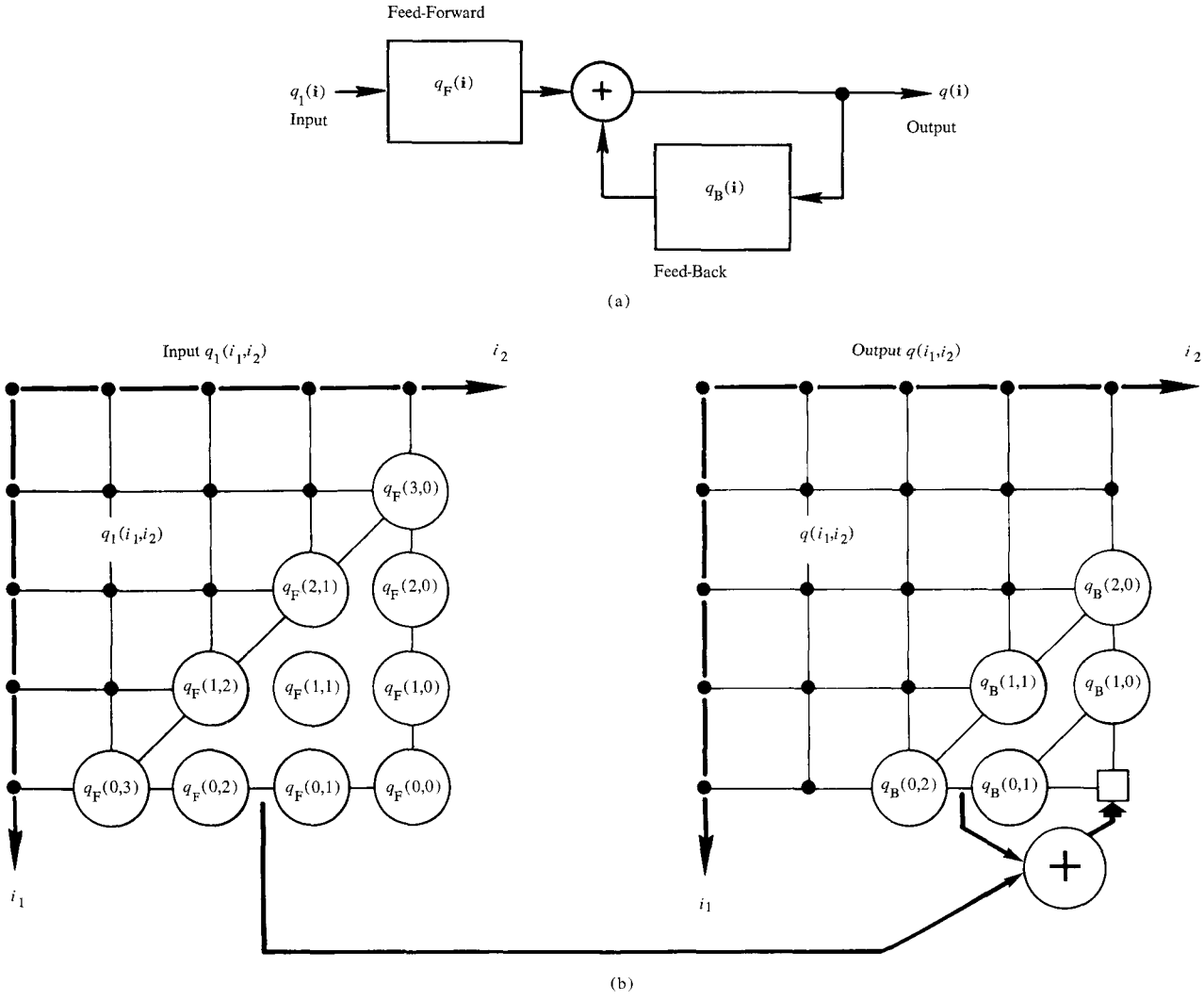


Figure 2 (a) Schematic representation of a feed-forward/feed-back filter, (b) Computation diagram for a spatial (i.e., $m = 2$) feed-forward/feed-back filter with $d_F = 3$ and $d_B = 2$. The circles symbolize multipliers which act on the underlying grid values.

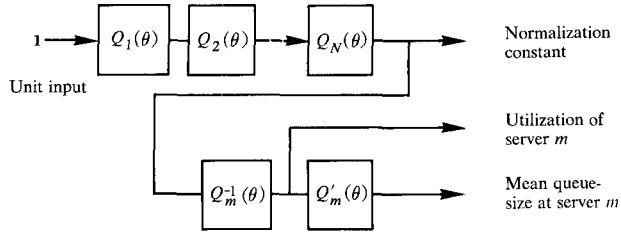
The process $q(\mathbf{i})$ can also be viewed as a multidimensional analogue of an *autoregressive moving-average* process [17]: The first term represents the moving average of the input $q_1(\mathbf{i})$ and the second, the autoregressive term. The operation count of (49) is $\mathcal{O}[(d_F{}^m + d_B{}^m) K_1 K_2 \cdots K_m]$. Of special interest to us is $d_F = 0$ and $d_B = 1$, in which case we have an operation count of $\mathcal{O}(K_1 K_2 \cdots K_m)$.

• *General algorithm*

From the preceding discussion, it is clear that the desired coefficient $q(\mathbf{K}) = \partial_{\Theta}(\mathbf{K}) \cdot Q(\Theta)$ is the impulse response evaluated at $\mathbf{i} = \mathbf{K}$ of a cascade of N linear filters with transfer functions $Q_n(\Theta) = \Psi_n(\rho_n{}^c \cdot \Theta)$ (see Fig. 3). Since we are not interested in the output for any $\mathbf{i} > \mathbf{K}$, we can always write $Q_n(\Theta)$ in the form of a feed-forward/feed-back filter with a feed-forward part $Q_F(\Theta)$ of the form

$$Q_F(\Theta) = \begin{cases} 1 \text{ for PS, LCFS, and fixed rates} \\ \psi_n(\rho_n{}^c \cdot \Theta) \text{ for limited queue-dependent rates} \\ \sum_{i=1}^{K} (1/i!)(\rho_n{}^c \cdot \Theta)^i \text{ for IS,} \end{cases} \tag{50}$$

Figure 3 Signal flow diagram of the general algorithm to compute normalization constant, server utilization, and mean queue sizes.



with $K = K_1 + K_2 + \cdots + K_{I_c}$ and $\psi_n(\zeta) = \phi_n(\rho_n^{\,0} + \zeta)$. The feed-back part $Q_B(\Theta)$ similarly becomes

$$Q_B(\Theta) = \begin{cases} (\boldsymbol{\rho}_n^{\,c} \cdot \Theta) \text{ for fixed rates,} \\ 1 \text{ for IS.} \end{cases} \tag{51}$$

Apparently, for an IS service center, we have a feed-forward filter only which extends over the entire past history.

We summarize this procedure in algorithmic notation:

Step 1 Set up the m-dimensional arrays Q and Q_1 with index bounds $\mathbf{0} \leq \mathbf{i} \leq \mathbf{K}$.

Step 2 Initialize $Q \leftarrow$ (1 for $\mathbf{i} = \mathbf{0}$, 0 otherwise) and $n \leftarrow 1$.

Step 3 $Q_1 \leftarrow Q$.

Step 4 Compute filter coefficients and store them in Q_F and Q_B [Eqs. (51) and (52)].

Step 5 Compute filtered output for the input in Q_1 and store it in Q [Eq. (50)].

Step 6 $n \leftarrow n + 1$. If $n > N$ then stop; else go to step 3.

After completion of the algorithm, $\partial_\Theta(\mathbf{K}) \cdot Q(\Theta)$ is found in array location $Q(\mathbf{K})$. The signal flow diagram of this algorithm is given in Fig. 3.

The operation count is $\mathcal{O}[(N_1 \Pi K_l + N_2)\Pi K_l]$, where N_1 is the number of general queue-dependent service centers and N_2 is the number of fixed-rate rate-limited queue-dependent-rate service centers. The storage requirement is $\mathcal{O}(2\Pi K_l)$ if there are queue-dependent service centers, $\mathcal{O}(\Pi K_l)$ otherwise. Both computational effort and storage requirement grow rapidly with the number of closed subchains and also with the population therein. It should be noted that the sums for the convolution are over positive terms only and therefore are *numerically stable*.

The high computational effort can often be reduced by taking advantage of special properties of $\mathcal{N}$. For example:

1. The filtering can be restricted to lower dimensional subspaces of the index space if all N service centers are not visited by every subchain.
2. More than one IS service center can be easily lumped together by $\Pi \exp(\boldsymbol{\rho}_n^{\,c} \cdot \Theta) = \exp[(\Sigma \boldsymbol{\rho}_n^{\,c}) \cdot \Theta]$ where product and sum are over all IS service centers. The combined exponential function should be applied to the unit input, i.e., to the corresponding filter used first in the cascade of N filters. Then the high computational effort of the general convolution (48) can be totally avoided.
3. If results are required for several different subchain populations $\mathbf{K}$, these are computed simultaneously for all $\mathbf{K} \leq \mathbf{K}_{max}$ where $\mathbf{K}_{max}$ is the original $\mathbf{K}$ value.

• *Fast computation of marginal queue-size distributions and their moments*

We showed in the previous section that the marginal queue-size distribution at service center m and its moments can be obtained by expressions of the form $\partial_\Theta(\mathbf{K}) \cdot Q_m^*(\Theta) \prod_{n=1,\, n\neq m}^{N} Q_n(\Theta)$. In terms of linear filtering, we apply a filter with transfer function $Q_n^*(\Theta)$ [18] to the output of $\prod_{n=1,\, n\neq m}^{N} Q_n(\Theta)$. For the service center labeled $n = N$ this output is part of the computation of the normalization constant, and therefore little additional effort is needed to obtain the desired results.

In general, however, the output of the filter with transfer functions $\prod_{n=1,\, n\neq m}^{N} Q_m(\Theta)$ has to be computed separately for each m. An efficient way to perform this computation is to apply the inverse filter $Q_m^{-1}(\Theta)$ to the output of $\prod_{n=1}^{N} Q_n(\Theta)$, which we have obtained as a by-product of the evaluation of the normalization constant C. The inverse filter is again a feed-forward/feed-back filter with transfer function

$$Q_n^{-1}(\Theta) = \frac{q_F(\mathbf{0})[1 - Q_B(\Theta)]}{1 - [1 - Q_F(\Theta)]} . \tag{53}$$

Note that the inverse filter Eq. (53) must be used with caution because it may be numerically unstable. Fixed-rate service centers always have stable inverse filters of the simple feed-forward form $Q_n^{-1}(\Theta) = 1 - \boldsymbol{\rho}_n^{\,c}\Theta$.

The computation of the mean queue size becomes especially simple for a fixed-rate service center. In this case, we have $Q_m^*(\Theta) = Q_m'(\Theta) = Q_m^{\,2}(\Theta)$. Therefore, we simply apply the filter $Q_m(\Theta)$ to the output of the entire filter cascade with transfer function $\prod_{n=1}^{N} Q_n(\Theta)$. In the case of an IS service center, the computation of the mean queue size is even simpler, since $Q_n^*(\Theta) = Q_m(\Theta)$, and therefore additional filtering is not required.

Appendix

Let $\mathcal{S}([nr]^-)$, $\mathcal{S}([nr]^+)$, and $\mathcal{S}([nr]^+, [n'r']^-)$ be states such that the following state transitions are possible between them:

1. $\mathcal{S}([nr]^-) \rightarrow \mathcal{S}$ upon arrival of a class r customer at service center n.
2. $\mathcal{S}([nr]^+) \rightarrow \mathcal{S}$ upon departure of a class r customer from service center n.
3. $\mathcal{S}([nr]^+, [n'r']^-) \rightarrow \mathcal{S}$ upon departure of a class r customer from service center n and his subsequent arrival at service center n' after a class change to r'.

With these definitions, the overall balance equations for a network of FCFS service centers are

$$\left[\sum_{l=1}^{L}\lambda_l(K_l)+\sum_{n=1}^{N}\sum_{r=1}^{R}\mu_n(k_n)(1-p_{(nr),\,(nr)})\right]P\{\mathscr{S}\}$$
$$=\sum_{n=1}^{N}\sum_{r=1}^{R}\mu_n(k_n+1)p_{(nr),\,l}P\{\mathscr{S}([nr]^+)\}$$
$$+\sum_{n=1}^{N}\sum_{r=1}^{R}\lambda_l(K_l-1)p_{l,\,(nr)}P\{\mathscr{S}([nr]^-)\}$$
$$+\sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{r=1}^{R}\sum_{r'=1}^{R}\mu_{n'}+1)p_{(n'r'),\,(nr)}$$
$$\times P\{\mathscr{S}([n'r']^+,[nr]^-)\},\qquad(A1)$$

where we assume that all l are such that $(nr)\in\mathscr{M}_l$. The principle of *local balance* or *individual balance* equations [9] is simply to equate a subset of the left-hand terms in (A1) with a subset of the right-hand terms. For example, if $A+B+C=D+E+F$ is a global balance equation, then $A=D$, $B=E$, $C=F$ are local balance equations. Clearly a set of local balance equations is a sufficient condition for the global balance equation but it is not a necessary condition. In the general case the local balance equations may be contradictory to each other. We use the principle of local balance with great care in the subsequent derivation.

First we equate the rate with which the system leaves the state $\mathscr{S}$ due to customer arrivals with the rate with which the system enters the state $\mathscr{S}$ due to customer departures, viz.,

$$\sum_{l=1}^{L}\lambda_l(K_l)\,P\{\mathscr{S}\}=\sum_{n=1}^{N}\sum_{r=1}^{R}\mu_n(k_n+1)\,p_{(nr),\,l}\,P\{\mathscr{S}([nr]^+)\}$$
$$=\sum_{l=1}^{L}\sum_{(nr)\in\mathscr{M}_l}\mu_n(k_n+1)\,p_{(nr),\,l}\,P\{\mathscr{S}([nr]^+)\}.\qquad(A2)$$

This relation is a balance between $\mathscr{N}$ and the outside. We next use local balance again and separate (A2) into L individual equations,

$$\lambda_l(K_l)P\{\mathscr{S}\}=\sum_{(nr)\in\mathscr{M}_l}\mu_n(k_n+1)p_{(nr),\,l}P\{\mathscr{S}([nr]^+)\}.\qquad(A3)$$

By multiplying the left side of (A3) by $\sum_{(nr)\in\mathscr{M}_l}p_{l,\,(nr)}\lambda_l(K_l)=1$ and by substituting the relations

$$p_{(nr),\,l}=1-\sum_{(n'r')\in\mathscr{M}_l}p_{(nr),\,(n'r')}$$

and

$$p_{l,\,(nr)}=e_{nr}-\sum_{(n'r')\in\mathscr{M}_l}e_{n'r'}p_{(n'r'),\,(n,r)}$$

[see Eq. (3)] into (A3) we partition the resultant equation into the following set of local balance equations, for all $(nr)\in\mathscr{M}_l$,

$$\lambda_l(K_l)e_{nr}P\{\mathscr{S}\}=\mu_n(k_n+1)P\{\mathscr{S}([nr]^+)\}\qquad(A4)$$

and

$$\lambda_l(K_l)\sum_{(n'r')\in\mathscr{M}_l}e_{n'r'}p_{(nr),\,(n'r')}P\{\mathscr{S}\}$$
$$=\mu_n(k_n+1)\sum_{(n'r')\in\mathscr{M}_l}p_{(nr),\,(n'r')}P\{\mathscr{S}([nr]^+)\}.\qquad(A5)$$

Evidently (A4) is equivalent to (A5).

We now turn our attention to the second set of terms in (A1), namely the balance between the rate of system transition out of state $\mathscr{S}$ due to service completion and the rate of transition into $\mathscr{S}$ due to customer arrivals from outside or due to interval transitions:

$$\sum_{n=1}^{N}\sum_{r=1}^{R}\mu(k_n)[1-p_{(nr),\,(nr)}]P\{\mathscr{S}\}$$
$$=\sum_{n=1}^{N}\sum_{r=1}^{R}\lambda_l(K_l-1)p_{l,\,(nr)}P\{\mathscr{S}([nr]^-)\}$$
$$+\sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{n'=1}^{N}\sum_{r'=1}^{R}\mu_{n'}(k_{n'}+1)\,p_{(n'r'),\,(nr)}$$
$$\times P\{\mathscr{S}([n'r']^+,[nr]^-)\}.\qquad(A6)$$

We first split (A6) into $N\times R$ balance equations. We then proceed in the same manner as in the derivation of (A4) and obtain

$$\lambda_l(K_l-1)e_{nr}\,P\{\mathscr{S}([nr]^-)\}=\mu_n(k_n)P\{\mathscr{S}\}\qquad(A7)$$

and

$$\lambda_l(K_l-1)e_{nr}\,P\{\mathscr{S}([nr]^-)\}$$
$$=\mu_n(k_n+1)\,P\{\mathscr{S}([n'r']^+,[nr]^-)\}.\qquad(A8)$$

We now have three equations which are, surprisingly enough, equivalent to each other. Thus, if one of them, say (A4), is satisfied for all $l=1,2,\cdots,L$ and $(nr)\in\mathscr{M}_l$, then the overall balance equations are also met. Note that if we let the rate be class-dependent, $\mu_{nr}(k_{nr})$ say, then the three equations are contradictory and no product-form solution exists. Finally, it remains to be shown that the recurrence equations (A4) and (A7–A8) also hold for the boundary, an exercise which we do not carry out in detail.

Acknowledgment

The authors are grateful to Kasra Hazeghi for pointing out mistakes in the analysis of FCFS centers in our earlier paper [19].

References and notes

1. J. R. Jackson, "Job Shop-like Queueing Systems," *Management Sci.* **10,** 131 (1963).
2. J. P. Buzen, "Analysis of System Bottlenecks Using a Queueing Network Model," *Proc. ACM-SIGOPS Workshop on System Performance Evaluation*, April 1971, pp. 82-103.
3. S. R. Arora and A. Gallo, "The Optimal Organization of Multiprogrammed Multi-level Memory," *ibid.*, pp. 104-141.
4. F. Baskett and R. R. Muntz, "Queuing Network Models with Different Classes of Customers," *Proceedings of the*

Sixth Annual IEEE Computer Society International Conference, September 1972, pp. 205-209.
5. F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers," *J. ACM* (to be published).
6. The processor shared queue discipline may be viewed as the limiting case of round robin scheduling with the quantum size approaching zero.
7. At an infinite server queue there are always enough servers to accommodate every job. Terminals and time delays are examples where IS queues may be useful.
8. D. R. Cox, "A Use of Complex Probabilities in the Theory of Stochastic Processes," *Proc. Camb. Phil. Soc.* **51,** 313 (1955).
9. K. M. Chandy, "The Analysis and Solution for General Queuing Networks," Proceedings of the Seventh Annual Princeton Conference on Information Sciences and Systems, Princeton University, March 1973, pp. 428-434.
10. M. Reiser and A. G. Konheim, "Blocking in a Queueing Network with Two Exponential Servers," *Research Report RJ 1360*, IBM Research Laboratory, San Jose, California, March 1974.
11. F. R. Moore, "Computational Model of a Closed Queuing Network with Exponential Servers," *IBM J. Res. Develop.* **16,** 567 (1972).
12. J. P. Buzen, "Computational Algorithms for Closed Queueing Networks with Exponential Servers," *Commun. ACM* **16,** 527 (1973).
13. M. Reiser and H. Kobayashi, "Recursive Algorithms for General Queueing Networks with Exponential Servers," *Research Report RC 4254*, IBM Thomas J. Watson Research Center, Yorktown Heights, New York, March 1973.
14. M. Reiser and H. Kobayashi, "Numerical Solution of Semiclosed Exponential Server Queueing Networks," *Proceedings of the Seventh Asilomar Conference on Circuits, Systems, and Computers*, Pacific Grove, CA, November 1973, pp. 308-312.
15. When $m = 1$ we can regard the index r_i as discrete time; thus an exact analogy to discrete time linear filtering holds here. When $m = 2$ we can view i_1 and i_2 as x and y coordinates; then $q(\mathbf{i})$ can be considered as a spatial filter.
16. The notion of causality defined here is, of course, a multidimensional analogue of the causality usually used for a *time-domain* filter.
17. G. F. Box and G. M. Jenkins, *Time Series Analysis*, Holden-Day Publishing Co., San Francisco, California, 1970.
18. The form of the additional transfer (or filter) function $Q^*(\boldsymbol{\Theta})$ depends on the desired result and is given explicitly in (35) to (38).
19. H. Kobayashi and M. Reiser, "Some Results on Queuing Models with Different Classes of Customers," *Proceedings of the Eighth Annual Princeton Conference on Information Sciences and Systems*, Princeton University, March 1974, to be published.

Received July 15, 1974

The authors are located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.