The Mathematics of Product Form Queuing Networks

RANDOLPH D. NELSON

IBM Research Division, T. J. Watson Research Center, P.O. Box 704, Room H2-D26, Yorktown Heights, NY 10598

Markov processes that have a product form solution have become an important computer performance modeling tool. The fact that such a simple solution exists for seemingly complex Markov processes is surprising at first encounter and can be established by showing that balance equations are satisfied. In this article we attempt to provide insight as to why such a solution form exists and demonstrate that product form and companion results, such as the arrival theorem and Norton's theorem, are consequences of four properties satisfied by queues that satisfy partial balance. Notions of reverse processes, reversibility, and quasireversibility are developed to establish the four properties.

Categories and Subject Descriptors: C.4 [**Computer Systems Organization**]: Performance of Systems—*modeling techniques*; G.3 [**Mathematics of Computing**]: Probability and Statistics; I.6 [**Computing Methodologies**]: Simulation and Modeling

General Terms: Performance

Additional Key Words and Phrases: Networks, partial balance, product form, quasireversibility, queuing theory, reversibility

1. INTRODUCTION

The discovery that certain queuing networks have tractable *product form solutions* [Baskett et al. 1975; Gordon and Newell 1967; Jackson 1963; Whittle 1967] has had a profound influence on computer performance modeling. In such systems the stationary distribution of the network is composed of a product of the distributions of each queue analyzed in isolation (subject to a normalization constant). When first encountered, such a solution is difficult to understand since for open networks it implies independence (of the stationary distributions) of the individual queues, and for closed networks it implies that the dependence between the queues is captured by normalizing the independent solution over a truncated state space. The purpose of this article is to provide some insight into why such solutions are obtained. We provide this insight by showing that product form and related results, such as the arrival theorem and Norton's theorem, follow from four properties of queues that satisfy partial balance. Each of these four properties can be understood within the context of a simple queuing system. The algebra for how such queues can be formed into a network while still retain-

CONTENTS

ing product form is shown to follow from the four properties. These properties unify the approach to product form, and, since establishing them does not initially burden intuition with excessive notation or algebraic manipulation, they bring to light the reasons underpinning the solution form. A by-product of this approach is that we clear up a possible confusion regarding the differences between reversible and quasireversible processes.

Networks of queues have been extensively used to model computer and communications networks. Work includes models of specific computer systems such as IBM mainframes running the VM [Bard 1978a; 1978b] and MVS [Buzen 1978] operating systems, subsystems such as DASD [Bard 1980; Brandwajn and McCormack 1984], memory and interconnection networks [Brown et al. 1977; Lam 1977; Lazowska and Zahorjan 1982; Thomasian and Bay 1984; Towsley 1983; 1986], systems that have features of parallel or concurrent processing [Heidelberger and Trivedi 1982; 1983; Le Boudec 1985; Sauer 1981; Thomasian and Bay 1986], and models that include blocking [Hordijk and Van Dijk 1981]. Often, the difficulty in calculating stationary measures for product form networks lies in the complexity of the numerical calculations required to calculate normalization constants. Much research has focused on creating new and efficient methods for determining performance measures [Conway and Georganas 1986; Conway et al. 1989; de Souza e Silva and Lavenberg 1989; Hoyme et al. 1986; Lam and Lien 1983; Reiser and Kobayashi 1975; Reiser and Lavenberg 1980] or by creating approximate techniques that can be used when product form does not hold [Bryant et al. 1984; Chandy and Sauer 1978; Chandy and Neuse 1982; de Souza e Silva et al. 1986; Eager and Lipscomb 1988; Krzesinski and Greyling 1984; Krzesinski and Teunissen 1985; Schweitzer 1979; Zahorjan et al. 1988]. Modeling of communication networks using product form networks includes work found in Henderson and Taylor [1989], Nelson and Kleinrock [1985], Reiser [1979], Van Dijk [1990a; 1990b; 1991], and Wong [1978].

The subject of product form queuing networks is mature, and there are several surveys of the area [Disney 1975; Disney and König 1985; Gelenbe and Muntz 1976; Lemoine 1977] and books which have sections devoted to various aspects of the subject [Disney and Kiessler 1987; Kelly 1979; Lavenberg 1983; Lazowska et al. 1984; Ross 1983; Sauer and Chandy 1981; Walrand 1988; Whittle 1986a]. The approach in this article is heavily indebted to Kelly's elegant treatment of the subject. Other main sources for the material come from two excellent books, Walrand [1988] and Whittle [1986a], and a thorough survey by Disney and Konig [1985]. Arguments often rely on viewing a Markov process in reverse time. Kolmogorov [1936] appears to be the first to consider such processes, and the theory was later extended in Reich [1957]. Viewing systems in reverse time yields important insight into the input-output behavior of queuing systems. Burke [1956] first established that the departure process of an M/M/1 queue is Poisson and is independent of the state of the queue. Such a queue thus produces Poisson outputs when presented with Poisson inputs. Muntz [1972; 1973] called this the $M \Rightarrow$

M property and first observed that product form solutions arise from input/output properties of queues. This approach allows one to ignore internal details of a queuing structure. In Muntz [1972], the $M \Rightarrow M$ property was used to establish that partial and local balance (as defined later in the paper) hold for product form networks and was also used to establish algebraically that certain queues satisfied properties later collectively termed quasireversibility by Kelly [1976a]. Further work characterizing input-output properties of queues can be found in Burke [1972; 1976], Daley [1976], Disney [1975], Kelly [1976b; 1983], and Kelly and Pollett [1983], and theoretical work on product form networks can be found in Chandy and Martin [1983], Harrison and Williams [1990], Hordijk and Van Dijk [1983a; 1983b], Kelly [1982], Pittel [1979], and Serfozo [1989]. Previous to much of this research, Koenigsberg [1958] showed that cyclic networks of queues had product form solutions, and Jackson [1963] established this fact for a larger class of networks. These results were extended to closed queuing networks by Gordon and Newell [1967] and in terms of migration models by Whittle [1967; 1986a] (where the notion of partial balance was first introduced) and by Kingman [1969]. The important **BCMP** paper [Baskett et al. 1975] (the acronym is composed of a concatenation of the first letters of the last names of the authors) established that a useful class of queuing networks satisfied partial balance and also satisfied product form. This had a profound influence on computer performance modeling and set a direction for further work (see Barbour [1976] and Chandy et al. [1977] for instance). Kelly [1975; 1976a] also independently established that product form holds for certain classes of networks.

Two types of networks are commonly distinguished: *open networks*, which have external arrival streams for all classes of customers, and *closed networks*, in which there is a fixed population of customers for all classes of customers. For open networks, product form implies that the stationary distribution is a product of individual queue distributions obtained by analyzing each queue with an appropriately modified arrival rate to reflect the routing of traffic in the network. For closed networks, computational difficulties emerge since a normalization constant must be calculated so that probabilities sum to unity over a restricted state space [Buzen 1973; Chandy and Sauer 1980; McKenna and Mitra 1982; Reiser and Kobayashi 1975]. An important property of closed networks is that the distribution of the system seen by a customer *in transit* between queues (but not yet resident in any queue) is the same as the stationary distribution of a system that does not contain the transit customer. This is a form of an *arrival theorem* [Lavenberg and Reiser 1980; Melamed 1982; Sevcik and Mitrani 1981]. This insight led to an important paper by Reiser and Lavenberg [1980] that derives a set of recurrence equations that can be used to compute derived quantities, such as mean queue lengths, without calculating a normalization constant. This procedure, called **Mean Value Analysis**, has had a major influence on the application of closed queuing networks to model computer systems.

In this article, we concentrate on the mathematics leading up to product form and do not consider difficulties associated with the computational aspects of the problem. The basic mathematical structures that we consider are Markov processes which satisfy a set of balance equations called partial balance. For such processes, the underlying mathematics that implies product form is often expressed in terms of probabilistic relationships that are found between events of the Markov process and specified sets of states. Such a representation does not require an interpretation of how such sets and transitions are reflected in the system that is modeled by the Markov process. In the special case of a queuing network, one can create a correspondence between states of the queues with sets of states found in the underlying Markov process. Such a correspondence

can also be established between customer events in the queuing network and transition events in the Markov process. Thus, the abstract theory developed for Markov processes can be applied to that of the actual queuing network.

As an example of this correspondence, suppose the state of a Markov process that models a queuing network is a vector $\mathbf{n} = (n_1, n_2, \ldots, n_K)$, $K \geq 1$, where n_i, $1 \leq i \leq K$, is the number of class i customers in the system. The set of all possible queuing states that can exist when a class i customer arrives to the system corresponds to the set of all possible states of the Markov process that can exist when a transition causes the value of n_i to increase by 1. Although we concentrate on queuing-network applications of product form networks (there are many other applications that do not correspond to queuing networks such as polymerization models and genetic models) we find it more convenient to derive the basic mathematics of product form in the abstract paradigm of Markov process theory rather than in the specific terms of queuing networks. This approach has the advantage of presenting the underlying mathematics of product form in a general, abstract, setting. Later, when discussing how these results can be used to model networks, we switch over to the ideology of queuing networks.

In Section 2 we present preliminary results regarding Markov processes and introduce the notions of reverse time processes, reversibility, and quasireversibility. The four properties of partial balance that are heavily used in product form networks are also established in this section. Partial balance is an important principle behind product form, and we fully explore its ramifications in this section. Our emphasis is on product form as found in networks of quasireversible queues. Although product form can hold when quasireversibility is not satisfied, quasireversible networks are frequently found in applications, and concentrating our attention to this type of network is not, in our opinion, unduly restrictive. In Section 3 we show how individual quasireversible queues can be joined into networks that preserve quasireversibility. Such networks have product form solutions. In Section 4 we present our conclusions.

2. PRELIMINARY RESULTS

In this section we define our notation and establish fundamental properties of Markov processes and their time reversal counterparts. We denote sets in calligraphic type style (i.e., $\mathcal{S}, \mathcal{U}$) and will denote the union of sets $\mathcal{U}$ and $\mathcal{V}$ as $\{\mathcal{U}, \mathcal{V}\}$. The notation $\{\mathcal{S}_c\}_{c=1}^{C}$ will represent the union of sets $\mathcal{S}_c$ over c, i.e., $\cup_{c=1}^{C} \mathcal{S}_c$, and $\bar{\mathcal{U}}$ will denote the complement of set $\mathcal{U}$ with respect to some universal set.

2.1 Preliminary Definitions

We let $\mathcal{S}$ be a countable set of states and let $X(t)$, $-\infty < t < \infty$, be a Markov process defined on $\mathcal{S}$. Throughout this article we assume that $X(t)$ is time homogeneous, irreducible, and stationary [Ross 1983]. We will sometimes suppress the time dependency in our notation of $X(t)$. The *state transition rate* from state i to state j, $i, j \in \mathcal{S}$, is defined as

$$q(i,j) = \begin{cases} \lim_{\tau \to 0} \dfrac{P[X(t+\tau) = j | X(t) = i]}{\tau}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{1}$$

and we define the *total transition rate* from state i as

$$q(i) \equiv \sum_{j \in \mathcal{S}} q(i,j). \tag{2}$$

The stationary distribution of X is denoted by $\pi(i)$, $i \in \mathcal{S}$, and is equal to the fraction of time that the process spends in state i. For $t_1 < t_2 < \ldots < t_m$, $m \geq 1$, the joint distribution of X is defined to be

$$\mathcal{D}(i_1, i_2, \ldots, i_m; t_1, t_2, \ldots, t_m) \equiv P[X(t_1) = i_1, X(t_2) = i_2, \ldots, X(t_m) = i_m]. \quad (3)$$

The joint distribution is the total probability of the set of paths where X is found in state i_j at times $t_j, 1 \le j \le m$.

The *probability flux* from state i to state j is defined as

$$\mathbf{F}(i, j) \equiv \pi(i)q(i, j), \quad i \in \mathcal{S} \quad (4)$$

and is equal to the time average transition rate out of state i to state j. More generally, we define the probability flux between two subsets of $\mathcal{S}$, $\mathcal{U}$, and $\mathcal{V}$, as

$$\mathbf{F}(\mathcal{U}, \mathcal{V}) = \sum_{u \in \mathcal{U}, v \in \mathcal{V}} \mathbf{F}(u, v). \quad (5)$$

The time reversal of a stationary Markov process X on the state space about the time τ is defined to be the process $X^r(t) \equiv X(\tau - t)$ and corresponds to viewing the process backward in time about the pivot τ. Under the given assumptions of X it can be shown that X^r is also a time homogeneous, irreducible, and stationary Markov process [Kelly 1979]. These assumptions imply that selection of τ is arbitrary since, as in the forward process, properties of the reverse process do not depend on absolute time. We henceforth set the pivot of time reversal to be $\tau = 0$ and will call process $X(t)$ and $X^r(t) \equiv X(-t)$ the *forward* and *reverse* processes, respectively.

We let $q^r(i, j)$ and $q^r(i), i, j \in \mathcal{S}$, be the transition rates and total transition rate, respectively, of the reverse process and let $\pi^r(i)$ be its stationary distribution. In general, the transition rates of the reverse process differ from that of the forward process. As an example, suppose that $q(i, j) \neq 0$ and $q(j, i) = 0$ for some process. Viewing the process in reverse time shows that $q^r(j, i) \neq 0$, and thus the forward and reverse transition rates are not equal. Since reversing time does not affect the fraction of time a process spends in a state, the stationary distributions of both the forward and reverse process are identical, i.e., $\pi^r(i) = \pi(i)$, $i \in \mathcal{S}$. The joint distribution of the reverse process and probability flux of the reverse process are defined similarly to that of the forward process, i.e.,

$$\mathcal{D}^r(i_1, i_2, \ldots, i_m; t_1, t_2, \ldots, t_m) \equiv P[X^r(t_1) = i_1, X^r(t_2) = i_2, \ldots, X^r(t_m) = i_m] = P[X(-t_m) = i_m, X(-t_{m-1}) = i_{m-1}, \ldots, X(-t_1) = i_1] \quad (6)$$

and

$$\mathbf{F}^r(i, j) \equiv \pi^r(i)q^r(i, j) = \pi(i)q^r(i, j), i, j \in \mathcal{S}, \quad (7)$$

with a similar form of probability flux for sets (5). We note that, by definition, the joint distributions for the forward and reverse process satisfy

$$\mathcal{D}(i_1, i_2, \ldots, i_m; t_1, t_2, \ldots, t_m) = \mathcal{D}^r(i_m, i_{m-1}, \ldots, i_2, i_1; -t_m, -t_{m-1}, \ldots, -t_2, -t_1). \quad (8)$$

It is important to note that equality of the stationary probabilities for the forward and reverse process does not imply equivalence of the joint distribution of these processes. The joint distribution is a more precise characterization of the process since it specifies the probability of evolving along a certain set of paths rather than simply specifying the fraction of time the process spends in a given set of states.

Processes that have the property that the joint distribution of the forward and reverse process are equal are said to be *reversible*. Specifically, a process is reversible if

$$\mathcal{D}(i_1, i_2, \ldots, i_m; t_1, t_2, \ldots, t_m) = \mathcal{D}^r(i_1, i_2, \ldots, i_m; t_1, t_2, \ldots, t_m) \quad (9)$$

$$= \mathcal{D}(i_m, i_{m-1}, \ldots, i_1; -t_m, -t_{m-1}, \ldots, -t_1). \quad (10)$$

This definition implies that reversible

processes are statistically identical in forward or reverse time and implies that $\pi(i)q(i, j) = \pi^r(i)q^r(i, j)$ and consequently that $q(i, j) = q^r(i, j)$. We will frequently make use of this characteristic of reversible processes to derive results that are difficult to obtain through other means. We record here that, because of this statistical equivalence, the probability flux of the forward and reverse process must be equal, i.e.,

$$\mathbf{F}(\mathscr{U}, \mathscr{V}) - \mathbf{F}^r(\mathscr{U}, \mathscr{V}) = 0.$$

Reversibility Balance Equations (11)

2.2 Balance Equations for the Forward and Reverse Process

To determine the stationary distribution of the process we must find values of $\pi(i), i \in \mathscr{S}$, that satisfy *global balance equations*. These equations can be expressed as a conservation law of probability flux and are given by

$$\mathbf{F}(\mathscr{U}, \bar{\mathscr{U}}) - \mathbf{F}(\bar{\mathscr{U}}, \mathscr{U}) = 0,$$

Global Balance Equations (12)

where $\mathscr{U}$ is any set in $\mathscr{S}$ and where $\bar{\mathscr{U}}$ is its complement with respect to $\mathscr{S}$. A solution to (12) for all $\mathscr{U}$ that is normalized to sum to unity is the *unique* stationary distribution of the process [Ross 1983]. A convenient way to represent (12) is shown in Figure 1. An arc from state i to state j in this figure is assumed to have a directed probability flux on it equal to $\mathbf{F}(i, j)$. Equation (12) shows that the probability flux into and out of any subset of states is equal. By careful selection of set $\mathscr{U}$, we can sometimes use the structure of a Markov process to determine a possible solution (a guess) of the global balance equations and then normalize it to sum to unity. The solution can then be checked by showing that it satisfies (12).

As an example of such a procedure, consider a *birth–death* process. Birth transitions in state i occur at rate $\lambda_i, i \geq 0$, and death transitions in state i occur at rate $\mu_i, i > 0$. We will think of this process as a queuing system in which the action of the scheduling and servicing policy of the queue in state i is such that customers arrive at rate λ_i and depart at rate $\mu_i (\mu_0 = 0)$. One context for such a system is a first-come first-serve, single-server queue where customers require a unit exponential service and the server works at rate μ_i when there are i customers in the queue. Other service and scheduling assumptions also lead to the same Markov process.

Setting $\mathscr{U} = \{0, 1, \ldots, i - 1\}, i \geq 1$, and using the global balance equations (12) show that

$$\mathbf{F}(\mathscr{U}, \bar{\mathscr{U}}) - \mathbf{F}(\bar{\mathscr{U}}, \mathscr{U}) = \mathbf{F}(i - 1, i) - \mathbf{F}(i, i - 1) = 0, \quad i \geq 1, \quad (13)$$

and thus that

$$\frac{\pi(i)}{\pi(i - 1)} = \frac{\lambda_{i-1}}{\mu_i}, \quad i \geq 1. \quad (14)$$

Equations (14) suggest that $\pi(i)$ has the form of a product of the ratio of transition rates. We guess a solution of the form

$$\psi(i) = \begin{cases} 1, & i = 0, \\ \prod_{k=1}^{i} \lambda_{k-1}/\mu_k, & i \geq 1. \end{cases} \quad (15)$$

Normalizing it with

$$\sum_{i=0}^{\infty} \prod_{k=1}^{i} \lambda_{k-1}/\mu_k \quad (16)$$

implies that the guessed stationary probabilities are given by

$$\pi(i) = \frac{\prod_{k=1}^{i} \lambda_{k-1}/\mu_k}{\sum_{i=0}^{\infty} \prod_{k=1}^{i} \lambda_{k-1}/\mu_k}, \quad i \geq 0, \quad (17)$$

(a product with an empty range is defined to be equal to 1). It is easy to check (17) to show that it satisfies the global balance equations and that the probabilities sum to 1. Thus (17) is the stationary distribution of the process.

We shall shortly see that, like birth–

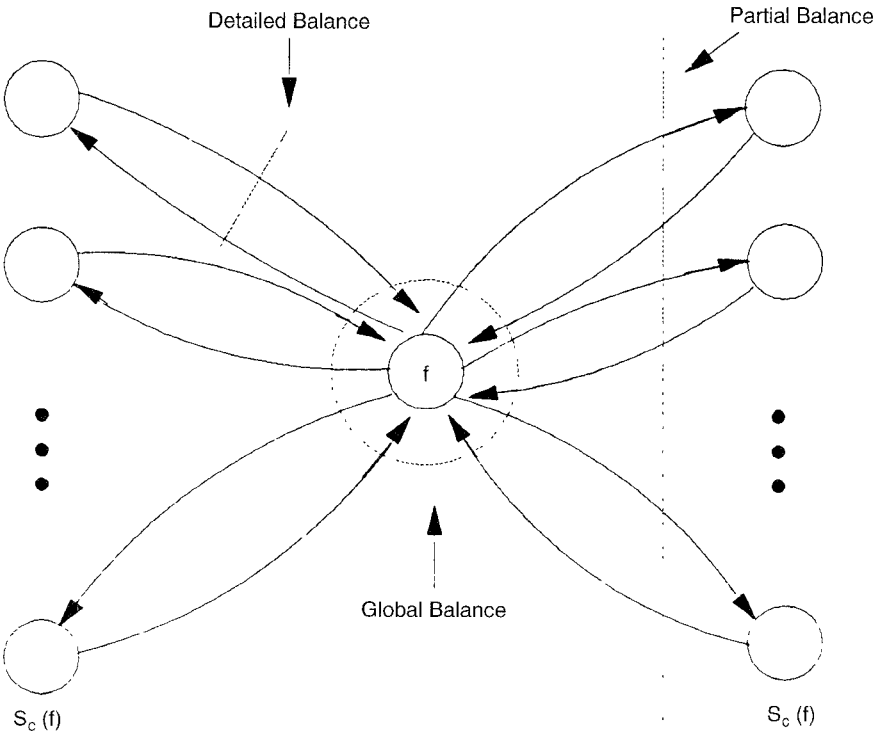

Figure 1. Balance equations.

death processes, all reversible Markov processes have solutions that can be written as products of the ratios of transitions rates (15). Before proceeding to this, we first derive equations satisfied by the reverse transition rates. Use (8) to write

$$\begin{aligned} P[X(t+dt)=k, X(t)=i] &= P[X^r(-t)=i, \\ &\quad X^r(-t-dt)=k] \qquad (18) \\ &= P[X^r(t+dt)=i, X^r(t)=k], \qquad (19) \end{aligned}$$

where (19) arises from (18) by shifting the process by $2t + dt$ time units. Since the process is time homogeneous and stationary, this shift in time does not change the joint probability. We rewrite (19) as

$$\begin{aligned} \pi(i)P[X(t+dt)=k|X(t)=i] &= \pi^r(k)P[X^r(t+dt) \\ &= i|X^r(t)=k] \qquad (20) \end{aligned}$$

which, after dividing by dt and letting $dt \to 0$, and using (1) implies that

$$\pi(i)q(i,k) = \pi(k)q^r(k,i). \qquad (21)$$

Summing this equation over k and using global balance for the reverse process also shows that

$$q(i) = q^r(i). \qquad (22)$$

It is convenient to record the result of (21) in terms of probability flux for any two disjoint subsets $\mathscr{U}, \mathscr{V} \in \mathscr{S}$,

$$\mathbf{F}(\mathscr{U}, \mathscr{V}) - \mathbf{F}^r(\mathscr{V}, \mathscr{U}) = 0$$

Reverse Balance Equations (23)

Equations (21–23) must be satisfied by the reverse transition rates. They show that we can deduce the reverse transition rates after we know the stationary distribution, or alternatively we can deduce the stationary distribution after we know the reverse transition rates. Since both

the stationary distribution and reverse transition rates are not known before hand, it might appear at first that the reverse balance equations are not practically useful. But suppose we can simultaneously guess values for *both* the reverse transition rates and the stationary probabilities that satisfy (21) and (22) (and consequently that also satisfies (23)). Does our guess correspond to the actual values of the stationary probability distribution and the reverse transition rates (i.e., are the guessed values of $\pi(i)$ and $q'(i,j)$ correct)? To see that they are correct, we sum (21) over i

$$\sum_{i \in \mathcal{S}} \pi(i)q(i,k) = \sum_{i \in \mathcal{S}} \pi(k)q^r(k,i) \tag{24}$$

$$= \pi(k) \sum_{i \in \mathcal{S}} q(k,i) \quad \text{From (22)}, \tag{25}$$

and thus the guessed values of $\pi(i)$ and $q^r(i,j)$ satisfy global balance.

Often it is the case that the structure of the reverse process is evident by imagining the process running backward in time, and we can often guess a form for the structure of the stationary probabilities. In such cases, algebraic experimentation can often be used to hone in on the solutions for the stationary probabilities and reverse transition rates, and then the reverse balance equations can be used to validate the derived solution. Although it might seem futile to attempt heuristic guessing to determine solutions for complex processes, we will later derive the stationary distribution for the entire family of quasireversible queuing networks in exactly this manner. The surprising thing about obtaining a result of this magnitude in this ad hoc manner is that the proposed guess is motivated largely by the desire to have a tractable solution rather than by any deep, penetrating insight. The reader has perhaps already anticipated, from the introductory material, the form of the guess.

2.3 Reversibility and Detailed Balance

We can deduce the reverse transition rates for one class of Markov processes. Comparing the reversibility balance equations (11) to the reverse balance equations (23) shows that for reversible processes

$$\mathbf{F}(\mathcal{U}, \mathcal{V}) = \mathbf{F}(\mathcal{V}, \mathcal{U}),$$

Detailed Balance Equations (26)

and thus that

$$\pi(i)q(i,j) = \pi(j)q(j,i), \quad i,j \in \mathcal{S}. \tag{27}$$

It can be shown that detailed balance is a necessary and sufficient condition for a process to be reversible [Kelly 1979].

The stationary probabilities for these reversible processes can be obtained in the same manner as in the birth–death example. Pick a starting state $s \in \mathcal{S}$, and for each state $i \in \mathcal{S}$ find *any* sequence of states $s = j_{i,1}, j_{i,2}, \ldots, j_{i,m_i} = i, m_i \geq 1$, so that $q(j_{i,k}, j_{i,k+1}) > 0, 1 \leq k \leq m_i - 1$. Such a sequence exists for all states $i \in \mathcal{S}$ since X is irreducible. Analogous to (15) and (17) let

$$\psi(i) = \begin{cases} 1, & i = s, \\ \prod_{k=1}^{m_i} \alpha_i(k), & i \neq s, \quad i \in \mathcal{S}, \end{cases} \tag{28}$$

where

$$\alpha_i(k) \equiv \frac{q(j_{i,k}, j_{i,k+1})}{q(j_{i,k+1}, j_{i,k})}, \quad k = 1,2,\ldots,m_1 - 1, \tag{29}$$

is the ratio of the forward to the reverse rate for step k along the sequence of states selected for the ith state. Normalizing (28) we obtain

$$\pi(i) = \frac{\prod_{k=1}^{m_i} \alpha_i(k)}{\sum_{i \in \mathcal{S}} \prod_{k=1}^{m_i} \alpha_i(k)}, \quad i \in \mathcal{S}. \tag{30}$$

Thus the form of the solution for the stationary distribution for reversible processes is a ratio of products of rates as in the *birth–death* example. Since we can

pick any sequence of states to derive (30), it follows that the rates along any *spanning tree* of the state space of a reversible process are sufficient to determine its stationary distribution. This observation leads to some interesting conclusions. For example, for a given spanning tree it shows that changes to transition rates not contained in the spanning tree that keep the process reversible do not affect the stationary distribution of the process.

Detailed balance is a restrictive conservation law that is graphically shown in Figure 1. From it we can derive a way to determine if a Markov process is reversible from inspection of its transition rates. Let $j_1, j_2, \ldots j_{m+1}, m \geq 0$, be any sequence of states that satisfies $j_{m+1} = j_1$. If the process is reversible, then detailed balance implies that

$$\pi(j_k)q(j_k, j_{k+1}) = \pi(j_{k+1})q(j_{k+1}, j_k), \quad 1 \leq k < m, \quad (31)$$

and

$$\pi(j_m)q(j_m, j_1) = \pi(j_1)q(j_1, j_m). \quad (32)$$

Multiplying the left and right-hand sides of (31) for $1 \leq k \leq m$ and (32) and cancelling the common probability factors implies that

$$q(j_1, j_2)q(j_2, j_3)\ldots q(j_m, j_1) = q(j_1, j_m)q(j_m, j_{m-1})\ldots q(j_2, j_1) \quad (33)$$

are satisfied if the process is reversible. Another argument can be used to establish that if (33) is satisfied for all sequences of states, then the process is reversible. This criterion is called *Kolmogorov's criterion* [Kolmogorov 1936] and is often used to establish the reversibility of a process. Each side of (33) can be thought of as a flow of transition rates along one direction, and the equality thus implies that there is no net circulation of this flow in the state space. It immediately follows that all Markov processes that have a state transition diagram that forms a tree with bidirectional arcs, regardless of transition values, are reversible. Here Equation (33) is trivially satisfied.

Returning to the birth–death example, since the state space is a tree, it immediately follows that the process is reversible. Thus the process is statistically identical in forward and reverse time. We apply this result to establish that the arrival and departure processes from the queue are statistically identical. The arrival and departure processes are defined to be the times at which customers join and leave the queue, respectively. Arrivals and departures in the forward process cause X to increase or decrease by one customer, respectively. An increase (decrease) of one customer in forward time, however, corresponds to a decrease (increase) of one customer in reverse time. Thus, arrivals and departures of the forward process correspond to departures and arrivals in the reverse process, respectively. Since both the forward and reverse processes are statistically identical, the arrival process of the forward process is statistically identical to the arrival process of the reverse process. The correspondence of reverse arrivals and forward departures thus implies that the arrival and departure processes of the forward process are statistically identical, as claimed. There is a striking contrast to the difficulty in obtaining this result using algebraic techniques [Burke 1956; 1972; 1976] with the ease of the above argument.

To continue the above line of reasoning, assume that arrival rates are independent of the state, $\lambda_i = \lambda, i \geq 0$. It then follows that the arrival and departure processes are both Poisson with rate λ. This conclusion initially appears to violate intuition since it is invariant to the selection of the values of the service rates as long as the system permits a stationary distribution. Although it is clear that the average customer departure rate must be λ since the queue is stationary, it is not clear that the departure process must have independent interdeparture intervals. We are led to believe that we could arrange service rates so as to force departures to occur in clusters. For ex-

ample, consider setting the service rates as follows

$$\mu_i = \begin{cases} \epsilon, & 1 \le i \le N, \\ \mu, & N < i, \end{cases} \tag{34}$$

for a given value of $N, 0 \le N < \infty$, and $\epsilon > 0$. The system is stationary provided that $\lambda/\mu < 1$. For any value of N and $\delta, 0 < \delta < 1$, we can select ϵ so that the probability of having at least N customers in the queue is greater than δ. As an example, suppose we select $N = 10^{134}$ and $\delta = 1 - 10^{-431}$. Thus the process has fewer than 10^{134} customers less than $10^{-431} \times 100$ percent of the time. This seems to imply that the departure process would frequently consist of a series of exponential interdeparture intervals at rate μ when the queue length is greater than N with exponential interdepartures at rate ϵ during the infrequent times that the queue length was below N. How then can the departure process have independent interdeparture intervals and be Poisson at rate λ as we have already shown?

Intuition leads us astray in the above argument since it misses a subtle dependency. The departure process is determined not by the stationary distribution, but rather by the joint queue length distribution. But the departure process and the time during which the queue length is greater than N (which is also determined by the joint queue length distribution) are not independent of each other. The "intuitive argument" presented above ignores this dependency by focusing only on the nature of the stationary distribution.

2.4 Quasireversibility

One important property arises when we consider the birth–death example in the case where the arrival rates are independent of the state of the system, i.e., $\lambda_i = \lambda, i \ge 0$. Since we will have recourse to discuss this process repeatedly in the article, we will call it the *constant arrival birth–death process*. Suppose that, at a random time, we observe both the state of the system and its future-arrival stream. Clearly these are independent since, for all states, arrivals are Poisson with rate λ. But since arrivals in the forward process correspond to departures in the reverse process and since the process is reversible, it must be the case that the state of the system at a random time is independent of the departure process prior to that time, and thus both the arrival and departure processes are Poisson. It is important here to note that if arrivals *are* state dependent, then the state of the system at a random time and its future arrivals *are not* independent (here the probability of an arrival within dt seconds in state i is $\lambda_i dt$ which clearly is state dependent).

A new property, that was first identified by Kelly [1976a], emerges from the above discussion, which captures a type of independence of both the arrival and departure processes from the state of the system [Burke 1956; Muntz 1972; 1973]. We will define this property within the context of a multiple-class queue. Suppose there are $C, C \ge 1$, classes of customers that arrive and are serviced by a queuing system. The Markov process associated with this system is said to be *quasireversible* if the state of the process, for all $c, 1 \le c \le C$, at time t is independent of the arrival process of class c customers after time t and is also independent of the departure process of class c customers prior to time t. Note that this definition, with the identification of arrivals (resp., departures) in the forward process with departures (resp., arrivals) in the reverse process, implies that the reverse process of a quasireversible queue is also quasireversible. If we think of a queue as a filter which takes a set of input processes and produces a set of output processes, then, as shown below, quasireversible queues have the property that Poisson streams pass through such a filter statistically unchanged (this is the $M \Rightarrow M$ property of Muntz [1972]). Queuing systems are typically easier to solve if governed by Poisson arrival and departure processes. This, with the fact that the state of a quasireversible queue

is independent of both of these processes, suggests that a network of such queues would also have states that were mutually independent. These observations provide the first glimpse as to why input-output relationships lead to queuing networks that have tractable analysis [Kelly 1976a; Muntz 1972; 1973].

The above arguments show that the constant-arrival birth–death process is quasireversible whereas the birth–death process with state-dependent arrivals is not. Both systems, it is important to note, are reversible. It is easy to construct examples of Markov processes that are quasireversible but not reversible. Consider the birth–death queue with constant arrival rates where we "split" state 1 into two states, say states 1_a and 1_b. Set the transition from state 0 to state 1_a (resp., state 1_b) to be equal to $p\lambda$ (resp., $(1-p)\lambda$) and the transition from state 2 to state 1_a (resp., state 1_b) to be equal to $(1-p)\mu$ (resp., $p\mu$). The transition rates from state 1_a and 1_b are similar to the original birth–death queue (i.e., the rate from these states to state 0 (resp., 2) is given by μ (resp., λ)). It is simple to see that the sum of the stationary probabilities of state 1_a and 1_b in this modified process is equal to this stationary probability of state 1 in the original birth–death process. It is also clear that this splitting of state 1 does not influence the departure process from the system. The modified process, however, is not reversible. This is easily seen from Kolmogorov's criteria by comparing the product of transition rates in both directions of the cycle $0 \to 1_a \to 2 \to 1_b \to 0$. Equality of these two products is only achieved if $p = 1/2$. These examples show that reversibility and quasireversibility are entirely separate notions even though the word "quasireversibility" seems to imply a superset relationship with "reversibility." In general, quasireversibility (resp., reversibility) does not imply reversibility (resp., quasireversibility).

We have now established that the constant-arrival birth–death process is quasireversible and also has Poisson arrival and departure processes that are independent of the state of the queue. We now show that this input-output property is shared by all quasireversible queues. The details of the following derivation were first presented by Muntz [1972] and the approach here follows Kelly [1979]. Let t be a random time, and let $\mathcal{S}_c(i), 1 \le c \le C, i \in \mathcal{S}$, be the set of states that contain one more class c customer than in state i with the same number of customers of other classes. The arrival rate of class c customers given that $X(t) = i$ is given by

$$\lambda(c,i) = \sum_{k \in \mathcal{S}_c(i)} q(i,k), \quad i \in \mathcal{S}, \tag{35}$$

and thus the average arrival rate of class c customers is given by

$$\lambda(c) = \sum_{i \in \mathcal{S}} \pi(i)\lambda(c,i). \tag{36}$$

By definition of quasireversibility, however, the arrival process of class c customers subsequent to t is independent of the state at time t, and thus $\lambda(c,i)$ is independent of i. Using (35) and (36) we write

$$\lambda(c) = \sum_{k \in \mathcal{S}_c(i)} q(i,k), \tag{37}$$

where i is any state in $\mathcal{S}$. The probability of an arrival of a class c customer within $(t, t+dt)$ is independent of any event prior to time t and is given by $\lambda(c)dt$, which shows that the arrival of class c customers is a Poisson process.

Assume that all arriving customers leave the system and that the queue is in equilibrium. This, combined with the identification of arrivals (departures) of the forward process with departures (arrivals) of the reverse process and the fact that the queue viewed in reverse time is also quasireversible, implies that the departure process is also Poisson with rate $\lambda(c)$. This argument also shows that

$$\lambda(c) = \sum_{k \in \mathcal{S}_c(j)} q^r(j,k), \quad j \in \mathcal{S}. \tag{38}$$

We can obtain substantially more from the above argument. The reverse balance equations (23) show that

$$\pi(i) \sum_{k \in \mathcal{S}_c'(i)} q^r(i,k) = \sum_{k \in \mathcal{S}_c'(i)} \pi(k) q(k,i). \tag{39}$$

Using (37) and (38) in (39) shows that

$$\pi(i) \sum_{k \in \mathcal{S}_c'(i)} q(i,k) = \sum_{k \in \mathcal{S}_c'(i)} \pi(k) q(k,i) \tag{40}$$

and also (obtained by subtracting this from a global balance) that

$$\pi(i) \sum_{k \in \overline{\mathcal{S}_c'(i)}} q(i,k) = \sum_{k \in \overline{\mathcal{S}_c'(i)}} \pi(k) q(k,i). \tag{41}$$

Rewriting in terms of probability flux we have for all $c, 1 \le c \le C$,

$$\mathbf{F}(i, \mathcal{S}_c'(i)) - \mathbf{F}(\mathcal{S}_c'(i), i) = 0,$$

$$\mathbf{F}\left(i, \overline{\mathcal{S}_c'(i)}\right) - \mathbf{F}\left(\overline{\mathcal{S}_c'(i)}, i\right) = 0, \quad i \in \mathcal{S}'. \tag{42}$$

Partial Balance Equations

These equations imply that the probability flux due to arrivals of class c jobs from a state i is equal to the probability flux due to departures of class c jobs that result in state i. In contrast to detailed balance which was necessary and sufficient for the reversibility of a process, partial balance is only a necessary condition for quasireversibility. There are processes that satisfy partial balance that are not quasireversible (e.g., those that do not have Poisson arrival and departure processes). The essence of product form, as will be seen later, is found in the partial balance equations. Networks of quasireversible queues have product form solutions because they also satisfy partial balance. It is important to note, however, that product form can exist in systems that do not satisfy quasireversibility.

To discuss other versions of partial balance, let $\mathcal{V}_c'(i)$ be the set of states that have one less class c customer than state i, and let $\mathcal{Y}(i)$ be the set of states that have the same number of class c' customers as state i for $c' = 1, 2, \ldots, C$. Transitions between state i and set $\mathcal{Y}(i)$ will be termed *internal transitions* since they can be viewed as transitions within a queue that do not change the number of its customers. We will also term *external transitions* as being those that correspond to external arrival or departure events. If all transitions cause class changes for some class, i.e., if all transitions are external transitions, then set $\mathcal{Y}(i)$ is empty. State transition from i are contained in the set $\{\mathcal{Y}(i), \mathcal{S}_c'(i), \mathcal{V}_c'(i)\}_{c=1}^{C}$. Global balance implies that

$$\mathbf{F}\left(i, \{\mathcal{Y}(i), \mathcal{S}_c'(i), \mathcal{V}_c'(i)\}_{c=1}^{C}\right) - \mathbf{F}\left(\{\mathcal{Y}(i), \mathcal{S}_c'(i), \mathcal{V}_c'(i)\}_{c=1}^{C}, i\right) = 0, \quad i \in \mathcal{S}'. \tag{43}$$

Summing (42) over all c implies that partial balance holds for the set

$$\{\mathcal{S}_c'(i)\}_{c=1}^{C} \text{ and that}$$

$$\mathbf{F}\left(i, \{\mathcal{S}_c'(i)\}_{c=1}^{C}\right) - \mathbf{F}\left(\{\mathcal{S}_c'(i)\}_{c=1}^{C}, i\right) = 0.$$

Using this in (43) shows that for systems that satisfy partial balance, the following balance equation also holds

$$\mathbf{F}\left(i, \{\mathcal{Y}(i), \mathcal{V}_c'(i)\}_{c=1}^{C}\right) - \mathbf{F}\left(\{\mathcal{Y}(i), \mathcal{V}_c'(i)\}_{c=1}^{C}, i\right) = 0, \quad i \in \mathcal{S}'. \tag{44}$$

The partial balance equations (42) along with (44) are sometimes collectively termed *local balance* equations. Equation (44) implies that the probability flux due to internal transitions and departures of customers from state i is equal to the probability flux due to internal transitions and arrivals that result in state i.

A more restrictive form of balance holds when *station balance* equations are satisfied. These equations are given by

$$\mathbf{F}(i, \mathcal{Y}(i)) - \mathbf{F}(\mathcal{Y}(i), i) = 0 \tag{45}$$

$$\mathbf{F}(i, \mathcal{V}_c'(i)) - \mathbf{F}(\mathcal{V}_c'(i), i) = 0 \quad i \in \mathcal{S}'. \tag{46}$$

Station Balance Equations

Equation (45) implies that the probability flux due to internal transitions is balanced, and (46) implies that the probability flux due to departures of class c customers from state i is equal to the probability flux due to arrivals of class c customers that result in state i. It is clear that if either partial balance or station balance is satisfied then (44) holds. Equation (44), however, can hold without either partial balance or station balance being satisfied. Station balance (resp., partial balance) does not imply that partial balance (resp., station balance) holds.

As an example where partial balance is satisfied but station balance does not hold, consider a multiple class M/M/K queue with C classes of customers. Assume that class c customers arrive with a rate of λ_c and have an exponential service rate of μ. The state of the system can be written as $\underline{s}$ where $\underline{s} = (s_1, s_2, \ldots, s_K, s_{K+1}, \ldots)$ represents classes of customers resident in the system. Here the first K components of $\underline{s}$ correspond to the classes of customers that are being served, and the remaining components correspond to the classes of customers that wait in the queue (we set $s_i = 0$ if no customer exists in position i). Let the total arrival rate of customers of all classes be given by $\lambda = \sum_{c=1}^{C} \lambda_c$, and let $p_c = \lambda_c/\lambda$ be the probability that an arrival is of class c. Let $\alpha(n)$ be the stationary probability of having n customers in a single class M/M/K queue with an arrival rate of λ. Then a simple calculation shows that the stationary probability is given by

$$\pi(\underline{s}) = \alpha(|\underline{s}|) \prod_{i=1}^{\infty} p_{c_i} \qquad (47)$$

where $|\underline{s}|$ is the number of customers in state $\underline{s}$. It can be shown algebraically that this queue satisfies partial balance. Station balance, however, is not satisfied. To see this consider a state

$$\underline{s} = (\underbrace{c, c, \ldots c}_{K}, s_{K+1}, s_{K+2}, \ldots). \qquad (48)$$

In this state all servers are processing a class c customer. A departure of a class c customer then results in a state

$$\underline{s}' = (\underbrace{c, c, \ldots c}_{K-1}, s_{K+1}, s_{K+2}, \ldots). \qquad (49)$$

It is clear that if $s_{K+1} \neq c$ no arrival of a class c can result in a $\underline{s}' \rightarrow \underline{s}$ transition, and thus station balance cannot be satisfied.

Product form solutions, as we will later see, arise from systems that satisfy partial balance. We show that one such class of networks are those consisting of quasireversible queues and in the next section explore the properties of partial balance that will allow us to develop such results. Consequences of station balance include insensitivity of the stationary distribution to higher moments of service time. These results lie outside the scope of this article and can be found in Chandy et al. [1977], Disney and Kiessler [1987], and Jansen and König [1980].

2.5 Partial Balance

The notion of customer classes is a useful paradigm within which to couch partial balance, but it can be expressed for arbitrary sets. More generally (see Whittle [1967; 1968] for the first definition of partial balance) we say that partial balance holds on set $\mathscr{Y}$ if

$$\mathbf{F}(i, \mathscr{Y}) - \mathbf{F}(\mathscr{Y}, i) = 0$$
$$\mathbf{F}(i, \bar{\mathscr{Y}}) - \mathbf{F}(\bar{\mathscr{Y}}, i) = 0 \quad \forall i \in \mathscr{Y}. \qquad (50)$$

Observe that global balance for state i results by summing these two equations. It is important to note that if partial balance holds over set $\mathscr{Y}$ then it does not necessarily hold over all possible sets. We will always specify which sets we mean when discussing partial balance and make it the convention that for quasireversible systems we mean partial balance as given by the sets $\mathscr{S}_c(i)$ as specified in (42). The relationship between partial, local, and station balance for a two-class system is shown in Figure 2. Since partial balance and quasireversibility play key roles in the rest of the article, we spend some time here to

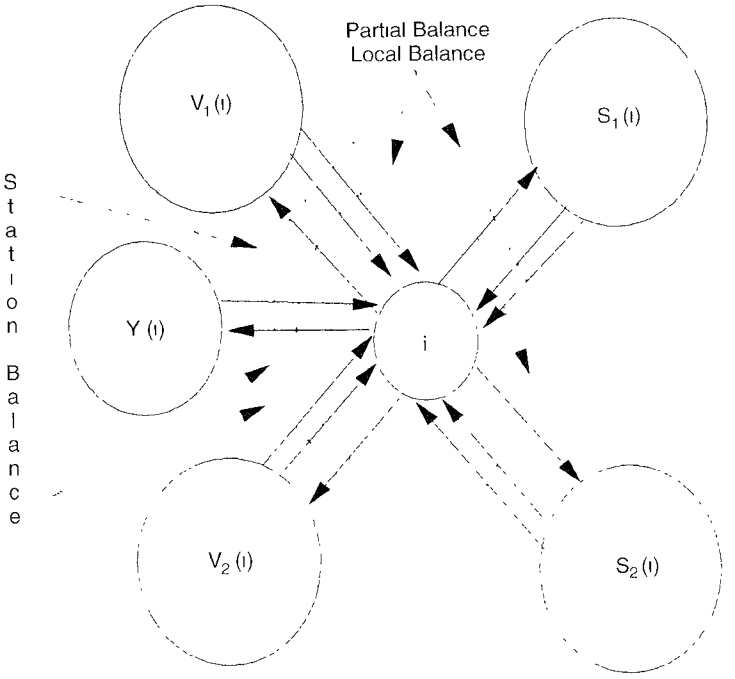

Figure 2. Partial, local, and station balance equations

explore some of their properties. The first property below requires that the process is quasireversible. The following three properties hold generally for all processes that satisfy partial balance. We make the convention that when we say a quasireversible queue satisfies partial balance, we mean that for all states $i, i \in \mathcal{S}$, Equation (42) is satisfied. When partial balance is said to hold over a set $\mathcal{U} \subset \mathcal{S}$ we mean that Equation (50) is satisfied over $\mathcal{U}$.

The Distribution Property

The first property requires that quasireversibility, and thus also that partial balance, is satisfied. Suppose we observe an arrival of a class c customer to a quasireversible queue that is in state i. We know that this arrival causes a state transition to some state $i' \in \mathcal{S}_c(i)$, and here we wish to derive its distribution. This probability is written as

$$\begin{aligned} &P[X(t+dt) = i' | X(t) = i, \text{Arrival of class } c \text{ in } dt] \\ &\quad = \frac{P[X(t+dt) = i', X(t) = i, \text{Arrival of class } c \text{ in } dt]}{P[X(t) = i, \text{Arrival of class } c \text{ in } dt]} \\ &\quad = \frac{\pi(i) q(i,i') dt}{\pi(i) \lambda(c) dt} = \frac{q(i,i')}{\lambda(c)}, \end{aligned} \tag{51}$$

where we have used the fact that the arrival rate of class c customers is independent of the state of the system (37). We call this the *distribution property* of quasireversible queues. Notice that this property holds only for quasireversible queues and does not hold in general for processes that satisfy partial balance.

Application of the Distribution Property

This property is, in some sense, the mechanism that allows us to join quasi-

reversible queues together into a network and still preserve that fact that the resultant network is quasireversible. We will see in Section 3 that this implies that the network has a product form solution.

The State Truncation Property

The second property of partial balance arises when we truncate the state space of the process. Let $\mathscr{A}$ be a set of states and consider the process $\mathscr{Y}$ that restricts X to $\mathscr{A}$ by setting

$$q_Y(i,k) = \begin{cases} q(i,k), & i,k \in \mathscr{A}, \\ 0, & i \in \bar{\mathscr{A}} \text{ or } k \in \bar{\mathscr{A}}. \end{cases} \tag{52}$$

If partial balance holds over $\mathscr{A}$ and if Y is irreducible then

$$\mathbf{F}(i,\mathscr{A}) - \mathbf{F}(\mathscr{A},i) = 0, \quad i \in \mathscr{A}, \tag{53}$$

and it is easy to see that (53) is unchanged if $\pi(i)$ is replaced by

$$\pi_Y(i) = C\pi(i), \quad i \in \mathscr{A}, \tag{54}$$

where $C = 1/\sum_{1 \in \mathscr{A}} \pi(i)$ is a normalizing constant. Since Y is restricted to $\mathscr{A}$, this shows that $\pi_Y(i)$ is its stationary distribution. Conversely, suppose that (54) is Y's stationary distribution and thus satisfies the global balance equations,

$$\pi_Y(i) \sum_{J \in \mathscr{A}} q_Y(i,j) = \sum_{J \in \mathscr{A}} \pi_Y(j) q_Y(j,i), \quad i \in \mathscr{A}. \tag{55}$$

Global balance for X can be written as

$$\pi(i)\left\{ \sum_{J \in \mathscr{A}} q(i,j) + \sum_{k \in \bar{\mathscr{A}}} q(i,k) \right\} = \sum_{J \in \mathscr{A}} \pi(j) q(j,i) + \sum_{k \in \bar{\mathscr{A}}} \pi(k) q(k,i), \quad i \in \mathscr{A}. \tag{56}$$

Substituting (52 and 54) into (55) and subtracting it from (56) shows that partial balance is satisfied. Thus, stationary probabilities are identical (up to a normalization) if a process is truncated to a set $\mathscr{A}$ if and only if partial balance holds over $\mathscr{A}$. We call this property the *state truncation* property of processes that satisfy partial balance.

Application of the State Truncation Property

Two applications of the state truncation property are finite-buffer models for communication networks (see Henderson and Taylor [1989] for an interesting example in communications) and closed, fixed-population queuing networks. As an example of a finite-buffer model, consider an infinite-server queue with Poisson arrivals at rate λ_c and exponential service times with expectations of $1/\mu_c$ for classes $c = 1, 2, \ldots, C$ (the results presented here also hold if the service times are generally distributed with expectation $1/\mu_c$; see Kelly's [1979] treatment of symmetric queues for a sketch of a proof). It is well known that this finite-buffer system satisfies partial balance and that the stationary distribution that there are n_c class c customers in the system is given by

$$\pi(\underline{n}) = \prod_{c=1}^{C} \frac{\rho_c^{n_c}}{n_c!} e^{-\rho_c}, \tag{57}$$

where $\rho_c \equiv \lambda_c/\mu_c$ and $\underline{n} = (n_1, n_2, \ldots, n_C)$. Suppose we restrict the total number of customers in the system, regardless of class, to be not greater than N. Customer arrivals to the system when there are N servers busy are assumed to be lost. Let $\mathscr{A}$ be the set of all feasible states that contain N customers or less. The stationary distribution for this system is then given by

The difficulty of analyzing truncated systems, as demonstrated in the above example, arises from the complexity of calculating the normalization constant over the truncated state space (i.e., the denominator in (58)).

For applications of the state truncation property to closed systems, consider an open multiple-class queuing network; let $\mathcal{N}_c(n_c) \subset \mathcal{S}$ be a subset of states having $n_c, n_c \geq 1$, class $c, c = 1, 2, \ldots, C$, customers; and let Y be the process X restricted to the set of states having $n_c, 1 \leq c \leq C$, class c customers. This restriction corresponds to a closed network where the underlying Markov process is truncated to the set

$$\mathcal{Z} = \bigcap_{c=1}^{C} \mathcal{N}_c(n_c). \tag{59}$$

Notice that (59) implies that

$$\bigcup_{i \in \mathcal{Z}} \mathcal{Y}(i) = \mathcal{Z}. \tag{60}$$

Suppose that Y is irreducible on $\mathcal{Z}$ and that X satisfies partial balance on set $\mathcal{Z}$. Using (60) this is equivalent to having (45) satisfied, which follows if station balance holds. Then the state truncation property implies that the stationary distribution of the closed system Y is a renormalization of the stationary distribution of the open system X.

Application of the Distribution and State Truncation Properties

We continue the closed-network application of the state truncation property. If the process Y defined on set $\mathcal{Z}$ is not irreducible or if X does not satisfy partial balance on $\mathcal{Z}$ then we cannot immediately apply station truncation. A closed system, however, can be thought of as being derived from an open system by modifying the open system's external state transitions while retaining the internal transitions of the open system. In particular, transitions in the open system that result in having more than n_c class c customers are assumed to be lost in the closed system; departures of class c in having more than n_c class c customers are eliminated; and transitions that reduce the number of class c customers to be less than n_c are modified so that they keep the number of class c customers equal to n_c in the closed system. We can thus think of arrivals in the open system as being "lost" in the closed system and departures of class c customers in the open system as causing an *immediate arrival* of a class c customer in the closed system. We require a precise specification of the transition rates of these immediate arrivals in the closed system. To do this, assume that the open system satisfies partial balance, and thus specifically that (44) is satisfied, and define the sets consisting of one less class c customer as

$$\mathcal{Z}_c \equiv \mathcal{N}_c(n_c - 1) \bigcap_{l=1, l \neq c}^{C} \mathcal{N}_l(n_l), \quad 1 \leq c \leq C. \tag{61}$$

Observe that in terms of the sets $\mathcal{Y}_c(i)$ we can write $\bigcup_{i \in \mathcal{Z}} \mathcal{Y}_c(i) = \mathcal{Z}_c$. We assume the immediate arrivals of the closed system, being similar to external arrivals of the open system, satisfy the distribution property. This implies that transitions for these immediate arrivals, which are denoted by $q_Y(\ )$, are given by

$$q_Y(i,k) = q(i,k) + \sum_{c=1}^{C} \sum_{j \in \mathcal{Z}_c} q(i,j) \frac{q(j,k)}{\lambda(c)}, \quad i, k \in \mathcal{Z}. \tag{62}$$

In words, (62) can be explained as follows. The value $q(i,k)$ in the first part of (62) corresponds to the internal transition rate found in the open process which is retained in the closed process. The second part of (62) accounts for external departure transitions from the open system which are altered in the closed process. Class c transitions from state $i, i \in \mathcal{Z}$, must first enter some state j in set $\mathcal{Z}_c$ which occurs at rate $q(i,j)$. This reduces the number of class c customers to $n_c - 1$ and in the closed system causes an immediate arrival of a class c customer. Using the distribution property, the probability that this new arrival causes a transition into state $k, k \in \mathcal{Z}$, is given by $q(j,k)/\lambda(c)$. Summing over all possibilities yields (62).

The closed network can thus be thought of as a truncation of the open process with modified external transitions that satisfy the distribution property. We now show that, as in the state truncation property, the stationary distribution for the closed process is simply a renormalization of the stationary distribution for the open process, i.e., that (54) is satisfied with $\mathscr{H}$ being set $\mathscr{Z}$. From (54) then it suffices to show that

$$\pi(i)\sum_{k\in\mathscr{Z}} q_Y(i,k) = \sum_{k\in\mathscr{Z}} \pi(k)q_Y(k,i), \qquad i\in\mathscr{Z}. \quad (63)$$

Using (62) we write the left-hand side of (63) as

$$\begin{aligned}
&\pi(i)\sum_{k\in\mathscr{Z}} q_Y(i,k)\\
&= \pi(i)\sum_{k\in\mathscr{Z}} q(i,k)\\
&\quad + \sum_{c=1}^{C}\pi(i)\sum_{j\in\mathscr{Z}_c} q(i,j)\sum_{k\in\mathscr{Z}}\frac{q(j,k)}{\lambda(c)} \qquad (64)\\
&= \pi(i)\sum_{k\in\mathscr{Z}} q(i,k)\\
&\quad + \sum_{c=1}^{C}\pi(i)\sum_{j\in\mathscr{Z}_c} q(i,j) \qquad (65)\\
&= \mathbf{F}\left(i,\{\mathscr{Z},\mathscr{Z}_c\}_{c=1}^{C}\right). \qquad (66)
\end{aligned}$$

Equation (65) follows from Equation (64) by an application of (37).

Using (62) we can write the right-hand side of (63) as

$$\begin{aligned}
&\sum_{k\in\mathscr{Z}}\pi(k)q_Y(k,i)\\
&= \sum_{k\in\mathscr{Z}}\pi(k)q(k,i)\\
&\quad + \sum_{c=1}^{C}\sum_{k\in\mathscr{Z}}\pi(k)\sum_{j\in\mathscr{Z}_c} q(k,j)\frac{q(j,i)}{\lambda(c)} \qquad (67)\\
&= \sum_{k\in\mathscr{Z}}\pi(k)q(k,i)\\
&\quad + \sum_{c=1}^{C}\sum_{k\in\mathscr{Z}}\sum_{j\in\mathscr{Z}_c}\pi(j)\\
&\qquad \times q^r(j,k)\frac{q(j,i)}{\lambda(c)} \qquad (68)\\
&= \sum_{k\in\mathscr{Z}}\pi(k)q(k,i)\\
&\quad + \sum_{c=1}^{C}\sum_{j\in\mathscr{Z}_c}\pi(j)q(j,i)\\
&\quad \times \sum_{k\in\mathscr{Z}}\frac{q^r(j,k)}{\lambda(c)} \qquad (69)\\
&= \mathbf{F}\left(\{\mathscr{Z},\mathscr{Z}_c\}_{c=1}^{C},i\right). \qquad (70)
\end{aligned}$$

Equation (68) follows from (67) from the reverse balance equations (21), and Equation (69) follows from (38). Equation (63) thus follows from the fact that (44) is assumed to hold in the original process. We call such a system a *closed quasireversible queue.*

The Arrival-Departure Property

The equation given in (50) permits an interesting probabilistic interpretation. Suppose one defines the point process Y_d by observing X just before making a transition that leaves a set $\mathscr{H}$. Let $\psi_{Y_d}(i), i\in\mathscr{H}$, be the probability that the system is in state i just prior to the transition. Similarly we let Y_a be the point process formed by observing X just after a transition into $\mathscr{H}$ and let $\psi_{Y_a}(i), i\in\mathscr{H}$, be its distribution. We will speak of the distribution of Y_d and Y_a as being the distribution of states as seen by a transition out of and into set $\mathscr{H}$, respectively. What is the relationship between these two distributions?

We can write the following equations for $i\in\mathscr{H}$,

$$\begin{aligned}
\psi_{Y_d}(i) &= \frac{\pi(i)\sum_{k\in\bar{\mathscr{H}}}q(i,k)}{\sum_{l\in\mathscr{H}}\pi(i)\sum_{k\in\bar{\mathscr{H}}}q(i,k)}\\
&= \frac{\mathbf{F}(i,\bar{\mathscr{H}})}{\mathbf{F}(\mathscr{H},\bar{\mathscr{H}})}, \qquad (71)
\end{aligned}$$

$$\psi_{Y_a}(i) = \frac{\sum_{k \in \bar{\mathscr{H}}} \pi(k) q(k,i)}{\sum_{k \in \bar{\mathscr{H}}} \pi(k) \sum_{l \in \mathscr{H}} q(k,i)} = \frac{\mathbf{F}(\bar{\mathscr{H}}, i)}{\mathbf{F}(\bar{\mathscr{H}}, \mathscr{H})}. \quad (72)$$

Global balance (12) shows that the denominators of (71) and (72) are equal. Partial balance shows that the numerators are also equal, and thus $\psi_{Y_d}(i) = \psi_{Y_a}(i), i \in \mathscr{H}$. It is easy to see, conversely, that if the above two distributions are equal then partial balance must hold. Thus the distribution as seen by transitions out of set $\mathscr{H}$ is identical to that seen by transitions into set $\mathscr{H}$ if and only if partial balance holds on set $\mathscr{H}$.

We call this property the *partial balance arrival-departure property*. Note that, in general, the distributions seen by transitions out of and into a set $\mathscr{H}$ that satisfies partial balance are not equal to the stationary distribution of the process. As a simple counterexample, assume that only one state in $\mathscr{H}$, say $u \in \mathscr{H}$, permits transitions out of $\mathscr{H}$. Then it must be the case that $\psi_{Y_d}(u) = 1$ which is clearly not equal to $\pi(u)$.

Quasireversible queues inherit the partial balance arrival-departure property since they satisfy partial balance. They also have the additional property that the distributions seen by transitions out of and into set $\mathscr{H}$ *are equal* to the stationary distribution. We show this by recasting the above argument in terms of customer classes and will talk of distributions seen by transitions out of (resp., into) a set $\mathscr{H}$ in terms of distributions seen by customer departures (resp., arrivals) that leave (resp., enter) set $\mathscr{H}$. Let $Z_a(\iota), i \in \mathscr{S}'$ be the point process seen by an arriving class c customer which, similar to above, has a distribution given by

$$\psi_{Z_a}(i) = \frac{\pi(i) \sum_{k \in \mathscr{S}_c(\iota)} q(i,k)}{\sum_{l \in \mathscr{S}'} \pi(i) \sum_{k \in \mathscr{S}_c(\iota)} q(i,k)}. \quad (73)$$

From equation (37) however

$$\lambda(c) = \sum_{k \in \mathscr{S}_c(\iota)} q(i,k) = \sum_{i \in \mathscr{S}} \pi(i) \sum_{k \in \mathscr{S}_c(\iota)} q(i,k) \quad (74)$$

and substituting this into (73) shows that $\psi_{Z_a}(i) = \pi(i)$ as claimed. A similar argument shows that departing class c customers also see the system in equilibrium. Analogous to (73) we write

$$\psi_{Z_d}(i) = \frac{\sum_{k \in \mathscr{S}_c(\iota)} \pi(k) q(k,i)}{\sum_{k \in \mathscr{S}_c(\iota)} \pi(k) \sum_{l \in \mathscr{S}} q(k,i)} = \frac{\pi(i) \sum_{k \in \mathscr{S}_c(\iota)} q^r(i,k)}{\sum_{l \in \mathscr{S}} \pi(i) \sum_{k \in \mathscr{S}_c(\iota)} q^r(i,k)} \quad (75)$$

which, using (38), shows that $\psi_{Z_d}(i) = \pi(i)$ from the partial balance arrival-departure property. Thus for quasireversible queues, arrivals and departures of class c queues see the system in equilibrium. We call this property, the *arrival-departure property of open quasireversible queues*. We will sometimes refer to this as the *arrival theorem* for open networks (see Lavenberg and Reiser [1980] and Sevcik and Mitrani [1981] for theorems of this type).

Application of the State Truncation and Arrival-Departure Properties

Suppose the system has a fixed population as in the closed-network application of state truncation property or the closed-network application of the distribution and the state truncation properties. The process Y corresponding to a quasireversible queue restricted to $\mathscr{Z}$ thus has a stationary distribution equal to

$$\pi_Y(i) = C\pi(i) \quad i \in \mathscr{Z}, \quad (76)$$

where $C = 1/\sum_{k \in \mathscr{Z}} \pi(k)$. What are the distributions seen by a class c arrival or departure from this system? When we talk about such a customer, we are assuming the customer is in transit between queues (i.e., not resident at any queue). Consider a departing class c customer and denote the distribution it sees at time of departure by $\psi_d(i)$. We first note that there must be a class c customer for one to depart and that there is

one less class c customer in the system after departure. Thus the states that can be seen by the departing customer are in set $\mathcal{Z}_c$. We can write the distribution as

$$\psi_d(i) \propto \sum_{k \in \mathcal{Y}_c'(i)} \pi(k) q(k,i), \quad (77)$$

$$\propto \pi(i) \sum_{k \in \mathcal{Y}_c(i)} q'(i,k) = \pi(i)\lambda_c, \quad i \in \mathcal{Z}_c, \quad (78)$$

where we have used (38) and (39). This has to be normalized over $\mathcal{Z}_c$, and thus the class c departure sees the system in equilibrium with one less class c customer. Considering the process in reverse time shows that this is also true for an arriving class c customer, and thus arrivals or departures of class c customers see the system in equilibrium with one less class c customer. We call this property the *arrival-departure property of closed quasireversible queues.*

Mean Value Analysis

Consider a closed quasireversible queuing network that consists of J different service centers, and suppose that there are n customers of a single class. Suppose that service center j is a single-server FCFS queue with exponential service times with expectation $1/\mu_j$ and that θ_j is the frequency with which a customer visits queue j relative to the frequency with which it visits queue 1. Let $R_j(n)$, $L_j(n)$, and $\Lambda_j(n)$ be the expected response time, expected queue length, and throughput, respectively, for service center j when the population of the network is n. Note that $\Lambda_j(n) = \theta_j \Lambda_1(n)$. Now consider an arrival to queue j. From the arrival-departure property of closed quasireversible queues it follows that the expected number of customers in queue j found by this arrival (while in transit and not in any queue) is equal to the expected number of customers in queue j when the population is equal to $n - 1$. Using this we can write the expected response time for the newly arrived customer as

$$R_j(n) = \frac{1}{\mu_j}\big(1 + L_j(n-1)\big). \quad (79)$$

Applying Little's [1961] result to the individual queues implies that

$$L_j(n) = \theta_j \Lambda_1(n) R_j(n), \quad (80)$$

and summing (80) over all queues yields

$$\Lambda_1(n) = \frac{n}{\sum_{j=1}^{J} \theta_j R_j(n)}. \quad (81)$$

Equations (80) and (81) are a special case of the celebrated *Mean Value Analysis* (MVA) equations [Reiser and Lavenberg 1980]. These equations, with the obvious boundary condition of $L_j(0) = 0$, can be used to recursively calculate the expected response time and queue length for increasing values of n without the difficulties of calculating a normalizing constant.

The State Aggregation Property

The last property of partial balance is related to state truncation. Suppose we partition $\mathcal{S}'$ into sets $\mathcal{U}^n$, $n = 0, 1, \ldots, N$, $N \geq 1$, and let X^n denote X truncated to $\mathcal{U}^n$. We assume that X^n is irreducible and let $\pi^n(i)$, $i \in \mathcal{U}^n$, denote the stationary distribution of X^n analyzed in isolation. Also assume that the sets satisfy nearest-neighbor transitions, i.e., $q(i,k) = 0$, $i \in \mathcal{U}^n$, $k \in \mathcal{U}^m$, if $|n - m| > 1$. We call each set an *aggregated state* and define a birth–death process with transition rates, $Q(n,m)$, given by

$$Q(n,m) \equiv \sum_{i \in \mathcal{U}^n} \sum_{k \in \mathcal{U}^m} \pi^n(i) q(i,k), \quad n, m \geq 0, |n - m| = 1. \quad (82)$$

Let $\Pi(n)$, $n \geq 0$, be the stationary distribution of this birth–death process. These values satisfy the following detailed balance equations

$$\Pi(n)Q(n,n+1) - \Pi(n+1)Q(n+1,n) = 0, \quad n \geq 0, \quad (83)$$

and thus can be easily solved using $Q(n, m)$.

What relationship does the distribution of this aggregated birth–death process have to the original process? We claim that if partial balance holds on sets $\mathscr{Y}^n$ then $\Pi(n) = \sum_{i \in \mathscr{Y}^n} \pi(i)$. To show this, observe that if partial balance is satisfied then the state truncation property implies that

$$\pi^n(i) = \frac{\pi(i)}{\sum_{i \in \mathscr{Y}^n} \pi(i)}, \quad i \in \mathscr{Y}^n, \quad (84)$$

and thus substituting this into (82) shows that (83) is satisfied with $\Pi(n) = \sum_{i \in \mathscr{Y}^n} \pi(i)$. Note that, if partial balance holds, $Q(n, m)$ is equal to the average transition rate from states in set $\mathscr{Y}^n$ to states in set $\mathscr{Y}^m$ in the original process. Summarizing this result, we say that if partial balance holds on sets $\mathscr{Y}^n$ then the distribution of the aggregated process is identical to what would be obtained in the original process by summing the stationary distribution of the aggregated states. We call this property the *state aggregation* property of processes that satisfy partial balance.

Viewing this in terms of an open quasireversible network with customer classes, note that partial balance among classes (42) implies that

$$\mathbf{F}(\mathscr{N}_c(n), \mathscr{N}_c(n+1)) - \mathbf{F}(\mathscr{N}_c(n+1), \mathscr{N}_c(n)) = 0, \quad 0 \le n. \quad (85)$$

Let states in the aggregated system correspond to the number of class c customers, and let $\Pi_c(n) \equiv \sum_{i \in \mathscr{I}_c(n)} \pi(i)$ be the probability that there are n class c customers in the original system. Since we assume the process is quasireversible, the arrival rate of class c customers, $\lambda(c)$, is independent of the state of the system. Let $\mu_c(n)$ be the average departure rate of class c customers conditioned on n class c customers being in the queue. This is given by

$$\mu_c(n) = \sum_{i \in \mathscr{I}_c(n)} \sum_{k \in \mathscr{I}_c(n-1)} \frac{\pi(i)}{\Pi_c(n)} q(i, k), \quad n > 0. \quad (86)$$

Thus the detailed balance equations satisfied by the aggregated process (analogous to (83)) are

$$\Pi_c(n)\lambda_c - \Pi_c(n+1)\mu_c(n+1) = 0, \quad n \ge 0. \quad (87)$$

Notice that (17) implies that we can write the solution to (87) as

$$\Pi_c(n) \propto \prod_{j=1}^{n} \frac{\lambda_c}{\mu_c(j)}, \quad (88)$$

and thus, as in the birth–death example with state-independent arrival rate, the arrival rate and departure rates of the aggregated process determine its stationary distribution. Thus, if some property of a given quasireversible queue depends only on the distribution of its aggregated process, then the stationary statistics for that property are identical to that of a system where we replace the given queue by a simple birth–death queue that has the appropriate state-dependent service rates. This result has been termed Norton's theorem [Chandy et al. 1975; Krzesinski and Teunissen 1985] and will be discussed below.

Application of the State Aggregation Property

The main application of the state aggregation property is to create a *flow-equivalent* server for a complex set of queues in a network. The flow-equivalent server is equivalent to the birth–death queue with state-dependent service rates mentioned above. For example, suppose we consider a model of a computer system consisting of a CPU subsystem and a disk subsystem. Assume that the disk subsystem consists of K M/M/1 queues, each corresponding to a disk with a queue of work, and assume that customers are routed with uniform probability to any one of the disks. Let customers at the disks be of class c; define $\Pi_c(n)$ to be the

probability that there are a total of n class c customers in the system (i.e., at the disk subsystem); and let $\mu_c(n)$ be defined as in (86). Suppose now that we wish to study the changes in performance of the entire system as a function of parameters of the CPU subsystem. Since there are no changes in the disk subsystem, the parametric study can be computationally facilitated by replacing the disk subsystem with a flow-equivalent server consisting of a single-server FCFS queuing system with a state-dependent service rate of $\mu_c(n)$.

We review in words the above properties.

- **Distribution Property**. In an open quasireversible network, the probability that an arriving class c customer causes a state transition to a given state is equal to the ratio of a transition rate to the arrival rate of class c customers.
- **State Truncation Property**. The stationary distribution for a system restricted to a subset of the states is a normalization of the unrestricted stationary distribution if partial balance holds on the subset. When open quasireversible queues are closed such that external transitions of the corresponding open system satisfy the distribution property, then the closed system is a truncated version of the open system and has a stationary distribution which is a normalization of the open system's distribution.
- **Arrival-Departure Property**. The distribution seen by transitions out of a set $\mathscr{A}$ is identical to that seen by transitions into set $\mathscr{A}$ if and only if partial balance on set $\mathscr{A}$ holds. If partial balance holds on set $\mathscr{A}$ and if the queue is quasireversible then for **open** systems, arriving (resp., departing) class c customers that enter (resp., leave) set $\mathscr{A}$ see the stationary distribution, and for **closed** systems, arriving (resp., departing) class c customers that enter (resp., leave) set $\mathscr{A}$ see the stationary distribution, calculated as if they were not in the system.
- **State Aggregation Property**. The stationary distribution of an aggregated system $\mathscr{A}^n$, for $n = 0, 1, \ldots, N$, with constant arrival and departure rates is the same as would be found by summing up stationary probabilities of the aggregated states in the original process if partial balance holds on sets $\mathscr{A}^n$.

We have already seen one queue that is quasireversible, the birth–death process with state-independent Poisson arrivals. Other queues that are useful in computer modeling and are quasireversible are the classical BCMP queues [Baskett et al. 1975] which allow general service time distributions and include the following scheduling policies: last-come first-serve preemptive resume, processor sharing, and infinite server. Symmetric queues [Kelly 1979] include these queuing disciplines as special cases. We refer the reader to Kelly [1979] and Walrand [1988] for the proofs that these queues are quasireversible. We just mention here a typical way to establish that a queue is quasireversible is to use the form of the reverse process. Assume that the forward process has Poisson arrivals that are independent of the state of the system. Often it is the case that the reverse process has a queuing structure that is a mirror image of the original system. If this is the case then the correspondence between arrivals (resp., departures) of the forward process and departures (resp., arrivals) of the reverse process implies that the departure process of the forward process is also independent of the state of the system and is also Poisson. This establishes that the queue is quasireversible, and typically the stationary distribution of the process can be guessed and checked using the reverse balance equations (21) and (22).

We close this section by reviewing the differences between reversible and quasireversible queuing systems in the context of a queuing system with C customer classes. Reversible systems satisfy detailed balance and have arrival and departure processes that are statistically

Table 1. Characterization of Reversible and Quasireversible Processes

Property	Quasireversible	Reversible
Arrival Rates	Exponential Interarrival Times	
Balance Equations	$\alpha(c)$ is Independent of State (Stronger Condition) Partial Balance (Weaker Condition)	$\alpha(c, i)$ can be State Dependent (Weaker Condition) Detailed Balance (Stronger Condition)

identical. The arrival and departure processes of class c customers, $\alpha(c, i), i \in \mathcal{S}$, are generally state dependent. A reversible system is quasireversible only if arrivals of customers to the queue are Poisson with state-independent rates. Quasireversible queues satisfy partial balance, a less restrictive condition than detailed balance, and always have Poisson arrival and departure processes. A queuing system can be reversible without being quasireversible (as in the birth–death queue with state-dependent arrival rates), and a quasireversible queue is not necessarily reversible since partial balance does not imply detailed balance. We summarize these statements in Table 1.

3. NETWORKS OF QUASIREVERSIBLE QUEUES

The previous section established four properties of queuing systems that satisfied partial balance and quasireversibility. Suppose that we join a set of quasireversible queues into a network so that the resultant system is also quasireversible. This network would then also satisfy these properties, namely, it would satisfy the distribution, state truncation, arrival-departure, and state aggregation properties. Clearly we cannot join queues in an arbitrary fashion and still preserve quasireversibility, but the *algebra* for how such queues can be joined is surprisingly flexible. In this section we first analyze two simple models of quasireversible networks to derive properties of their stationary distributions. This allows us to derive basic properties of such systems without being burdened with excessive notation. We then indicate how all of the results found in these simple models generalize to more complex models.

3.1 Tandem Queues

Suppose we consider an open network consisting of two quasireversible queues in tandem. Suppose arrivals to the first queue are Poisson with rate λ. The state of the system is (x_1, x_2) where x_i, $i = 1, 2$, is the state of queue i. What are the stationary state probabilities, $\pi(x_1, x_2)$?

To answer this, we note that because the first queue is quasireversible, $x_1(t)$ is independent of the departure process from queue 1 prior to time t. Departures from queue 1, however, form the arrivals to queue 2 and thus determine the value of $x_2(t)$. Thus $x_1(t)$ and $x_2(t)$ are independent and act as if it were in isolation. The stationary probabilities satisfy a product form, $\pi(x_1, x_2) = \pi_1(x_1)\pi_2(x_2)$, where $\pi_i(\ )$ is the stationary distribution for queue i analyzed in isolation. The arrival-departure property of open quasireversible queues implies that an arriving customer to the second queue sees the same distribution of (x_1, x_2) that is seen by a departing customer from the second queue. Both of these distributions are equal to the stationary distribution of the process.

Recall that we specified no scheduling or service policy for the birth–death model with state-dependent servicing rates. The existence of product form depends on external properties (the input-output properties) of queues rather than internal properties (scheduling disciplines for example) [Kelly 1979; Muntz 1972]. The fact that is is quasireversible does not require any notion of the operation of the queue. To show that a system

is quasireversible we must first define a system state and then specify which transitions correspond to class c arrivals and departures and then demonstrate that the conditions for quasireversibility are satisfied. This procedure permits great flexibility and creativity in defining such processes [Kelly 1979]. It is sometimes misleading, however, to speak of such systems as "queues" since very little queuing in the traditional sense takes place in many quasireversible systems. For example, the arrival and departure processes, from the tandem queuing system are Poisson streams that are independent of the state of the system. Thus, considering the tandem as one unit shows that it is also a quasireversible queue. This would hardly qualify as a "queue" in the normal sense of the word.

Within the above arguments are the seeds for an algebra of quasireversibility in which quasireversible queues can be joined in a manner that preserves quasireversibility. It is easy to see, in the above system, that joining queues in series can be performed any number of times and still lead to a quasireversible queue. The stationary distribution after such operations is a product of terms where each term is the stationary distribution of the individual queues analyzed as if they were in isolation. It is intriguing to question the generality under which such properties hold. Before we address this issue, we first consider the tandem model under slightly different assumptions.

Assume now that the network is closed and thus that there are a fixed number of jobs in the queuing system, $N \geq 1$, so that once a job finishes executing at the second queue, it immediately cycles back as an arrival to the first queue. Let $n_i(x_i)$, $i = 1, 2$ be the number of customers in queue i when that queue is in state x_i. Observe that several states of a queue could correspond to having the same number of customers. Clearly, now the states of the queues are dependent since $n_2(x_2) = N - n_1(x_1)$. We are compelled here to perpetuate the existing nomenclature for such a system and call it a *closed network of quasireversible queues.* Clearly, the notion of Poisson arrival and departure processes does not exist in a closed network, and also the arrival and departure processes from a queue of this system are not independent of its state. Each queue of the network thus violates the properties of quasireversibility. What we mean by calling the system a closed network of quasireversible queues is that the network consists of queues that would be quasireversible *if* each queue were considered in isolation with Poisson input processes.

Here we view the system as a network of quasireversible queues that is restricted to have only N customers. The state truncation property combined with the solution for the open system immediately yields a product form solution, $\pi(x_1, x_2) = C\pi_1(x_1)\pi_1(x_2)$, where C is a normalization constant calculated over all (x_1, x_2) so that $n_1(x_1) + n_2(x_2) = N$. Without knowing the state truncation property, it would be surprising to have such a solution. Let us take this moment to clear up a possible confusion regarding the solution. We know that the queues are mutually dependent, and yet the solution is of product form which might seem to imply independence. There is no contradiction, however, because the normalization constant implies that the factors of the product do not correspond to the distributions of the individual queues. In other words, there is no way to separate the normalization term into a product of factors, say $C = C_1C_2$, so that $C_i\pi_1(x_i)$, $i = 1, 2$ is the stationary distribution for queue i.

Continuing the discussion of the closed system above, we now revisit the state aggregation property of partial balance. Suppose we consider queue 2 in isolation and aggregate its states according to its number of customers. State n here thus corresponds to the aggregation of all states x_2 satisfying $n_2(x_2) = n$. If we solve for the stationary distribution of the aggregated states then (88) shows that

$$\Pi_2(n) \propto \prod_{j=0}^{n} \frac{\lambda}{\mu_2(j)}, \qquad (89)$$

where $\Pi_2(n)$ is the stationary probability of aggregated state n and where $\mu_2(n)$ is the average departure rate of customers calculated for aggregated state n. Suppose now that one replaces queue 2 with a queue that processes customers at rate $\mu_2(n)$ when it contains n customers. This is the flow-equivalent server previously mentioned and corresponds to replacing the original process by a birth–death process with queue-dependent service rates. Clearly, this has the same stationary distribution as (89), and thus, as stated in Section 2, any function that depends on queue 2 only through its aggregation will not change with this replacement of a flow-equivalent server. One such function is the probability of observing queue 1 in state x_1, denoted by $\pi_1^e(x_1)$. Let $\mathscr{X}_2(x_1) = \{x_2 | n_2(x_2) = N - n_1(x_1)\}$ be the set of states of queue 2 that contain the customers not found in queue 1 when in state x_1. Then we can write the probability as

$$\pi_1^e(x_1) \propto \sum_{x_2 \in \mathscr{X}_2(x_1)} \pi_1(x_1)\pi_2(x_2) \quad (90)$$

$$\propto \pi_1(x_1)\Pi_2(N - n_1(x_1)) \quad (91)$$

$$\propto \pi_1(x_1) \prod_{j=0}^{N-n_1(x_1)} \frac{\lambda}{\mu_2(j)}. \quad (92)$$

Thus $\pi_1^e(x_1)$ depends only on the flow-equivalent rates $\mu_2(n)$ of queue 2. This is an application of *Norton's theorem* mentioned in Section 2.

3.2 Feedback Queues

We next consider a simple modification of the routing scheme given above. Suppose we have a single queue with two classes of customers. Customers of both classes are assumed to be served in the order in which they enter the tail of the queue. Arrivals to the queue are of class 1 and are Poisson with rate λ. After receiving service, a class 1 customer returns to the tail of the queue as a class 2 customer. Class 2 customers leave the system after receiving service, and we assume that the service times for both classes of customers are exponential with rate μ. The state of the system is the sequence of classes in the queue, $\underline{c} = (c_1, c_2, \ldots, c_n)$, where c_i is the class of the customer in position i in the queue, and n is the number of customers in the queue. We denote the state corresponding to an empty system by 0 and will let $|\underline{c}|$ be the total number of customers found in state $\underline{c}$. The transition rates for this system are given by

$$q(\underline{c}, \underline{c}') = \begin{cases} \lambda, \underline{c}' = (\underline{c}, 1), & \text{(Arrival)} \\ \mu, \underline{c} = (2, \underline{c}'), & \text{(Departure)} \\ \mu, \underline{c} = (1, \hat{\underline{c}}), \underline{c}' = (\hat{\underline{c}}, 2), & \\ \quad \text{(Feed Back)}. & \end{cases} \quad (93)$$

We claim that this process is quasireversible. To see this we first guess the form of the stationary distribution and the reverse transition rates and then show that (21) and (22) are satisfied.

What would the reverse process look like? The obvious guess is that customers arrive at the system at rate λ as class 2 customers, are fed back, and then leave the system as class 1 customers. This implies transition rates given by

$$q^r(\underline{c}', \underline{c}) = \begin{cases} \mu, \underline{c}' = (\underline{c}, 1), & \text{(Departure)} \\ \lambda, \underline{c} = (2, \underline{c}'), & \text{(Arrival)} \\ \mu, \underline{c} = (1, \hat{\underline{c}}), \underline{c}' = (\hat{\underline{c}}, 2), & \\ \quad \text{(Feed Back)}. & \end{cases} \quad (94)$$

Thus we have interchanged arrivals and departures for the reverse transition rates.

Since both classes of customers arrive at an average rate of λ and have the same service time distributions, it is plausible that states with the same number of total customers have the same probability, i.e., that $\pi(\underline{c}) = \pi(\underline{c}')$ if $|\underline{c}| = |\underline{c}'|$. We thus aggregate states according to their number of customers. Let $\mathscr{X}^n = \{\underline{c} | n = |\underline{c}|\}$ be the set of states with n customers, and let $\Pi(n)$ be the probability of aggregated state n. Since each state $\underline{c} \in \mathscr{X}^n$ is assumed to have the same

probability, it follows that

$$\pi(\underline{c}) = \left(\frac{1}{2}\right)^{|\underline{c}|} \Pi(|\underline{c}|). \qquad (95)$$

Suppose that we then make a guess that the distribution of the aggregated states are identical to that of an M/M/1 system with an arrival rate of 2λ. Thus we guess that

$$\Pi(n) = (1 - \rho)\rho^n, \quad n \geq 0, \quad (96)$$

where $\rho \equiv 2\lambda/\mu$.

It is clear that the transition rates (93) and (94) satisfy (22), and thus we only have to show that the reverse balance equations are satisfied (21) to show that the values (95) are correct. For forward arrivals (reverse departures) this implies that

$$(1 - \rho)\left(\frac{\rho}{2}\right)^{|\underline{c}|} \lambda = (1 - \rho)\left(\frac{\rho}{2}\right)^{|\underline{c}|+1} \mu \quad (97)$$

which is clearly true. It similarly follows that (21) is satisfied for forward departures (reverse arrivals) and for fed back customers, and thus (95) is correct. The nature of the reverse process shows that the departure process is Poisson and similarly that the queue is quasireversible as claimed.

We make an important observation about this system. In the open tandem system considered in Section 3.1, the input process to each queue was Poisson, and thus it was not surprising that the queues were quasireversible. Although the external arrival process to feedback queue is Poisson its input process *is not* Poisson. We will now provide a proof of this by creating a mapping between our feedback model, henceforth called *the fixed-feedback model* and that of a model of *Bernoulli feedback*. In Bernoulli feedback, each customer is fed back with probability $p, 0 \leq p < 1$, after receiving service. Burke [1976] considered such a model and showed that the stationary distribution of states as seen by any arriving customer (external or fed back) is the same as that of an M/M/1 queue with an arrival rate given by $\lambda/(1 - p)$ and also that the total arrival stream of customers to the queue is not Poisson. We show here that the total arrival stream of customers in the feedback model is statistically identical to that of a Bernoulli feedback system with $p = 1/2$, and thus it follows from Burke's result that the total arrival stream is not Poisson as claimed. Equation (96) shows that the stationary distributions for the number of customers in the system are identical for the fixed feedback and Bernoulli feedback models when $p = 1/2$.

Consider now the fixed-feedback model. Since the queue is quasireversible; the arrival-departure property of quasireversible queues shows that the probability distribution seen by any arriving customer (either external or fed back) is equal to the stationary distribution. Suppose now that we randomly select an arriving customer, customer J, and assume that at the time of its arrival it sees $n, n \geq 0$, customers in the system. Note that J joins the queue in position $n + 1$. Since every customer that arrives externally will be fed back exactly once, the probability that J is of class 1 (external arrival) or class 2 (fed back) is identical and equal to 1/2. Thus, the probability that the customer in position $n + 1$ (i.e., customer J) is of class 2 is equal to 1/2. We claim that this is also the case for the customer in position $j, 1 \leq j \leq n$, of the queue. This follows the fact that all states in $\mathscr{C}^n$ are equally likely with probability given by (95). Thus, the probability that any customer in the queue after J's arrival is of class 2 is given by 1/2 so that the system is equivalent to a Bernoulli feedback system with $p = 1/2$ as claimed. Thus, the total arrival stream of customers to the queue is not Poisson. This result can be established more generally (see Walrand [1983]). It is *still* the case, however, that the external arrival and departure processes are Poisson. The non-Poisson flow within the queue do not violate the notion of quasireversibility.

3.3 Product Form or Quasireversible Queuing Networks

The algebra hinted at in the previous examples suggests that quasireversible queues can be joined together so as to form new quasireversible queues. More general routing mechanisms also preserve quasireversibility. For example, in Markov routing [Walrand 1988] class c customers at queue i depart and become class c' customers at queue i' with probability $r(c, i; c', i')$. This routing is said to be Markovian because routing decisions for a customer depend only on its current class and queue and are independent of the rest of the state of the network. More general routing decisions that still preserve quasireversibility and allow certain types of state dependencies can be found in Kelly [1979], Krzesinski [1987], and Towsley [1980]. Any network created using such routing policies leads to a network of queues that is quasireversible and thus will have product form solutions and will satisfy the four properties derived in Section 2. The proof of these claims is obscured by the notations needed to define the general process and by the algebra needed to establish that global balance is satisfied. In keeping with the tenor of this article, we content ourselves with investigating a simplified case that illuminates the procedures used to prove the general case.

We consider an open network consisting of M quasireversible queues and C customer classes. Suppose that we know the stationary distribution $\pi_m(x)$ and the forward and reverse transition rates, $q_m(\ )$ and $q_m^r(\ )$, respectively, of queue m when analyzed in isolation. In the network, external arrivals of class c customers are Poisson with rate $\gamma(c)$. Class c customers start at queue $r_c(1)$ and sequentially visit queues $r_c(j)$, $1 \le j \le l_c$, $l_c \ge 1$, and then leave the system. There are three different types of transitions: external arrivals and external departures, each of which causes a state change at only one queue, and internal transitions, where a departure from one queue corresponds to an arrival to another queue. Internal transitions change state values of two queues in the network. The state of queue m is a vector containing the number of class c customers resident in the queue, and the state of the network is a concatenation of all the states of all the queues. For a given transition that causes queue m to change state, we denote its state before the transition by x_m and its state after the transition by y_m.

To analyze the process we must specify transition rates for each of the above transitions. To do this we expand our previous notations to include a queue index. Let $\mathcal{S}_c^m(x)$ be the set of states for queue m that have one more class c customer than state x with the same number of customers of other classes. Transition rates for queue m will be denoted by $q_m(\ )$, and the arrival rate of class c customers to the queue is given by $\lambda_m(c)$. It is clear that $\lambda_m(c) = j\gamma(c)$, $j \ge 0$, if class c jobs visit queue m exactly j times.

We now specify the rates for each of the above types of transitions. External arrivals of class c customers to queue $m = r_c(1)$ occur at rate $q_m(x_m, y_m)$ where $y_m \in \mathcal{S}_c^m(x_m)$. Similarly, external departures occur at rate $q_m(x_m, y_m)$ where $x_m \in \mathcal{S}_c^m(y_m)$, $m = r_c(l_c)$. Internal transitions of a class c customer from queue $m = r_c(j)$, $1 < j < l_c - 1$, to queue $m' = r_c(j+1)$ are *assumed* to occur at rate

$$q_m(x_m, y_m)\frac{q_{m'}(x_{m'}, y_{m'})}{\lambda_{m'}(c)}, \quad x_m \in \mathcal{S}_c^m(y_m), \quad y_{m'} \in \mathcal{S}_c^{m'}(x_{m'}), \quad (98)$$

and thus satisfy the distribution property of quasireversible queues (51). It is important to note that the rates (98) are defined for the process in such a way that they satisfy the distribution property. If we arbitrarily join quasireversible queues together in a network, it is not true that they necessarily satisfy this property.

We are now in the position to make good our previous promise to "guess" the correct solution for networks of quasire-

versible queues. We guess the only tractable solution, that the distribution of the network is the product of the distributions of the individual queues analyzed in isolation,

$$\pi(x_1, x_2, \ldots, x_M) = \pi_1(x_1)\pi_2(x_2)\ldots\pi_M(x_M). \quad (99)$$

We next guess the transition rates for the reverse process. It seems natural to suppose that, in the reverse process, class c customers backtrack on the forward route, i.e., they enter the system at queue $r_c(l_c)$ according to a Poisson process with rate $\gamma(c)$, sequentially visit queues $r_c(l_c - 1), r_c(l_c - 2), \ldots, r_c(1)$, and then exit the system. The reverse transition rates are then just the "reverse" of that given above, i.e., $q^r_m(y_m, x_m)$, $x_m \in \mathcal{S}_c(y_m)$, $m = r_c(l_c)$, for reverse external arrivals, $q^r_m(y_m, x_m)$, $y_m \in \mathcal{S}_c(x_m)$, $m = r_c(1)$, for reverse external departures, and for $m' = r_c(j+1), m = r_c(j), 1 < j < l_c - 1$,

$$q^r_{m'}(y_{m'}, x_{m'})\frac{q^r_m(y_m, x_m)}{\lambda_m(c)}, \quad x_m \in \mathcal{S}^m_c(y_m), \quad y_{m'} \in \mathcal{S}^{m'}_c(x_{m'}), \quad (100)$$

for reverse internal transitions.

To show that our guess is correct, we must show that the reverse balance equations, (21) and (22), are satisfied, and this is a matter of algebra. For external arrival (reverse departures) transitions this implies checking that

$$\pi_m(x_m)q_m(x_m, y_m) = \pi_m(y_m)q^r_m(y_m, x_m), \quad y_m \in \mathcal{S}^m_c(x_m), \quad (101)$$

which is clearly satisfied since the stationary distribution and reverse transition rates for queue m in isolation are given. External departures (reverse arrivals) are just as easily checked. For internal transitions we must show that

$$\pi_m(x_m)\pi_{m'}(x_{m'})q_m(x_m, y_m)\frac{q_{m'}(x_{m'}, y_{m'})}{\lambda_{m'}(c)} = \pi_m(y_m)\pi_{m'}(y_{m'}) \times q^r_{m'}(y_{m'}, x_{m'})\frac{q^r_m(y_m, x_m)}{\lambda_m(c)}, \quad x_m \in \mathcal{S}^m_c(y_m), \quad y_{m'} \in \mathcal{S}^{m'}_c(x_{m'}). \quad (102)$$

But this follows from the fact that (21) is satisfied for each individual queue. We have only (22) to check. The total transition rate from a network state for the forward process is given by

$$\sum_{m=1}^{M} q(x_m) + \sum_{c=1}^{C} \gamma(c). \quad (103)$$

Equation (22) is thus satisfied since $q^r_m(x_m) = q_m(x_m)$.

The above derivation embodies the methodology that is used to establish that product form holds for more general networks. Specifically, we set up routing between queues so that the distribution property holds. The "obvious" reversed routes are then guessed as is a product form solution for the stationary distribution. We then use the reverse balance equations, (21) and (22), to show that the guess is correct. The state truncation property is invoked if the network is closed. The resultant network is itself quasireversible and thus possesses the properties derived in Section 2. Thus, closing the system leads to a product form solution: the arrival theorem holds. For some systems a generalized form of Mean Value Analysis can be developed, and Norton's Theorem can be applied to create flow-equivalent servers.

Although, externally, an open quasireversible network has Poisson arrival and departure processes, the internal flow of customers within the network is not necessarily Poisson [Kelly 1979; Walrand 1983]. This does not violate the property of quasireversible since, as previously mentioned, such a notion requires no specification of the internal workings of the queue.

4. CONCLUSIONS

The fact that networks of quasireversible queues have product form solutions fol-

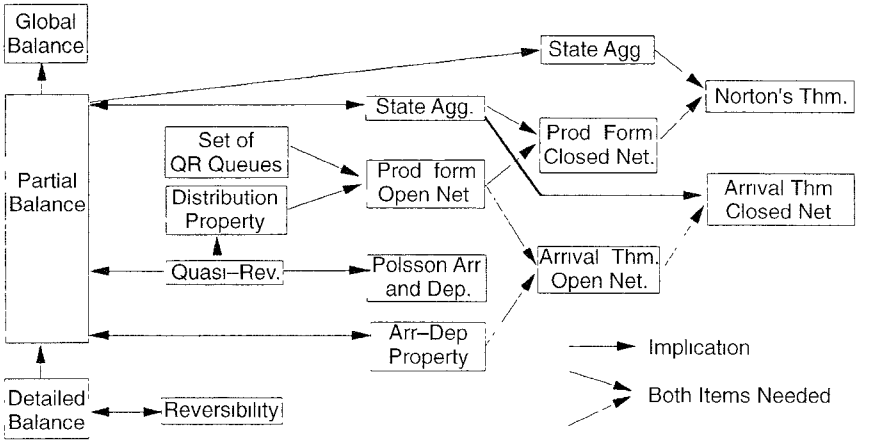

Figure 3. Summary of implications.

lows from fundamental properties of partial balance in general and quasireversibility in particular. Properties such as the arrival-departure property are preserved when queues are joined into networks in a manner that preserves quasireversibility, and we have seen that the algebra for joining such queues is flexible. The results derived in this article can be used as a starting point for studying more general forms of product form networks. Properties derived from the four properties established in Section 2 are summarized in Figure 3. The implications in this figure should be interpreted within the context of this article, and the assumptions used in their derivation can be found in the body of the article. There are many properties and features of product form networks that lie outside the scope of this article. For example, it can be shown that if partial balance is satisfied for set $\mathscr{C}$ then the equilibrium distribution of the system depends only on the mean of the distribution of time that is spent in states $\mathscr{C}$ [Whittle 1985; 1986b]. This property is called *insensitivity*, since the stationary distribution is insensitive to higher moments of sojourn time in $\mathscr{C}$.

Our focus was to develop the mathematics of product form from first principles. In so doing we have had to bypass a rich body of work devoted to using these results to model computer systems. As many applications do not satisfy the assumptions needed for product form, numerical or approximate solution techniques are required, and often these methods are based on intuition gained from exact solutions. Several sources [Lavenberg 1983; Lazowska et al. 1984; Sauer and Chandy 1981] provide a good starting point for investigating work along these lines.

ACKNOWLEDGMENTS

The author would like to thank S. S. Lavenberg for his cogent comments made after several careful readings of the manuscripts Special thanks go to C. S. Chang for numerous insightful and penetrating discussions. R. Muntz asked many probing questions that resulted in a more thorough treatment of partial and local balance and greatly enhanced the article's presentation. The author also thanks the reviewers for their criticisms of the article which led to an improved version

REFERENCES

BARD, Y. 1978a. The VM/370 performance predictor. *ACM Comput. Surv. 10*, 333–342.

BARD, Y 1978b. An analytic model of the VM/370. *IBM J. Res. Dev 22*, 498–508.

BARD, Y. 1980. A model of shared DASD and multipathing. *Commun. ACM 23*, 564–572.

BARBOUR, A D. 1976. Networks of queues and the methods of stages *Adv. Appl Prob. 8*, 584–591.

BASKETT, F., CHANDY, M., MUNTZ, R., AND PALACIOS, J. 1975. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM 22*, 248–260.

BOYSE, J. W., AND WARN, D. R. 1975. A straightforward model for computer performance prediction *ACM Comput. Surv. 7*.

BRANDWAJN, A., AND MCCORMACK, W. M. 1984. Efficient approximation for models of multiprogramming with shared domains. *Perf. Eval. Rev. 12*, 186–194

BROWN, R. M., BROWNE, J. C., AND CHANDY, K. M. 1977. Memory management and response time. *Commun. ACM 20*, 153–165.

BRYANT, R. M., KRZESINSKI, A. E., LAKSHMI, M. S., AND CHANDY, K. M. 1984. The MVA priority approximation *ACM Trans. Comput. Syst. 2*, 335–359.

BURKE, P. J. 1956. The output of a queueing system. *Oper. Res , 4*, 699–704.

BURKE, P. J. 1972. The output process of a stationary M/M/s queueing system. *An. Math. Stat. 39*, 1144–1152.

BURKE, P. J. 1976. Proof of a conjecture on the interarrival time distribution in an M/M/1 queue with feedback. *IEEE Trans. Commun. COM-24*, 575–576.

BUZEN, J. P. 1973. Computational algorithms for closed queueing networks with exponential servers *Commun. ACM, 16*, 527–531.

BUZEN, J. P. 1978. A queueing network model of MVS. *Comput. Surv. 10*, 319–331.

CHANDY, K. M., AND MARTIN, A. J. 1983. A characterization of product form queueing networks. *J. ACM 30*, 286–299.

CHANDY, K. M., AND NEUSE, D. 1982. Linearizer: A heuristic algorithm for queueing network models of computing systems. *Commun. ACM 25*, 126–133.

CHANDY, K. M., AND SAUER, C. H. 1980. Computational algorithms for product form queueing networks. *Commun. ACM 23*, 573–583.

CHANDY, K M., AND SAUER, C H. 1978. Approximate methods for analysis of queueing network models of computer systems. *Comput. Surv. 10*, 263–280.

CHANDY, K. M., HERZOG, U., AND WOO, L. S. 1975. Parametric analysis of queueing networks. *IBM J. Res. Devel. 19*, 43–49.

CHANDY, K. M., HOWARD, J. H., AND TOWSLEY, D. F. 1977. Product form and local balance in queueing networks. *J. ACM 24*, 250–263.

CONWAY, A. E , AND GEORGANAS, N D. 1986. RECAL—A new efficient algorithm for the exact analysis of multiple chain queuing networks. *J. ACM 33*, 768–791.

CONWAY, A. E., DE SOUZA E SILVA, E., AND LAVENBERG, S. S. 1989. Mean value analysis by chain of product form queueing networks. *IEEE Trans. Comput. 38*, 432–442.

DALEY, D. J. 1976. Queueing output processes. *Adv. Appl. Prob. 8*, 395–415.

DE SOUZA E SILVA, E., AND LAVENBERG, S. S. 1989. Calculating the joint queue-length distribution in product-form queueing networks. *J. ACM. 36*, 194–207.

DE SOUZA E SILVA, E., LAVENBERG, S. S., AND MUNTZ, R. R. 1986. A clustering approximation technique for queueing network models with a large number of chains. *IEEE Trans. Comput. C-35*, 419–430.

DISNEY, R. L 1975 Random flows in queueing networks: A review and a critique. *Trans. AIEE 7*, 268–288

DISNEY, R. L., AND KIESSLER, P. C. 1987. *Traffic Processes in Queueing Networks: A Markov Renewal Approach*. The John Hopkins Press, Baltimore, Md

DISNEY, R. L., AND KÖNIG, D 1985. Queueing networks: A survey of their random processes. *SIAM Rev. 27*, 335–403.

EAGER, D. L., AND LIPSCOMB, J. N. 1988. The AMVA priority approximation. *Perf. Eval. 8*, 173–193.

GELENBE, E., AND MUNTZ, R. R. 1976. Probabilistic models of computer systems—Part 1 (exact results). *Acta Informatica 7*, 35–60.

GORDON, W. J., AND NEWELL, G. F. 1967. Closed queueing systems with exponential servers. *Oper. Res. 15*, 254–265.

HARRISON, J M., AND WILLIAMS, R. J. 1990 On the quasireversibility of a multiclass Brownian service station *An. Prob. 18*, 1249–1268.

HEIDELBERGER, P. AND TRIVEDI, K. S. 1982. Queueing network models for parallel processing with asynchronous tasks. *IEEE Trans. Comput. C-31*, 1099–1108.

HEIDELBERGER, P., AND TRIVEDI, K. S. 1983. Analytic queueing models for programs with internal concurrency. *IEEE Trans. Comput. C-32*, 73–82.

HENDERSON, W., AND TAYLOR, P. 1989. Alternative routing networks and interruptions. *ITC-12*, 1352–1358.

HORDIJK, A., AND VAN DIJK, N. M. 1983a. Networks of queues. Part 1: Job-local-balance and the adjoint process. Part II: General routing and service characteristics. In *Lecture Notes in Control and Informational Sciences*, Vol. 60. Springer-Verlag, New York, 151–205.

HORDIJK, A., AND VAN DIJK, N. M. 1983b. Adjoint processes, job-local-balance and insensitivity of stochastic networks. *Bulletin of the 44th Session of the International Statistics Institute*, vol. 50. 776–788

HORDIJK, A., AND VAN DIJK, N. M. 1981. Networks of queues with blocking. In *Performance 81*. 51–65.

HOYME, K. P., BRUELL, S. C., AFSHARI, P. V., AND KAIN, R. Y. 1986. A tree-structured mean

value analysis algorithm. *ACM Trans. Comput. Syst. 4*, 178–185.

JACKSON, J. R. 1963. Jobshop-like queueing systems. *Manage. Sci. 10*, 131–142.

JANSEN, U., AND KÖNIG, D. 1980. Insensitivity and steady state probabilities in product form queueing networks. *Elektron. Informationsverarb. Kybernet 16*, 385–397. In German.

KELLY, F. P. 1983. The dependence of sojourn times in closed queueing networks. In the *International Workshop on Applied Mathematics and Perf. Rel. Models*. 57–65.

KELLY, F. P. 1982. Networks of quasireversible nodes. In *The Interface*, vol. 1. *Progress in Computer Science* 2. Birkhauser, Boston, 3–29.

KELLY, F. P. 1979. *Reversibility and Stochastic Networks*. Wiley, New York.

KELLY, F. P. 1976a. Networks of queues. *Adv. Appl. Prob. 8*, 416–432.

KELLY, F. P. 1976b. The departure process from a queueing system. *Math. Proc. Cambr. Phil. Soc. 80*, 283–285.

KELLY, F. P. 1975. Networks of queues with customers of different types. *J. Appl. Prob. 12*, 542–554.

KELLY, F. P., AND POLLETT, P. K. 1983. Sojourn times in closed queueing networks. *Adv. Appl. Prob. 15*, 638–656.

KINGMAN, J. F. C. 1969. Markov population processes. *J. Appl. Prob. 6*, 1–18.

KOENIGSBERG, E. 1958. Cyclic queues. *Oper. Res. Q. 9*, 22–35.

KOLMOGOROV, A. 1936. Zur theorie der Markoffschen ketten. *Mathematische Annalen 112*, 155–160.

KRZESINSKI, A. 1987. Multiclass queueing networks with state-dependent routing. *Perf. Eval. 7*, 125–144.

KRZESINSKI, A., AND GREYLING, J. 1984. Improved lineariser methods for queueing networks with queue dependent centres. In *Proceedings of the 1984 ACM Sigmetrics Conference*. ACM, New York, 41–51.

KRZESINSKI, A., AND TEUNISSEN, P. 1985. Multiclass queueing networks with population constrained subnetworks. In *Proceedings of the 1985 ACM Sigmetrics Conference*. ACM, New York, 128–139.

LAM, S. S. 1977. Queueing networks with population size constraints. *IBM J. Res. Devel. 21*, 370–378.

LAM, S. S., AND LIEN, Y. L. 1983. A tree convolution algorithm for the solution of queueing networks. *Commun. ACM 26*, 203–215.

LAVENBERG, S. S. 1983. *Computer Performance Modeling Handbook*. Academic Press, New York.

LAVENBERG, S. S., AND REISER, M. 1980. Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *J. Appl. Prob. 17*, 1048–1061.

LAZOWSKA, E., AND ZAHORJAN, J. 1982. Multiple class memory constrained queueing networks. In *Proceedings of the 1982 ACM Sigmetrics Conference*. ACM, New York, 130–140.

LAZOWSKA, E. D., ZAHORJAN, J., GRAHAM, G. S., AND SEVCIK, K. C. 1984. *Quantitative System Performance, Computer System Analysis using Queueing Network Models*. Prentice-Hall, Englewood Cliffs, N. J.

LE BOUDEC, J. Y. 1985. A BCMP extension to multiserver stations with concurrent classes of customers. In *Proceedings of the 1986 ACM Sigmetrics Conference*. ACM, New York, 128–139.

LEMOINE, A. J. 1977. Networks of queues–A survey of equilibrium analysis. *Manage. Sci. 24*, 464–481.

LITTLE, J. D. C. 1961. A proof of the queueing formula $L = \lambda W$. *Oper. Res. 9*, 383–387.

MCKENNA, J., AND MITRA, D. 1982. Integral representations and asymptotic expansions for closed Markovian queueing networks: Normal usage. *Bell Syst. Tech. J. 61*, 661–683.

MELAMED, B. 1982. On Markov jump processes embedded at jump epochs and their queueing theoretic applications. *Math. OR 7*, 111–128.

MUNTZ, R. 1972. Poisson departure processes and queueing networks. IBM Res. Rep. RC 4145. IBM, Yorktown Heights, N.Y.

MUNTZ, R. 1973. Poisson departure processes and queueing networks. In *Proceedings of the 7th Annual Conference on Information Science and Systems*. Princeton Univ., Princeton, N.J., 435–440.

NELSON, R., AND KLEINROCK, L. 1985. Rude-CSMA: A multihop channel access protocol. *IEEE Trans. Commun. 33*, 785–791.

PITTEL, B. 1979. Closed exponential networks of queues with saturation: The Jackson type stationary distribution and its asymptotic analysis. *Math. Oper. Res. 4*, 367–378.

REICH, E. 1957. Waiting times when queues are in tandem. *An. Math. Stat. 28*, 768–773.

REISER, M. 1979. A queueing network analysis of computer communications networks with window flow control. *IEEE Trans. Commun. C-27*, 1199–1209.

REISER, M., AND KOBAYASHI, H. 1975. Queuing networks with multiple closed chains: Theory and computational algorithms. *IBM J. Res. Devel. 19*, 283–294.

REISER, M., AND LAVENBERG, S. S. 1980. Mean value analysis of closed multichain queueing networks. *J. ACM 27*, 313–322.

ROSS, S. *Stochastic Processes*. Wiley, New York.

SAUER, C. H. 1981. Approximate solution of queueing networks with simultaneous resource possession. *IBM J. Res. Devel. 25*, 894–903.

SAUER, C., AND CHANDY, K. M. 1981. Computer systems performance modeling. Prentice-Hall, Englewood Cliffs, N.J.

SCHWEITZER, P. 1979. Approximate analysis of multiclass closed networks of queues. In *Proceedings of the International Conference on Stochastic Control and Optimization*.

SERFOZO, R. 1989. Markovian network processes: Congestion dependent routing and processing. *Queueing Syst. 5*, 5–36.

SEVCIK, K. C., AND MITRANI, I. 1981. The distribution of queueing network states at input and output instants *J. ACM 28*, 358–371

THOMASIAN, A., AND BAY, P. F. 1986. Analytic queueing networks for parallel processing of task systems. *IEEE Trans Comput. C-35*, 1045–1054.

THOMASIAN, A., AND BAY, P. F. 1984. Analysis of queueing network models with population size constraints and delayed blocked customers *Proceedings of the 1984 ACM Sigmetrics Conference*. ACM, New York, 202-216.

TOWSLEY, D 1986. Approximate models of multiple bus multiprocessor systems. *IEEE Trans. Comput C-35*, 220–228.

TOWSLEY, D 1983. An approximate analysis of multiprocessor systems. In *Proceedings of the 1983 ACM Sigmetrics Conference*. ACM, New York, 207–213.

TOWSLEY, D. 1980. Queueing network models with state-dependent routing *J. ACM 27*, 323–337.

VAN DIJK, N 1991 Product forms for random access schemes. *Comput Networks ISDN Syst. 22*, 303–317.

VAN DIJK, N. 1990a. Product forms for queueing networks with limited clusters. Series Research Memoranda, Vrije Universiteit, Amsterdam.

VAN DIJK, N. 1990b. An equivalence of communications protocols for interconnection networks. *Comput. Networks ISDN Syst. 20*, 277–283.

WALRAND, J. 1988 *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, N J.

WALRAND, J. 1983 A probabilistic look at networks of quasireversible queues. *IEEE Trans. Inf. Theor. IT-29*, 825–831.

WHITTLE, P 1986a. *Systems in Stochastic Equilibrium*. Wiley, New York.

WHITTLE, P. 1986b. Partial balance, insensitivity and weak coupling *Adv. Appl. Prob 18*, 706–723.

WHITTLE, P. 1985. Partial balance and insensitivity. *J. Appl. Prob. 22*, 168–176.

WHITTLE, P. 1968. Equilibrium distributions for an open migration process *J. App. Prob. 5*, 567–571.

WHITTLE, P. 1967. Nonlinear migration processes. *Bull. Int. Inst. Stat. 42*, 642–647.

WONG, J. W. 1978. Queueing network modeling of computer communications networks. *ACM Comput. Surv. 10*, 343–352

ZAHORJAN, J., EAGER, D. L., AND SWEILLAM, H. 1988. Accuracy, speed and convergence of approximate mean value analysis. *Perf. Eval. 8*, 255–270.

Received July 1991, final revision accepted February 1993.