Preface

These notes include the material from a series of nine lectures given at the Saint-Flour Probability Summer School, July 2 - July 15, 2006. Lectures were also given by Alice Guionnet and Steffen Lauritzen. It was my first visit to Saint-Flour, which I enjoyed considerably.

The topic of these notes, the stability of queueing networks, has been of interest to the queueing community since the early 1990s. At that point, little was known about this and other aspects of queueing networks in general settings. There is now a theory (albeit incomplete), with positive criteria for stability as well as examples where such stability fails. I felt that the time was right to combine in one place a summary of such work.

I wish to thank Jean Picard for his work in organizing the Saint-Flour Summer School and the other participants of the School. I thank the following individuals with whom I consulted at various points: John Baxter, Jim Dai, Michael Harrison, John Hasenbein, Haya Kaspi, Thomas Kurtz, Sean Meyn, and Ruth Williams. I also thank John Baxter for his technical help in preparing the manuscript.

The research that led to these notes was supported in part by U.S. National Science Foundation grants DMS-0226245 and CCF-0729537.

Minneapolis, Minnesota, U.S.A. *Maury Bramson*
March 2008

2

The Classical Networks

In this chapter, we discuss two families of queueing networks whose Markov processes are positive recurrent when $\rho < e$, and whose stationary distributions have explicit product-like formulas. The first family includes networks with the FIFO discipline, and the second family includes networks with the PS and LIFO disciplines, as well as infinite server (IS) networks. We introduced the first three networks in Section 1.2; we will define infinite server networks shortly. These four networks are sometimes known as the "classical networks". Together with their generalizations, they have had a major influence on the development of queueing theory because of the explicit nature of their stationary distributions. For this reason, we present the basic results for the accompanying theory here, although only the FIFO discipline is HL. These results are primarily from [BaCMP75], and papers by F. P. Kelly (such as [Ke75] and [Ke76]) that led to the book [Ke79].

In Section 2.1, we state the main results in the context of the above four networks. We first characterize the stationary distributions for networks consisting of a single station, whose jobs exit from the network when service is completed, without being routed to another class. We will refer to such a station as a *node*. We then characterize the stationary distribution for networks with multiple stations and general routing. Since all states will communicate, the Markov processes for the networks will be positive recurrent, and hence the networks will be stable.

In the remainder of the chapter, we present the background for these results and the accompanying theory. In Section 2.2, we give certain basic properties of stationary and reversible Markov processes on countable state spaces that we will use later. Sections 2.3 and 2.4 apply this material to obtain generalizations of the node-level results in Section 2.1 to the two families of interest to us. The first family, homogeneous nodes, includes FIFO nodes under certain restrictions, and the second family, symmetric nodes, includes PS, LIFO, and IS nodes.

The concept of quasi-reversibility is introduced in Section 2.5. Using quasi-reversibility, the stationary distributions of certain queueing networks can be

M. Bramson, *Stability of Queueing Networks*. Lecture Notes in Mathematics 1950.

written as the product of the stationary distributions of nodes that correspond to the stations "in isolation". These queueing networks include the FIFO, PS, LIFO, and IS networks, and so generalize the network-level results in Section 2.1. Except for Theorem 2.9, all of the material in this chapter is for queueing networks with a countable state space.

The main source of the material in this chapter is [Ke79]. Section 2.2 is essentially an abridged version of the material in Chapter 1 of [Ke79]. Most of the material in Sections 2.3-2.5 is from Sections 3.1-3.3 of [Ke79], with [Wa88], [ChY01], [As03], and lecture notes by J.M. Harrison having also been consulted. The order of presentation here, starting with nodes in Sections 2.3 and 2.4 and ending with quasi-stationarity in Section 2.5, is different.

2.1 Main Results

In this section, we will give explicit formulas for the stationary distributions of FIFO, PS, LIFO, and infinite server networks. Theorems 2.1 and 2.2 state these results for individual nodes, and Theorem 2.3 does so for networks. In Sections 2.3-2.5, we will prove generalizations of these results.

The range of disciplines that we consider here is of limited scope. On the other hand, the routing that is allowed for the network-level results will be completely general. As in Chapter 1, routing will be given by a mean transition matrix $P = \{P_{k,\ell},\ k, \ell = 1, \ldots, K\}$ for which the network is open.[1] For all queueing networks considered in this section, the interarrival times are assumed to be exponentially distributed. When the service times are also exponentially distributed, the evolution of these queueing networks can be expressed in terms of a countable state Markov process. By enriching the state space, more general service times can also be considered in the countable state space setting. This will be useful for the PS, LIFO, and IS networks.

In Section 1.1, we introduced the $M/M/1$ queue with external arrival rate $\alpha = 1$. By employing the reversibility of its Markov process when $m < 1$, we saw that its stationary distribution is given by the geometric distribution in (1.1). Allowing α to be arbitrary with $\alpha m < 1$, this generalizes to

$$\pi(n) = (1 - \alpha m)(\alpha m)^n \quad \text{for } n = 0, 1, 2, \ldots. \tag{2.1}$$

For a surprisingly large group of queueing networks, generalizations of (2.1) hold, with the stationary distribution being given by products of terms similar to those on the right side of (2.1).

[1] In the literature (such as in [Ke79]), deterministic routing is frequently employed. For these lectures, we prefer to use random routing, which was used in [BaCMP75] and has been promulgated in its current form by J. M. Harrison. By employing sufficiently many routes, one can show the two approaches are equivalent. We find the approach with random routing to be notationally more flexible. The formulation is also more amenable to problems involving dynamic scheduling, which we do not cover here.

We first consider queueing networks consisting of just a single node. That is, jobs at the unique station leave the network immediately upon completion of service, without being routed to other classes. Classes are labelled $k = 1, \ldots, K$. The nodes of interest to us in this chapter fall into two basic families, depending on the discipline.

The first family, homogeneous nodes, will be defined in Section 2.3. FIFO nodes, which are the canonical example for this family, will be considered here. For homogeneous nodes, including FIFO nodes, we need to assume that the mean service times m_k at all classes k are equal. In order to avoid confusion with the vector m, we label such a service time by m^s (with "s" standing for "station"). We will refer to such a node as a *FIFO node of Kelly type*. In addition to assuming the interarrival times are exponentially distributed, we assume the same is true for the service times.

The state x of the node at any time will be specified by an n-tuple of the form

$$(x(1), \ldots, x(n)), \tag{2.2a}$$

where n is the number of jobs in the node and

$$x(i) \in \{1, \ldots, K\} \quad \text{for } i = 1, \ldots, n \tag{2.2b}$$

gives the class of the job in the i^{th} position in the node. We interpret $i = 1, \ldots, n$ as giving the order of arrival of the jobs currently in the node; because of the FIFO discipline, all service is directed to the job at $i = 1$. The state space S_0 will be the union of these states. The stochastic process $X(t),\ t \geq 0$, thus defined will be Markov with a countable state space. For consistency with other chapters, we interpret vectors as column vectors, although this is not needed in the present chapter (since matrix multiplication is not employed).

All states communicate with the empty state. So, if $X(\cdot)$ has a stationary distribution, it will be unique. Since there is only a single station, a stationary distribution will exist when the node is subcritical. Theorem 2.1 below gives an explicit formula for the distribution. As in Chapter 1, α_k denotes the external arrival rates at the different classes k; ρ denotes the traffic intensity and is in the present setting given by the scalar

$$\rho = m^s \sum_k \alpha_k. \tag{2.3}$$

As elsewhere in this chapter, when we say that a node (or queueing network) has a stationary distribution, we mean that its Markov process, on the chosen state space, has this distribution. (For us, "distribution" is synonymous with the somewhat longer "probability measure".)

Theorem 2.1. *Each subcritical FIFO node of Kelly type has a stationary distribution π, which is given by*

$$\pi(x) = (1 - \rho) \prod_{i=1}^{n} m^s \alpha_{x(i)}, \tag{2.4}$$

for $x = (x(1), \ldots, x(n)) \in S_0$.

The stationary distribution π in Theorem 2.1 can be described as follows. The probability of there being a total of n jobs in the node is $(1-\rho)\rho^n$. Given a total of n jobs, the probability of there being $n_1, \ldots, n_K$ jobs at the classes $1, \ldots, K$, with $n = n_1 + \ldots + n_K$ and no attention being paid to their order, is

$$\rho^{-n} \binom{n}{n_1, \ldots, n_K} \prod_{k=1}^{K} (m^s \alpha_k)^{n_k}. \tag{2.5}$$

Moreover, given that there are $n_1, \ldots, n_K$ jobs at the classes $1, \ldots, K$, any ordering of the different classes of jobs is equally likely. Note that since all states have positive probability of occurring, the process $X(\cdot)$ is positive recurrent. Consequently, the node is stable.

The other family of nodes that will be discussed in this chapter, symmetric nodes, will be defined in Section 2.4. Standard members of this family are PS, LIFO, and IS nodes. The PS and LIFO disciplines were specified in Chapter 1. In an *infinite server* (IS) node, each job is assumed to start receiving service as soon as it enters the node, which it receives at rate 1. One can therefore think of there being an infinite number of unoccupied servers available to provide service, one of which is selected whenever a job enters the node. All other disciplines studied in these lectures will have only a single server at a given station. We note that although the PS discipline is not HL, it is related to the HLPPS discipline given at the end of Section 5.3.

We consider the stationary distributions of PS, LIFO and IS nodes. As with FIFO nodes, we need to assume that the interarrival times of jobs are exponentially distributed. If we assume that the service times are also exponentially distributed, then the process $X(\cdot)$ defined on S_0 will be Markov. As before, we interpret the coordinates $i = 1, \ldots, n$ in (2.2) as giving the order of jobs currently in the node. For LIFO nodes, this is also the order of arrival of jobs there. For reasons relating to the definition of symmetric nodes in Section 2.4, we will instead assume, for PS and IS nodes, that arriving jobs are, with equal probability $1/n$, placed at one of the n positions of the node, where n is the number of jobs present after the arrival of the job. Since in both cases, jobs are served at the same rate irrespective of their position in the node, the processes $X(\cdot)$ defined in this manner are equivalent to the processes defined by jobs always arriving at the rear of the node.

The analog of Theorem 2.1 holds for PS, LIFO, and IS nodes when the service times are exponentially distributed. We no longer need to assume that the service times have the same means, so in the present setting, the traffic intensity ρ is given by

$$\rho = \sum_k m_k \alpha_k. \tag{2.6}$$

For subcritical PS and LIFO nodes, the stationary distribution π is given by

$$\pi(x) = (1-\rho)\prod_{i=1}^{n} m_{x(i)}\alpha_{x(i)}, \tag{2.7}$$

for $x = (x(1), \ldots, x(n)) \in S_0$. For any IS node, the stationary distribution π is given by

$$\pi(x) = \frac{e^{-\rho}}{n!}\prod_{i=1}^{n} m_{x(i)}\alpha_{x(i)}. \tag{2.8}$$

The stability of the infinite server node for all values of ρ is not surprising, since the total rate of service at the node is proportional to the number of jobs presently there.

For PS, LIFO, and IS nodes, an analogous result still holds when exponential service times are replaced by service times with more general distributions. One employs the "method of stages", which is defined in Section 2.4. One enriches the state space S_0 to allow for different stages of service for each job, with a job advancing to its next stage of service after service at the previous stage has been completed. After service at the last stage has been completed, the job leaves the node. Since the service times at each stage are assumed to be exponentially distributed, the corresponding process $X(\cdot)$ for the node, on this enriched state space S_e, will still be Markov. On the other hand, the total service time required by a given job will be the sum of the exponential service times at its different stages, which we take to be i.i.d. Such service times are said to have *Erlang distributions*.

Using the method of stages, one can extend the formulas (2.7) and (2.8), for the stationary distributions of PS, LIFO, and IS nodes, to nodes that have service distributions which are countable mixtures of Erlang distributions. This result is stated in Theorem 2.2. The state space here for the Markov process $X(\cdot)$ of the node is S_e, which is defined in Section 2.4. The analog of this extension for FIFO nodes is not valid.

Theorem 2.2. *Each subcritical PS node and LIFO node, whose service time distributions are mixtures of Erlang distributions, has a stationary distribution π. The probability of there being n jobs in the node with classes $x(1), \ldots, x(n)$ is given by (2.7). The same is true for any IS node with these service time distributions, but with (2.8) replacing (2.7).*

Mixtures of Erlang distributions are dense in the set of distribution functions, so it is suggestive that a result analogous to Theorem 2.2 should hold for service times with arbitrary distributions. This is in fact the case, although one needs to be more careful here, since one needs to replace S_e with an uncountable state space which specifies the residual service times of jobs at the node. More detail on this setting is given at the end of Section 2.4. Because of the technical difficulties for uncountable state spaces, (2.7) and (2.8) are typically stated for mixtures of Erlang distributions or other related distributions on countable state spaces. Moreover, quasi-reversibility, which we discuss shortly, employs a countable state space setting.

So far in this section, we have restricted our attention to nodes. As mentioned at the beginning of the section, the results in Theorems 2.1 and 2.2 extend to analogous results for queueing networks, which are given in Theorem 2.3, below. FIFO, PS, LIFO, and IS queueing networks are the analogs of the respective nodes, with jobs at individual stations being subjected to the same service rules as before, and, upon completion of service at a class k, a job returning to class ℓ with probability $P_{k,\ell}$, which is given by the mean transition matrix P. In addition to applying to these queueing networks, Theorem 2.3 also applies to networks that are mixtures of such stations, with one of the above four rules holding for any particular station. We also note that the formula in Theorem 2.3 holds for Jackson networks, as a special case of FIFO networks. The product formula for Jackson networks in [Ja63] predates those for the other networks.

In Theorem 2.3, we assume that the service times are exponentially distributed when the station is FIFO, and are mixtures of Erlang distributions in the other three cases. The distribution function π is defined on the state space

$$S = S^1 \times \ldots \times S^J,$$

where $S^j = S_0$ if j is FIFO and $S^j = S_e$ for the other cases, and π^j is defined on S^j. (The choice of K in each factor depends on S^j.)

Theorem 2.3. *Suppose that each station j of a queueing network is either FIFO of Kelly type, PS, LIFO, or IS. Suppose that in the first three cases, the station is subcritical. Then, the queueing network has a stationary distribution π that is given by*

$$\pi(x) = \prod_{j=1}^{J} \pi^j(x^j), \tag{2.9}$$

for $x = (x^1, \ldots, x^J)$. Here, each π^j is either of the form (2.4), (2.7), or (2.8), depending on whether the station j is FIFO of Kelly type, PS or LIFO, or IS, and α_k in the formulas is replaced by λ_k.

Theorem 2.3 will be a consequence of Theorems 2.1 and 2.2, and of the quasi-reversibility of the nodes there. Quasi-reversibility will be introduced in Section 2.5. Using quasi-reversibility, it will be shown, in Theorem 2.11, that the stationary distributions of certain queueing networks can be written as the product of the stationary distributions of nodes that correspond to the individual stations "in isolation". This will mean that when service of a job at a class is completed, the job will leave the network rather than returning to another class (either at the same or a different station). The external arrival rates α_k at classes are replaced by the total arrival rates λ_k of the network in order to compensate for the loss of jobs that would return to them. Quasi-reversibility can be applied to queueing networks whose stations are FIFO, PS, LIFO, or IS. By employing Theorem 2.11, one obtains Theorem 2.3 as a special case of Theorem 2.12.

2.2 Stationarity and Reversibility

In this section, we will summarize certain basic results for countable state, continuous time Markov processes. We define stationarity and reversibility, and provide alternative characterizations. Proposition 2.6, in particular, will be used in the remainder of the chapter.

The Markov processes $X(t),\ t \geq 0$, we consider here will be assumed to be defined on a countable state space S. The space S will be assumed to be irreducible, that is, all states communicate. None of the states will be instantaneous; we will assume there are only a finite number of transitions after a finite time, and hence no explosions. Sample paths will therefore be right continuous with left limits. The transition rate between states x and y will be denoted by $q(x, y)$; the rate at which a transition occurs at x is therefore $q(x) \stackrel{\text{def}}{=} \sum_{y \in S} q(x, y)$. The *embedded jump chain* has mean transition matrix $\{p(x, y),\ x, y \in S\}$, where $p(x, y) \stackrel{\text{def}}{=} q(x, y)/q(x)$ is the probability $X(\cdot)$ next visits y from the state x. The time $X(\cdot)$ remains at a state x before a transition occurs is exponentially distributed with mean $1/q(x)$.

A stochastic process $X(t),\ t \geq 0$, is said to be *stationary* if $(X(t_1), \ldots, X(t_n))$ has the same distribution as $(X(t_1 + u), \ldots, X(t_n + u))$, for each nonnegative $t_1, \ldots, t_n$ and u. Such a process can always be extended to $-\infty < t < \infty$ so that it is stationary as before, but with $t_1, \ldots, t_n$ and u now being allowed to assume any real values. When $X(\cdot)$ is a Markov process, it suffices to consider just $n = 1$ in order to verify stationarity.

A *stationary distribution* $\pi = \{\pi(x),\ x \in S\}$ for a Markov process $X(\cdot)$ satisfies the *balance equations*

$$\pi(x) \sum_{y \in S} q(x, y) = \sum_{y \in S} \pi(y) q(y, x) \quad \text{for } x \in S, \tag{2.10}$$

which say that the rates at which mass leaves and enters a state x are the same. We are assuming here that $\sum_{x \in S} \pi(x) = 1$. Since all states are assumed to communicate, π will be unique. If π exists, then it is the limit of the distributions of the Markov process starting from any initial state. If a measure π satisfying (2.10) with $\sum_{x \in S} \pi(x) < \infty$ exists, then it can be normalized so that $\sum_{x \in S} \pi(x) = 1$. If $\sum_{x \in S} \pi(x) = \infty$, then there is no stationary distribution, and for all $x, y \in S$,

$$P(X(t) = \ell \mid X(0) = x) \to 0 \quad \text{as } t \to \infty.$$

(See, e.g., [Re92] for such basic theory.)

A stochastic process $X(t),\ -\infty < t < \infty$, is said to be *reversible* if $(X(t_1), \ldots, X(t_n))$ has the same distribution as $(X(u - t_1), \ldots, X(u - t_n))$, for each $t_1, \ldots, t_n$ and u. This condition says that the process is stochastically indistinguishable, whether it is run forward or backwards in time. It is easy to see that if $X(\cdot)$ is reversible, then it must be stationary. When $X(\cdot)$

is Markov, it suffices to consider $n = 2$ in order to verify reversibility. The Markov property can be formulated as saying that the past and future states of the process $X(\cdot)$ are independent given the present. It follows that the *reversed process* $\hat{X}(t) \stackrel{\text{def}}{=} X(-t)$ is Markov exactly when $X(\cdot)$ is. If $X(\cdot)$ has stationary measure π, then π is also stationary for $\hat{X}(\cdot)$ and the transition rates of $\hat{X}(\cdot)$ are given by

$$\hat{q}(x, y) = \frac{\pi(y)}{\pi(x)} q(y, z) \quad \text{for } x, y \in S. \tag{2.11}$$

A stationary Markov process $X(\cdot)$ with distribution π is reversible exactly when it satisfies the *detailed balance equations*

$$\pi(x)q(x, y) = \pi(y)q(y, x) \quad \text{for } x, y \in S. \tag{2.12}$$

This condition says that the rate at which mass moves from x to y is the same rate at which it moves in the reverse direction. This condition need not, of course, be satisfied for arbitrary stationary distributions. When (2.12) holds, it often enables one to express the stationary distribution in closed form, as, for example, for the $M/M/1$ queue in Section 1.1. It will always hold for the stationary distribution of any birth and death process, and, more generally, for the stationary distribution of any Markov process on a tree. By summing over ℓ, one obtains the balance equations in (2.10) from (2.12).

An alternative characterization of reversibility is given by Proposition 2.4. We will not employ the proposition elsewhere, but state it because it provides useful intuition for the concept.

Proposition 2.4. *A stationary Markov process is reversible if and only if its transition rates satisfy*

$$\begin{aligned} q(x_1, x_2)q(x_2, x_3) \cdots q(x_{n-1}, x_n)q(x_n, x_1) \\ = q(x_1, x_n)q(x_n, x_{n-1}) \cdots q(x_3, x_2)q(x_2, x_1), \end{aligned} \tag{2.13}$$

for any $x_1, x_2, \ldots, x_n$.

The equality (2.13) says that the joint transition rates of the Markov process are the same along a path if it starts and ends at the same point, irrespective of its direction along the path.

Proof of Proposition 2.4. The "only if" direction follows immediately by plugging (2.12) into (2.13).

For the "if" direction, fix x_0, and define

$$\pi(x) = \prod_{i=1}^{n} [q(x_{i-1}, x_i)/q(x_i, x_{i-1})], \tag{2.14}$$

where $x_n = x$ and $x_0, x_1, \ldots, x_n$ is any path from x_0 to x, with $q(x_i, x_{i-1}) > 0$. One can check using (2.13) that the right side of (2.14) does not depend on the

particular path that is chosen, and so $\pi(x)$ is well defined. To see this, let $\mathscr{P}_1$ and $\mathscr{P}_2$ be any two paths from x_0 to x, and $\hat{\mathscr{P}}_1$ and $\hat{\mathscr{P}}_2$ be the corresponding paths in the reverse directions. Then, the two paths from x_0 to itself formed by linking $\mathscr{P}_1$ to $\hat{\mathscr{P}}_2$, respectively $\mathscr{P}_2$ to $\hat{\mathscr{P}}_1$, satisfy (2.13), from whence the uniqueness in (2.14) will follow.

Assume that for given y, $q(y,x) > 0$. Multiplication of both sides of (2.14) by $q(x,y)$ implies that

$$\begin{aligned}\pi(x)q(x,y) &= q(y,x)\left(\prod_{i=1}^{n} q(x_{i-1},x_i)/q(x_i,x_{i-1}\right)(q(x,y)/q(y,x)) \\ &= q(y,x)\pi(y).\end{aligned}$$

This gives (2.12), since the case $q(x,y) = q(y,x) = 0$ is trivial. Since the process is assumed to be stationary, $\sum_x \pi(x) < \infty$, and so π can be scaled so that $\sum_x \pi(x) = 1$. ■

The following result says that under certain modifications of the transition rates $q(x,y)$, a reversible Markov process will still be reversible. It will not be needed for later work, but has interesting applications.

Proposition 2.5. *Suppose that the transition rates of a reversible Markov process $X(\cdot)$, with state space S and stationary distribution π, are altered by changing $q(x,y)$ to $q'(x,y) = bq(x,y)$ when $x \in A$ and $y \notin A$, for some $A \subseteq S$. Then, the resulting Markov process $X'(\cdot)$ is reversible and has stationary distribution*

$$\pi'(x) = \begin{cases} c\pi(x) & \text{for } x \in A, \\ cb\pi(x) & \text{for } x \notin A, \end{cases} \tag{2.15}$$

where c is chosen so that $\sum_x \pi'(x) = 1$. In particular, when the state space is restricted to A by setting $b = 0$, then the stationary distribution of $X'(\cdot)$ is given by

$$\pi'(x) = \pi(x) \Big/ \sum_{y \in A} \pi(y) \quad \text{for } x \in A. \tag{2.16}$$

Proof. It is easy to check that q' and π' satisfy the detailed balance equations in (2.12). ■

The following illustration of Proposition 2.5 is given in [Ke79].

Example 1. *Two queues with a joint waiting room.* Suppose that two independent $M/M/1$ queues are given, with external arrival rates α_i and mean service times m_i, and $\alpha_i m_i < 1$. Let $X_i(t)$ be the number of customers (or jobs) in each queue at time t. The Markov processes are each reversible with stationary distributions as in (2.1). It is easy to check that the joint Markov process $X(t) = (X_1(t), X_2(t))$, $-\infty < t < \infty$, is reversible, with stationary distribution

$$\pi(n_1, n_2) = (1 - \alpha_1 m_2)(1 - \alpha_2 m_2)(\alpha_1 m_1)^{n_1}(\alpha_2 m_2)^{n_2} \quad \text{for } n_i \in \mathbf{Z}_{+,0}.$$

Suppose now that the queues are required to share a common waiting room of size N, so that a customer who arrives to find N customers already there leaves without being served. This corresponds to restricting $X(\cdot)$ to the set A of states with $n_1 + n_2 \leq N$. By Proposition 2.5, the corresponding process $X'(\cdot)$ is reversible, and has stationary measure

$$\pi'(n_1, n_2) = \pi(0, 0)(\alpha_1 m_1)^{n_1}(\alpha_2 m_2)^{n_2} \quad \text{for } (n_1, n_2) \in A. \qquad \blacksquare$$

It is often tedious to check the balance equations (2.10) in order to determine that a Markov process $X(\cdot)$ is stationary. Proposition 2.6 gives the following alternative formulation. We abbreviate by setting

$$q(x) = \sum_{y \in S} q(x, y), \quad \hat{q}(x) = \sum_{y \in S} \hat{q}(x, y), \tag{2.17}$$

where $q(x, y)$ are the transition rates for $X(\cdot)$ and $\hat{q}(x, y) \geq 0$ are for the moment arbitrary. When $X(\cdot)$ has stationary distribution π and $\hat{q}(x, y)$ is given by (2.11), it is easy to check that

$$q(x) = \hat{q}(x) \quad \text{for all } x. \tag{2.18}$$

The proposition gives a converse to this. As elsewhere in this section, we are assuming that S is irreducible.

Proposition 2.6. *Let $X(t)$, $-\infty < t < \infty$, be a Markov process with transition rates $\{q(x, y),\ x, y \in S\}$. Suppose that for given quantities $\{\hat{q}(x, y),\ x, y \in S\}$ and $\{\pi(x),\ x \in S\}$, with $\hat{q}(x, y) \geq 0$, $\pi(x) > 0$, and $\sum_x \pi(x) = 1$, that q, $\hat{q}$, and π satisfy (2.11) and (2.18). Then, π is the stationary distribution of $X(\cdot)$ and $\hat{q}$ gives the transition rates of the reversed process.*

Proof. It follows, by applying (2.11) and then (2.18), that

$$\sum_{x \in S} \pi(x) q(x, y) = \pi(y) \sum_{x \in S} \hat{q}(y, x) = \pi(y) \hat{q}(y) = \pi(y) q(y).$$

So, π is stationary for $X(\cdot)$. The transition rates of the reversed process are therefore given by (2.11). $\blacksquare$

Proposition 2.6 simplifies the computations needed for the demonstration of stationarity by replacing the balance equations, that involve a large sum and the stationary distribution π, by two simpler equations, (2.18), which involves just a large sum, and (2.11), which involves just π. On the other hand, the application of Proposition 2.6 typically involves guessing $\hat{q}$ and π. In situations where certain choices suggest themselves, the proposition can be quite useful. It will be used repeatedly in the remainder of the chapter.

2.3 Homogeneous Nodes of Kelly Type

FIFO nodes of Kelly type belong to a larger family of nodes whose stationary distributions have similar properties. We will refer to such a node as a *homogeneous node of Kelly type*. Such nodes are defined as follows.

Consider a node with K classes. The state $x \in S_0$ of the node at any time is specified by an n-tuple as in (2.2), when there are n jobs present at the node. The ordering of the jobs is assumed to remain fixed between arrivals and service completions of jobs. When the job in position i completes its service, the position of the job is filled with the jobs in positions $i+1, \ldots, n$ moving up to positions $i, \ldots, n-1$, while retaining their previous order. Similarly, when a job arrives at the node, it is assigned some position i, with jobs previously at positions $i, \ldots, n$ being moved back to positions $i+1, \ldots, n+1$. Each job requires a given random amount of service; when this is attained, the job leaves the node. As throughout this chapter, interarrival times are required to be exponentially distributed. As elsewhere in these lectures, all interarrival and service times are assumed to be independent.

We will say that such a node is a *homogeneous node* if it also satisfies the following properties:

(a) The amount of service required by each job is exponentially distributed with mean m_k, where k is the class of the job.

(b) The total rate of service supplied at the node is $\phi(n)$, where n is the number of jobs currently there.

(c) The proportion of service that is directed at the job in position i is $\delta(i,n)$. Note that this proportion does not depend on the class of the job.

(d) When a job arrives at the node, it moves into position i, $i = 1, \ldots, n$, with probability $\beta(i,n)$, where n is the number of jobs in the node including this job. Note that this probability does not depend on the class of the job.

When the mean m_k does not depend on the class k, we will say that such a node is a *homogeneous node of Kelly type*. We will analyze these nodes in this section. We use m^s when the mean service times of a node are constant, as we did in Section 2.1.

The rate at which service is directed to a job in position i is $\delta(i,n)\phi(n)$. So, the rate at which service at the job is completed is $\delta(i,n)\phi(n)/m^s$. We will assume that $\phi(n) > 0$, except when $n = 0$. The rate at which a job arrives at a class k and position i from outside the node is $\alpha_k\beta(i,n)$. We use here the mneumonics β and δ to suggest births and deaths at a node. Of course, $\sum_{i=1}^{n} \beta(i,n) = \sum_{i=1}^{n} \delta(i,n) = 1$.

We have emphasized in the above definition that the external arrival rates and service rates β and δ do not depend on the class of the job. This is crucial for Theorem 2.7, the main result in this section. This restriction will also be

needed in Section 2.4 for symmetric nodes, as will be our assumption that the interarrival times are exponentially distributed. On the other hand, the assumptions that the service times be exponential and that their means m_k be constant, which are needed in this section, are not needed for symmetric nodes. We note that by scaling time by $1/m^s$, one can set $m^s = 1$, although we prefer the more general setup for comparison with symmetric nodes and for application in Section 2.5.

In Section 2.5, we will be interested in homogeneous queueing networks of Kelly type. *Homogeneous queueing networks* are defined analogously to homogeneous nodes. Jobs enter the network independently at the different stations according to exponentially distributed random variables, and are assigned positions at these stations as in (d). Jobs at different stations are served independently, as in (a)-(c), with departing jobs from class k being routed to class ℓ with probability $P_{k,\ell}$ and leaving the network with probability $1 - \Sigma_\ell P_{k,\ell}$. Jobs arriving at a class ℓ from within the network are assigned positions as in (d), according to the same rule as was applied for external arrivals. The external arrival rates and the quantities in (a)-(d) are allowed to depend on the station. When the mean service times m_k are assumed to depend only on the station $j = s(k)$, we may write m_j^s; we refer to such networks as *homogeneous queueing networks of Kelly type*.

The most important examples of homogeneous nodes are FIFO nodes. Here, one sets $\phi(n) \equiv 1$,

$$\beta(i,n) = \begin{cases} 1 & \text{for } i = n, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\delta(i,n) = \begin{cases} 1 & \text{for } i = 1, \\ 0 & \text{otherwise,} \end{cases}$$

for $n \in \mathbf{Z}_+$. That is, arriving jobs are always placed at the end of the queue and only the job at the front of the queue is served. Another example is given in [Ke79], where arriving jobs are again placed at the end of the queue, but where L servers are available to serve the first L jobs, for given L. In this setting, $\phi(n) = L \wedge n$, β is defined as above, and

$$\delta(i,n) = \begin{cases} 1/n & \text{for } i \le n \le L, \\ 1/L & \text{for } i \le L < n, \\ 0 & \text{for } i > L. \end{cases}$$

The main result in this section is Theorem 2.7, which is a generalization of Theorem 2.1. Since the total service rate ϕ that is provided at the node can vary, the condition

$$B \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} \left(\rho^n \Big/ \prod_{i=1}^{n} \phi(i) \right) < \infty \tag{2.19}$$

replaces the assumption in Theorem 2.1 that the node is subcritical. Here, $\rho = m^s \sum_{k=1}^{K} \alpha_k$ is the traffic intensity.

Theorem 2.7. *Suppose that a homogeneous node of Kelly type satisfies $B < \infty$ in (2.19). Then, it has a stationary distribution π that is given by*

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (m^s \alpha_{x(i)} / \phi(i)), \tag{2.20}$$

for $x = (x(1), \ldots, x(n)) \in S_0$.

As was the case in Theorem 2.1, the structure of the stationary distribution π in Theorem 2.7 exhibits independence at multiple levels. The probability of there being a total of n jobs at the node is $\rho^n / B \prod_{i=1}^{n} \phi(i)$. Given a total of n jobs at the node, the probability of there being $n_1, \ldots, n_k$ jobs of classes $1, \ldots, K$, respectively, is given by the multinomial distribution in (2.5). Moreover, the ordering of the different classes of jobs is equally likely. As was the case in Theorem 2.1, all states communicate with the empty state, and the Markov process $X(\cdot)$ for the node is positive recurrent.

Demonstration of Theorem 2.7

The proof of Theorem 2.7 that we will give is based on Proposition 2.6. In order to employ the proposition, we need to choose quantities $\hat{q}$ and π so that (2.11) and (2.18) are satisfied for them and the transition rates q of the Markov process $X(\cdot)$ for the node. It will then follow from Proposition 2.6 that π is the stationary distribution for the node and $\hat{q}$ gives the transition rates for the reversed process $\hat{X}(\cdot)$. A similar argument will be used again for symmetric nodes in Section 2.4 and for networks consisting of quasi-reversible nodes in Section 2.5. We will summarize a more probabilistic argument for Theorem 2.7 at the end of the section.

In order to demonstrate Theorem 2.7, we write q and our choices for $\hat{q}$ and π explicitly in terms of α, β, δ, and ϕ. To be able to reuse this argument in Section 2.4 for symmetric nodes, we write m_k for the mean service times, which in the present case reduces to the constant m^s.

The nonzero transition rates $q(x, y)$ take on two forms, depending on whether the state is obtained from x by the arrival or exit of a job. In the former case, we write $y = a_{k,i}(x)$ if a class k job arrives at position i; in the latter case, it follows that $x = a_{k,i}(y)$, if i is the position of the exiting class k job. One then has

$$q(x, y) = \begin{cases} \alpha_k \beta(i, n_y) & \text{for } y = a_{k,i}(x), \\ m_k^{-1} \delta(i, n_x) \phi(n_x) & \text{for } x = a_{k,i}(y), \end{cases} \tag{2.21}$$

where n_x and n_y are the number of jobs at the node for states x and y.

Finding the transition rates $\hat{q}(x, y)$ of the reversed process involves some guessing, motivated by our idea of what $\hat{X}(\cdot)$ should look like. It is reasonable

to guess that $\hat{X}(\cdot)$ is also the Markov process for a homogeneous node. The external arrival rates α_k and $\hat{\alpha}_k$ will then be the same for both processes, since arrivals for $\hat{X}(\cdot)$ correspond to exits for $X(\cdot)$, and under the stationary distribution π, the two rates must be the same. The mean service times m_k and $\hat{m}_k$ will also be the same. It is reasonable to guess that $\hat{\phi}(n) = \phi(n)$; this would be the case if $X(\cdot)$ were reversible, as it is for the $M/M/1$ queue. Running $X(\cdot)$ backwards in time mentally, it is also tempting to set

$$\hat{\beta}(i,n) = \delta(i,n),\ \hat{\delta}(i,n) = \beta(i,n) \quad \text{for } n \in \mathbf{Z}_+,\ i \le n.$$

For instance, if the original node is FIFO, then jobs arrive at $i = n$ and exit at $i = 1$; if the node is run backwards in time, jobs arrive at $i = 1$ and exit at $i = n$. Substitution of these choices for α, β, δ, ϕ, and m in (2.21) yields

$$\hat{q}(x,y) = \begin{cases} \alpha_k \delta(i, n_y) & \text{for } y = a_{k,i}(x), \\ m_k^{-1} \beta(i, n_x) \phi(n_x) & \text{for } x = a_{k,i}(y). \end{cases} \tag{2.22}$$

We still need to choose our candidate for the stationary distribution π of both $X(\cdot)$ and $\hat{X}(\cdot)$. The equality (2.11), that is needed for Proposition 2.6, is equivalent to

$$\pi(x) q(x,y) = \pi(y) \hat{q}(y,x) \tag{2.23}$$

holding whenever $y = a_{k,i}(x)$ or $x = a_{k,i}(y)$. When $y = a_{k,i}(x)$ and $q(x,y) > 0$, substitution of (2.21) and (2.22) into (2.23) implies that

$$\pi(y)/\pi(x) = m_k \alpha_k / \phi(n_y). \tag{2.24}$$

The case $x = a_{k,i}(y)$ yields the same equality, but with the roles of x and y reversed. Reasoning backwards, it is not difficult to see that (2.23) also follows from (2.24).

Set $x = (x(1), \ldots, x(n))$, for $n \in \mathbf{Z}_{+,0}$. One can repeatedly apply (2.24) by removing jobs from x one at a time, starting from the last, until the empty state is reached. We therefore choose π so that

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (m_{x(i)} \alpha_{x(i)} / \phi(i)), \tag{2.25}$$

where the normalizing constant $B = 1/\pi(\emptyset)$. Under (2.25), (2.24) must hold. We have therefore verified (2.11) for this choice of π. In particular, this holds for $m_k \equiv m^s$, as in Theorem 2.7.

In order to employ Proposition 2.6, we also need to verify (2.18). One can check that

$$\sum_y q(x,y) = \sum_k \alpha_k + \phi(n_x) \sum_i m_{x(i)}^{-1} \delta(i, n_x). \tag{2.26}$$

The first sum on the right side of (2.26) follows by summing the top line of (2.21) over all i and k. The relationship $x = a_{k,i}(y)$ implies that $x(i) = k$, and

so the last sum in (2.26) follows by summing the last line of (2.21) over all i. Using the same reasoning, one obtains the formula

$$\sum_y \hat{q}(x,y) = \sum_k \alpha_k + \phi(n_x) \sum_i m_{x(i)}^{-1} \beta(i, n_x) \tag{2.27}$$

from (2.22). As before, the first sum on the right side is obtained from arriving jobs and the last sum is obtained from exiting jobs.

We see from (2.26) and (2.27) that a sufficient condition for (2.18) is that

$$\sum_i m_{x(i)}^{-1} \delta(i, n_x) = \sum_i m_{x(i)}^{-1} \beta(i, n_x). \tag{2.28}$$

In Theorem 2.7, $m_k \equiv m^s$, which factors outside of the sum on both sides of (2.28). Since the resulting sums both equal 1, (2.28), and hence (2.18), holds in this setting. Note that this is the only point in the argument at which we need m_k to be constant.

We have shown that both (2.11) and (2.18) are satisfied for q and q' given by (2.21) and (2.22), and π given by (2.25), under the assumptions in Theorem 2.7. It therefore follows from Proposition 2.6 that π is the stationary distribution for the Markov process with transition rates q. This implies Theorem 2.7.

Some observations

One can generalize the above proof of Theorem 2.7 so that it applies to homogeneous queueing networks of Kelly type, rather than to just homogeneous nodes of Kelly type as in the theorem. Then, the stationary distribution π can be written as the product of stationary distributions π^j of nodes corresponding to the individual stations, when they operate "in isolation". This is done in Section 3.1 of [Ke79]. We prefer to postpone the treatment of homogeneous networks until Section 2.5, where they are considered within the context of quasi-reversibility.

One can give a more probabilistic proof of Theorem 2.7 that is based on the following intuitive argument. The rates $\beta(i,n)$, $\delta(i,n)$ and $\phi(n)$ governing the arrival and service rates of jobs, as well as the mean service time m^s, do not distinguish between classes of jobs. Jobs are therefore served as they would be for an $M/M/1$ queue modified to have the total rate of service $\phi(n)$, when there are n jobs, and having the arrival rate $\sum_k \alpha_k$. By randomly choosing the class of each job, with probability $\alpha_k / \sum_k \alpha_k$ for each k, at either time 0 or at some later time t, the distributions at time t of the two resulting processes will be the same. If the stationary distribution for the modified $M/M/1$ queue is chosen as its initial distribution, the resulting distribution π' for the K classes will therefore also be stationary.

One can show, by using reversibility, that the probability of there being n jobs for the stationary distribution of the modified $M/M/1$ queue is $\rho^n / B \prod_{i=1}^n \phi(i)$, where B is as in (2.19). Because of the random way in which

the classes of jobs are chosen above for π', the remaining properties in the alternative characterization of π in (2.20), that are given after the statement of Theorem 2.7, also hold. Therefore, $\pi' = \pi$, as desired.

We also note the following consequence of the proof of Theorem 2.7, that is a special case of phenomena that will be discussed in Section 2.5. (It also follows from the alternative argument that was sketched above.) The transition rates $\hat{q}$ of the reversed Markov process $\hat{X}(\cdot)$ in (2.22) are the rates for a homogeneous node of Kelly type. For this reversed node, arrivals are therefore given by K independent Poisson processes for the different classes, that are independent of the initial state. These arrivals correspond to exiting jobs for the original homogeneous node. It follows that the K different exit processes for the classes are also independent Poisson processes that are independent of any future state of the node.

2.4 Symmetric Nodes

PS, LIFO, and IS nodes all belong to the family of symmetric nodes. They are defined similarly to the homogeneous nodes of Kelly type in the previous section, with a few major differences. The basic framework is the same, with state space S_0 given by (2.2) and existing jobs being reordered as before upon the arrival and departure of jobs at the node. Moreover, the interarrival times are exponentially distributed.

In order for such a node to be a *symmetric node*, we require that it also satisfy the following properties:

(a) The amount of service required by each job is exponentially distributed with mean m_k. We will soon allow more general distributions, but this will require us to extend the state space.

(b) The total rate of service supplied at the node is $\phi(n)$, where n is the number of jobs currently there.

(c) The proportion of service that is directed at the job in position i is $\beta(i, n)$. Note that the proportion does not depend on the class of the job.

(d) When a job arrives at the node, it moves into position i, $i = 1, \ldots, n$, with probability $\beta(i, n)$, where n is the number of jobs in the node including this job. This function is the same as that given in (c).

In Section 2.5, we will also be interested in symmetric queueing networks. These networks are defined analogously, with properties (a)-(d) being assumed to hold at each station, and departing jobs from a class k being routed to a class ℓ with probability $P_{k,\ell}$. More detail is given in Section 2.3 for homogeneous networks, where the procedure is the same.

The properties (a)-(d) given here are more restrictive than the properties (a)-(d) in the previous section in that we now assume, in the notation of

Section 2.3, that $\delta = \beta$, in parts (c) and (d). These properties are more general in that the service time means m_k need no longer be equal at different classes. After comparing the stationary distributions of these nodes with those of Section 2.3, we proceed to generalize the exponential distributions of the service times in (a) in two steps, first to mixtures of Erlang distributions, and then to arbitrary distributions.

The PS, LIFO, and IS nodes are standard examples of symmetric nodes. For the PS discipline, one sets

$$\beta(i,n) = 1/n \quad \text{for } i \le n,$$

for $n \in \mathbf{Z}_+$, and for LIFO, one sets

$$\beta(i,n) = \begin{cases} 1 & \text{for } i = n, \\ 0 & \text{otherwise.} \end{cases}$$

In both cases, $\phi(n) \equiv 1$. For IS nodes, $\phi(n) = n$ and

$$\beta(i,n) = 1/n \quad \text{for } i \le n.$$

The analog of Theorem 2.7 holds for symmetric nodes when $B < \infty$, for B in (2.19), with the formula for the stationary distribution π,

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (m_{x(i)} \alpha_{x(i)} / \phi(i)), \tag{2.29}$$

for $x = (x(1), \ldots, x(n))$, replacing (2.20). One can check that the same argument as before is valid. One applies Proposition 2.6, for which one needs to verify (2.11) and (2.18). As before, the formulas for q and $\hat{q}$ are given by (2.21) and (2.22), but with $\delta = \beta$. The formula for π is given by (2.25).

The argument we gave for (2.11) involved no restrictions on m_k, and therefore holds in the present context as well. The argument we gave for (2.18) consisted of equating (2.26) and (2.27), and therefore verifying (2.28). Previously, (2.28) held since m_k^{-1} was constant, and so could be factored out of both sums. It is now satisfied since $\delta = \beta$, and so the summands are identical. One can therefore apply Proposition 2.6, from which the analog of Theorem 2.7 for symmetric nodes follows, with (2.29) replacing (2.20).

The method of stages

As mentioned earlier, the assumption that the service times be exponentially distributed is not necessary for symmetric nodes. By applying the *method of stages*, one can generalize the service time distributions to mixtures of Erlang distributions. (These are gamma distributions that are convolutions of identically distributed exponential distributions.) We will show that the analog of the formula (2.29) for the stationary distribution continues to hold in this more general setting.

In order for the stochastic process corresponding to the node to remain Markov for more general service times, we need to enrich the state space S_0. For this, we replace each coordinate $x(i)$ in (2.2) by the triple $\mathbf{x}(i) = (x(i), s(i), v(i))$, where

$$x(i) \in \{1, \ldots, K\}, \quad s(i) \in \mathbf{Z}_+, \quad v(i) \in \{1, \ldots, s(i)\}. \tag{2.30}$$

Such a triple gives the *refined class* of a job, with $x(i)$ denoting its class as before. The third coordinate $v(i)$ gives the current *stage* of a job, with $s(i)$ denoting the total number of stages the job visits before leaving the node. The state space S_e will consist of such n-tuples $x = (\mathbf{x}(1), \ldots, \mathbf{x}(n))$, $n \in \mathbf{Z}_{+,0}$, under a mild restriction to ensure all states are accessible.

The basic dynamics of the node are the same as before, which satisfies the properties for symmetric nodes given at the beginning of the section, including properties (b)-(d); we are generalizing here the assumption in (a). Instead of entering the node at a class k with rate α_k, jobs enter at a refined class (k, s, s) with rate $\alpha_k p_k(s)$, where $\sum_s p_k(s) = 1$. Once at a refined class (k, s, v), such a job moves to $(k, s, v-1)$ after completing its service requirement, which is exponentially distributed with mean $m_k(s)$. After a job completes its service at the stage $v = 1$, it leaves the node. The current stage v of a job can therefore be thought of as the residual number of stages remaining before the job leaves the node. We note that the proportion of service that is directed at a job in position i is $\beta(i, n)$, which does not depend on its class or refined class.

The distribution of the service time that is required for the job between entering and leaving class k is a mixture of Erlang distributions, and has mean

$$m_k \stackrel{\text{def}}{=} \sum_s s p_k(s) m_k(s). \tag{2.31}$$

The state space S_e mentioned earlier is defined to consist of n-tuples whose components (k, s, v) satisfy $p_k(s) > 0$, in order to exclude inaccessible states. Under this restriction, all states will communicate. The state space is of course countable. When $p_k(1) = 1$ at all k, the service times are all exponentially distributed and the model reduces to the one considered at the beginning of the section. As with homogeneous and symmetric nodes with exponentially distributed service times, the networks corresponding to symmetric nodes with stages can be defined in the natural way.

We wish to show that the nodes just defined have stationary distributions π that generalize (2.29). This result is stated in Theorem 2.8.

Theorem 2.8. *Suppose that the service times of a symmetric node are mixtures of Erlang distributions, and that the node satisfies $B < \infty$ in (2.19). Then, the node has a stationary distribution π that is given by*

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (p_{x(i)}(s(i)) m_{x(i)}(s(i)) \alpha_{x(i)} / \phi(i)), \tag{2.32}$$

where $x = (\mathbf{x}(1), \ldots, \mathbf{x}(n)) \in S_e$ and $\mathbf{x}(i) = (x(i), s(i), v(i))$, for $i = 1, \ldots, n$.

As was the case in Theorem 2.7 and in (2.29), the structure of the stationary distribution π is Theorem 2.8 exhibits independence at multiple levels. The probability of there being a total of n jobs at the node is $\rho^n / B \prod_{i=1}^{n} \phi(i)$. Given a total of n jobs at the node, the probability of there being $n_1, \ldots, n_k$ jobs at the classes $1, \ldots, K$ is given by the multinomial

$$\rho^{-n} \binom{n}{n_1, \ldots, n_K} \prod_{k=1}^{K} (m_k \alpha_k)^{n_k}. \tag{2.33}$$

The ordering of these classes is equally likely. Note that none of these quantities depends on the particular service time distributions, except for the means m_k.

The stationary distribution also has the following refined structure. Given the class of the job at each position i, the probability of the job at a given position, whose class is k, having refined class (k, x, v) is

$$p_k(s) m_k(s) / m_k,$$

and these events are independent at different i. Summing over all stages strictly greater than v and over all s, while keeping everything else fixed, this implies that the conditional probability of the job at position i being in a strictly earlier stage than v is

$$m_k^{-1} \sum_{s} (s - v) p_k(s) m_k(s). \tag{2.34}$$

Note that (2.34) depends on the actual service time distributions, and not just on their means.

Demonstration of Theorem 2.8

In order to demonstrate Theorem 2.8, we employ Proposition 2.6. To do so, we need to verify (2.11) and (2.18) for the transition rates q of the Markov process on S_e corresponding to the node, with an appropriate choice of the quantities $\hat{q}$ and π.

In order to specify q and $\hat{q}$, we modify the function $a_{k,i}(\cdot)$ we used for exponential service times on the state space S_0. Here, $a_{k,s,i}(x)$ will denote the state y obtained from state x by the arrival of a job at position i, with refined class (k, s, s). Since jobs exit from the node at refined classes of the form $(k, s, 1)$ (rather than at (k, s, s)), we need additional notation. With an eye on defining $\hat{q}$, we denote by $\hat{a}_{k,s,i}(x)$ the state y that is obtained from x by inserting a job with refined class $(k, s, 1)$ at i; the positions of jobs already at the node are shifted in the usual way. We also denote by $\mathfrak{s}_i(x)$ the state y obtained from a state x satisfying $2 \leq v(i) \leq s(i)$, when the stage at i advances to $v(i) - 1$. ($\mathfrak{s}_i(x)$ is not defined for other x.)

Using this notation, one can check that q is given by

$$q(x,y) = \begin{cases} \alpha_k p_k(s)\beta(i,n_y) & \text{for } y = a_{k,s,i}(x), \\ (m_k(s))^{-1}\beta(i,n_x)\phi(n_x) & \text{for } x = \hat{a}_{k,s,i}(y), \\ (m_k(s))^{-1}\beta(i,n_x)\phi(n_x) & \text{for } y = \mathfrak{s}_i(x), \end{cases} \tag{2.35}$$

with $q(x,y) = 0$ otherwise. Employing the same motivation as in Section 2.3, we choose $\hat{q}$ so that

$$\hat{q}(x,y) = \begin{cases} \alpha_k p_k(s)\beta(i,n_y) & \text{for } y = \hat{a}_{k,s,i}(x), \\ (m_k(s))^{-1}\beta(i,n_x)\phi(n_x) & \text{for } x = a_{k,s,i}(y), \\ (m_k(s))^{-1}\beta(i,n_y)\phi(n_y) & \text{for } x = \mathfrak{s}_i(y), \end{cases} \tag{2.36}$$

with $\hat{q}(x,y) = 0$ otherwise. The transition function $\hat{q}$ is the same as q, except that jobs arrive at the stage $v = 1$, exit at $v = s(i)$, with changes in stage occurring from $v-1$ to v, for $2 \le v \le s(i)$. We choose π as in (2.32).

The assumptions for Proposition 2.6 can be verified as they were in Section 2.3 for the space S_0, with only a small change in argument. The argument for (2.11) is the same when either $y = a_{k,s,i}(x)$ or $x = \hat{a}_{k,s,i}(y)$. For $y = a_{k,s,i}(x)$, the equality (2.24) is replaced by its analog

$$\pi(y)/\pi(x) = m_k(s)p_k(s)\alpha_k/\phi(n_y).$$

When $y \in \mathfrak{s}_i(x)$, one has

$$\pi(y)/\pi(x) = q(x,y)/\hat{q}(y,x) = 1, \tag{2.37}$$

in which case (2.11) is obvious. So, (2.11) holds in all cases. (Note that for the homogeneous nodes in Section 2.3 with distinct β and δ, the analog of (2.37) does not hold, and so the method of stages employed here for generalizing the exponential distributions will not work.)

The formula (2.18) holds for the same reasons as before, except that one now has

$$\sum_y q(x,y) = \sum_y \hat{q}(x,y) = \sum_k \alpha_k + \phi(n_x)\sum_i (m_{x(i)}(s(i)))^{-1}\beta(i,n_x),$$

with $m_{x(i)}(s(i))$ replacing $m_{x(i)}$. So, the assumptions for Proposition 2.6 hold. Application of the proposition therefore implies Theorem 2.8.

One can generalize the above argument so that it applies to symmetric queueing networks. Then, the stationary distribution π can be written as the product of stationary distributions π^j of nodes corresponding to the individual stations. As in the previous section, we choose to postpone the treatment of symmetric networks until Section 2.5, where they are considered within the context of quasi-reversibility.

Extensions to general distributions

We have employed the method of stages to generalize the formula (2.29), for the stationary distribution of symmetric nodes with exponentially distributed service times, to the formula (2.32), which holds for service times that are mixtures of Erlang distributions. The method of stages can also be employed to construct service times with other distributions. This approach is employed, for example, in Section 3.6 of [Wa88] and in Section 3.4 of [As03], where the more general *phase-type distributions* are constructed. [As03] also gives further background on the problem.

Let $\mathcal{H} = \bigcup_{N=1}^{\infty} \mathcal{H}_N$, where $\mathcal{H}_N$ denotes the family of mixtures of Erlang distributions, but with the restriction that $m_k(s) = 1/N$ for all k and s. It is not difficult to show that $\mathcal{H}$ is dense in the set of distribution functions, with respect to the weak topology; this result is given in Exercise 3.3.3 in [Ke79]. The basic idea is that the sum of Ns i.i.d. copies of an exponential distribution, with mean $1/N$, has mean s and variance s/N, and so, for large N, is concentrated around s. For large enough N, one can therefore approximate a given service time distribution function F_k as closely as desired, by setting

$$p_k^N(s) = F_k^N(s/N) - F_k^N((s-1)/N), \quad \text{for } s \in \mathbf{Z}_+, \tag{2.38}$$

equal to the probability that a job chooses a refined class with s stages, when it enters class k. Here, F_k^N is the distribution function satisfying $F_k^N(s') = F_k(s')$, for $Ns' \in \mathbf{Z}_+$, and which is constant off this lattice. For the same reason, the phase-type distributions that were mentioned in the previous paragraph are also dense.

Since the family $\mathcal{H}$ of mixtures of Erlang distributions is dense, it is tempting to infer that a stationary distribution will always exist for a symmetric node with any choice of service time distributions F_k satisfying (2.19), with m_k replacing m^s in the definition of ρ, and that this distribution has the same product structure as given below the statement of Theorem 2.8. Such a result holds, although the state space needs to be extended so that the last component v of the refined class (k, s, v) of a job can now take on any value in $(0, s]$; this corresponds to the residual service time of that job. The resulting state space S_∞ for the Markov process is uncountable; this causes technical problems which we discuss at the end of the section. Since the space is uncountable, it is most natural to formulate the result in a manner similar to that given below Theorem 2.8.

Theorem 2.9. *Suppose that a symmetric node satisfies $B < \infty$ in (2.19). Then, it has a stationary distribution π on S_∞ with*

$$\pi(x(i) = k(i),\ i = 1, \ldots, n) = B^{-1} \prod_{i=1}^{n} (m_{k(i)} \alpha_{k(i)} / \phi(i)). \tag{2.39}$$

Conditioned on any such set, the residual service times of the different jobs are independent, with the probability that a job of a class k has residual service time at most r being

$$F_k^*(r) = \frac{1}{m_k} \int_0^r (1 - F_k(s))ds. \tag{2.40}$$

One can motivate (2.39) by applying Theorem 2.8 to a sequence of nodes indexed by N, with p_k^N for each class k being given by (2.38). Since $F_k^N \Rightarrow F_k$ and $m_k^N \to m_k$ as $N \to \infty$, one should expect (2.39) to follow from (2.32). In order to motivate (2.40), one can reason as follows. Applying Theorem 2.8, one can check that, under the stationary distribution π^N and conditioned on the job at a given position being k, the probability that the stage there is strictly greater than v is

$$\sum_{s>v} (s - v)p_k^N(s) \Big/ \sum_s sp_k^N(s).$$

One can also check that $\sum_{s>v}(s-v)p_k^N(s) = \sum_{s>v} \bar{F}_k^N(s/N)$, where $\bar{F}_k^N(s) = 1 - F_k^N(s)$. So, the above quantity equals

$$\frac{1}{N} \sum_{s>v} \bar{F}_k^N(s/N) \Big/ \frac{1}{N} \sum_s \bar{F}_k^N(s/N).$$

By setting $v = Nr,\ r \geq 0$, and applying the Monotone Convergence Theorem to the numerator and denominator separately, one obtains the limit

$$\frac{1}{m_k} \int_r^\infty \bar{F}_k(s)ds. \tag{2.41}$$

On the other hand, the same reasoning as above (2.38) implies that, for large N, the stage v scaled by N typically approximates the residual service time. So, (2.41) will also give the limiting distribution of the residual service times as $N \to \infty$. Taking the complementary event, one obtains (2.40) from (2.41).

The same reasoning as above implies (2.40) is also the probability that the amount of service that has been received by a job is at least r. We point out that F_k^*, as in (2.40), is the distribution of the residual time for the stationary distribution of a renewal process, with lifetime distribution F_k.

The above reasoning, although suggestive, is not rigorous. In particular, implicit in the explanations for both (2.39) and (2.40) is the assumption that the stationary distribution π and the residual service time distributions $F_1^*, \ldots, F_K^*$ are continuous in $F_1, \ldots, F_K$. A rigorous justification for Theorem 2.9 is given in [Ba76]. There, the Markov processes $X^N(\cdot)$ corresponding to the above sequences of nodes are constructed on a common uncountable state space S, where the residual times of the jobs are included in the state. The Markov process that corresponds to the node in Theorem 2.9 is expressed as a weak limit of the processes $X^N(\cdot)$; this provides a rigorous justification for the convergence of π^N and $F_1^N, \ldots, F_K^N$ that is needed for the theorem. [Ba76] in fact demonstrates the analog of Theorem 2.9 in the more general context of symmetric queueing networks.

The above uncountable state space setting requires a more abstract framework than one typically wishes for a basic theory of symmetric networks. The countable state space setting is typically employed in the context of either mixtures of Erlang distributions, the more general phase-type distributions, or some other dense family of distributions. (See, for example, Section 3.4 of [As03], for more detail.) Quasi-reversibility, which we discuss in the next section, also employs a countable state space setting.

2.5 Quasi-Reversibility

In Sections 2.3 and 2.4, we showed that the stationary distributions of homogeneous nodes of Kelly type and symmetric nodes are of product form. Employing quasi-reversibility, it will follow that the stationary distributions of the corresponding queueing networks are also of product form, with the states at the individual stations being independent and the distributions there being given by Theorems 2.7 and 2.8.

Quasi-reversibility has two important consequences. When a queueing network can be decomposed in terms of nodes that are quasi-reversible, the stationary distribution of the network can be written as the product of the stationary distributions of these individual nodes. It will also follow from the "duality" present in quasi-reversibility that the exit processes of such networks are independent Poisson processes, a property that is inherited from the processes of external arrivals of the network. Quasi-reversibility does not depend on the routing in a network, but holds only under certain disciplines, like those mentioned in the first paragraph.

In this section, in order to avoid confusion, we will say that a departing job from a class that leaves the network *exits* from the network (as opposed to being routed to another class). For nodes, such as in the two previous sections, departures and exits are equivalent.

Before introducing quasi-reversibility, we first motivate the basic ideas with a finite sequence of $M/M/1$ queues that are placed in tandem:

$$\rightarrow 1 \rightarrow 2 \rightarrow \ldots \rightarrow k \rightarrow \ldots \rightarrow K \rightarrow . \tag{2.42}$$

Jobs are assumed to enter the one-class station, with $j = k = 1$, according to a rate-α Poisson process, and are served in the order of their arrival there. Upon leaving station 1, jobs enter station 2 and are served there, and so on, until leaving the network after having been served at station K. All jobs are assumed to have exponentially distributed service times which are independent, with means m_k so that $\alpha m_k < 1$.

An $M/M/1$ queue with external arrival rate α and mean service time m has stationary distribution given by (2.1), if $m\alpha < 1$. Under this distribution, the corresponding Markov process $X(t)$, $-\infty < t < \infty$, is reversible, and so is stochastically equivalent to its reversed process $\hat{X}(t) = X(-t)$. In particular,

the stationary process $X_1(\cdot)$ for the number of jobs at station 1 is stochastically equivalent to its reversed process $\hat{X}_1(\cdot)$. Interpreting $\hat{X}_1(\cdot)$ in terms of the arrival and departure of jobs, with the former corresponding to an increase and the latter a decrease of $\hat{X}_1(\cdot)$, jobs arrive according to a rate-α Poisson process. But, each arrival of a job for $\hat{X}_1(\cdot)$, at time t, corresponds to the departure of a job for $X_1(\cdot)$, at time $-t$. It follows that the departure process of jobs for $X_1(\cdot)$ is a Poisson rate-α process. Moreover, departures preceding any given time t_1 are independent of $X(t_1)$. What we are observing here, is that the specific nature of the Poisson *input* into station 1 results in an *output* of the same form.

Let $X_2(\cdot)$ denote the process for the number of jobs at station 2, and assume that $X_2(t_0)$ has the stationary distribution (2.1), with $m = m_2$, for a given t_0 and is independent of $X_1(t)$ for $t \geq t_0$. The arrival process of $X_2(\cdot)$ is also the departure process of $X_1(\cdot)$, which is a rate-α Poisson process, by the previous paragraph. It follows that $X_2(\cdot)$ is also the stationary process of an $M/M/1$ queue. Its arrivals, up to a given time t_1, with $t_1 \geq t_0$, are independent of $X_1(t_1)$ by the previous paragraph. Consequently, $X_1(t_1)$ and $X_2(t_1)$ are independent, each with the distribution in (2.1), with m_1 and m_2 replacing the mean m. Since t_1 was arbitrary, the joint process $(X_1(\cdot), X_2(\cdot))$ is Markov and stationary, for all $t \geq t_0$. Since t_0 was arbitrary, $(X_1(\cdot), X_2(\cdot))$ in fact defines a stationary Markov process over all t.

Continuing in this manner, one obtains a stationary Markov process $X(t) = (X_1(t), \ldots, X_K(t))$, $-\infty < t < \infty$, whose joint distribution at any time is a product of distributions of the form in (2.1). One therefore obtains the following result.

Theorem 2.10. *Assume that the interarrival time and the service times of the sequence of stations depicted in (2.42) are exponentially distributed, with $\alpha m_k < 1$ for each $k = 1, \ldots, K$. Then, the network has a stationary distribution π, which is given by*

$$\pi(n_1, \ldots, n_K) = \prod_{k=1}^{K} (1 - \alpha m_k)(\alpha m_k)^{n_k}, \tag{2.43}$$

for $n_k \in \mathbf{Z}_{+,0}$.

We note that although the components of the stationary distribution given by (2.43) are independent, this is not at all the case for the components $X_k(\cdot)$ of the corresponding stationary Markov process $X(\cdot)$. In particular, a departure at station k coincides with an arrival at station $k+1$.

Results leading up to Theorem 2.10 and the above proof are given in [Ja54], [Bu56], and [Re57]. More detail on the background of the problem is given on page 212 of [Ke79].

The same formula as in (2.43) holds when the routing in (2.42) is replaced by general routing, if the total arrival rates λ_k are substituted for α. More precisely, suppose that a subcritical Jackson network (i.e., a single

class network with exponentially distributed interarrival and service times) has external arrival rates $\alpha = \{\alpha_k,\ k = 1, \ldots, K\}$ and mean routing matrix $P = \{P_{k,\ell},\ k, \ell = 1, \ldots, K\}$. Then, it has the stationary distribution π, with

$$\pi(n_1, \ldots, n_K) = \prod_{k=1}^{K} (1 - \lambda_k m_k)(\lambda_k m_k)^{n_k}, \tag{2.44}$$

for $n_k \in \mathbf{Z}_{+,0}$. This result is no longer as easy to see as is (2.43); it was shown in the important work [Ja63]. The result will follow as a special case of Theorems 2.3 and 2.12.

Basics of quasi-reversibility

The "input equals output" behavior of the network in (2.42) was central to our ability to write the stationary distribution of the network as a product of the stationary distributions at its individual stations. *Quasi-reversibility* generalizes this concept, and leads to similar results for more general families of networks. Quasi-reversibility was first identified in [Mu72] and has been extensively employed in work by F.P. Kelly. The property can be defined in different equivalent ways; we use the following analytic formulation.

We consider a node for which arrivals at its classes $k = 1, \ldots, K$ are given by independent Poisson processes, with intensities $\alpha = \{\alpha_k,\ k = 1, \ldots, K\}$, that do not depend on the state of the node at earlier times, and for which the exits occur only one at a time and do not coincide with an arrival. The evolution of the node is assumed to be given by a Markov process $X(\cdot)$ with stationary distribution π defined on a countable state space S. Assume that all states communicate. Also, let q denote the transition function of $X(\cdot)$ and $\hat{q}$ the transition function of the reversed process $\hat{X}(\cdot)$ satisfying (2.11).

Under the above assumptions, any change in the state of the node due to a transition from x to y must be due to an increase by 1 in the number of jobs at some class k, for which we write $y \in A_k(x)$; a decrease by 1 at some k, for which we write $y \in E_k(x)$; or a transition that involves neither an increase nor a decrease, for which we write $y \in I(x)$, and which we refer to as an *internal transition*. Note that $y \in I(x)$ and $x \in I(y)$ are equivalent. An example of an internal transition is the "advance in stage" $y = \mathfrak{s}_i(x)$ in Section 2.4, although in the current setting far more general changes of state are allowed, including the simultaneous swapping of positions by many jobs. General changes of state are also allowed with the arrival or exit of a job.

We will say the node is *quasi-reversible* if for each class k and state x,

$$\sum_{y \in A_k(x)} \hat{q}(x, y) = \beta_k \tag{2.45}$$

for some $\beta_k \geq 0$. The equality (2.45) says that the rate of arrivals at each class k for the reversed process $\hat{X}(\cdot)$ does not depend on the state x. It is equivalent to the apparently stronger

$$\sum_{y \in A_k(x)} \hat{q}(x,y) = \sum_{y \in A_k(x)} q(x,y) = \alpha_k \tag{2.46}$$

for each k and x, which states that $\beta_k = \alpha_k$. (Note that the last equality follows automatically from the definition of α_k.)

To see (2.46), we note that by (2.45), the arrival times of $\hat{X}(\cdot)$ form independent rate-β_k Poisson processes at the K classes. The same reasoning that was applied to the sequence of $M/M/1$ queues in (2.42) implies that the exit times of $X(\cdot)$ also form independent rate-β_k Poisson processes. By assumption, only one exit occurs at each such time. Moreover, under the stationary distribution π, the rates at which jobs enter and leave a class are the same. Since the former is α_k, this implies $\beta_k = \alpha_k$, as needed for (2.46).

In the preceding argument, we have shown that the exit processes for $X(\cdot)$ form independent rate-α_k Poisson processes. Comparison with $\hat{X}(\cdot)$ also shows that exits for $X(\cdot)$ preceding any given time t_1 are independent of $X(t_1)$. These are important properties of quasi-reversible nodes. We have already employed them in the proof of Theorem 2.10.

The term quasi-reversible can also be applied to a queueing network rather than just to a node, with equation (2.45) again being employed as the defining property. (One should interpret $A_k(x)$ in terms of external arrivals at k.) In this setting, the stronger (2.46) need not hold, since departures from a class, that are not exits, may occur because of a job moving to another class within the network, and the reasoning in the paragraph below (2.46) is not valid. Nevertheless, external arrivals for the reversed process $\hat{X}(\cdot)$ correspond to jobs exiting the network for $X(\cdot)$. The same reasoning that was employed for quasi-reversible nodes therefore implies that the exiting processes at the classes k are independent rate-β_k Poisson processes.

Although we will not use this here, we also note that the *partial balance equations*

$$\pi(x) \sum_{y \in A_k(x)} q(x,y) = \sum_{y \in A_k(x)} \pi(y) q(y,x), \tag{2.47}$$

for each k and x, are equivalent to (2.46), and hence to the quasi-reversibility of a node. This follows immediately from the definition of $\hat{q}$ in (2.11) and the assumption $\pi(x) \neq 0$ for all x. These equations are weaker than the detailed balance equations, which correspond to reversibility, but include information not in the balance equations. The partial balance equations are often used as an alternative to quasi-reversibility.

Construction of networks from quasi-reversible nodes and applications

The main result on quasi-reversibility is Theorem 2.11, which states that when a queueing network satisfies certain conditions involving quasi-reversible nodes, its stationary distribution can be written as the product of the stationary distributions of these nodes. These nodes typically correspond to the stations of the network in a natural way. Such a queueing network is itself

quasi-reversible. Examples of these queueing networks are the sequence of $M/M/1$ queues in (2.42), Jackson networks, the homogeneous networks of Kelly type that were defined in Section 2.3, and the symmetric networks that were defined in Section 2.4.

We will consider queueing networks in the following framework. The network will consist of J stations and K classes on a countable state space S of the form

$$S = S^1 \times \ldots \times S^J,$$

where S^j is the state space corresponding to the j^{th} station. We will typically write $x = (x^1, \ldots, x^J)$ for $x \in S$, where $x^j \in S^j$. For concreteness, we will assume that for each j, S^j is one of the two spaces S_0 and S_e that were employed in the last two sections, although the theory holds more generally. As usual, the queueing network is assumed to have transition matrix $P = \{P_{k,\ell},\ k, \ell = 1, \ldots, K\}$ and external arrival rates $\alpha = \{\alpha_k,\ k = 1, \ldots, K\}$. Recall that $\lambda = Q\alpha$ denotes the total arrival rate, and satisfies the traffic equations given in (1.6).

We will employ notation similar to what was used earlier in the section, with $A_k(x), E_k(x)$, $I_j(x)$, and $R_{k,\ell}(x)$ denoting the states y obtained from x by the different types of transitions. As before, $A_k(x)$, $E_k(x)$, and $I_j(x)$ will denote the states obtained by an arrival into the network at k, an exit from the network at k, and an internal state change at j. For $y \in A_k(x)$ or $y \in E_k(x)$, we will require that $y^j = x^j$ for $j \neq s(k)$, and that the number of jobs at k increase or decrease by 1, and elsewhere remain the same. For $y \in I_j(x)$, we will require that $y^j = x^j$ for $j \neq s(k)$, and that the number of jobs at each class remain the same. We let $R_{k,\ell}(x)$ denote the set of y obtained from x by a job returning to class ℓ after being served at class k. We require that $y^j = x^j$ for $j \neq s(k)$ and $j \neq s(\ell)$, and that the number of jobs at k decrease by 1, at ℓ increase by 1, and elsewhere remain the same.

The queueing networks we will consider will be assumed to satisfy properties (2.48)-(2.52), which are given in terms of prechosen quasi-reversible nodes. Before listing these properties, we provide some motivation, recalling the sequence of 1-class stations given in (2.42), with the stationary distribution in (2.43). When a given station j, with $j = k$, is viewed "in isolation", it evolves as an $M/M/1$ queue with mean service time m_k and external arrival rate α, and has as its stationary distribution the stationary distribution of the corresponding $M/M/1$ queue. The stationary distribution of the sequence of stations is given by the product of the stationary distributions of the individual queues. Because of the specific structure of the network, Poisson arrivals into a given station result in Poisson departures, which then serve as Poisson arrivals for the next station. This property allowed us to view the stations "in isolation".

We will show that queueing networks satisfying properties (2.48)-(2.52) will have stationary distributions that are the product of the stationary distributions of the quasi-reversible nodes given there. Each such node can be

interpreted as the corresponding station evolving "in isolation". Here, "in isolation" will also mean that routing between classes at the same station is not permitted. The network in the previous paragraph, consisting of a sequence of 1-class stations, will be a special case of this more general setup.

Because of the more abstract setting now being considered, we will not explicitly follow the evolution of individual jobs; instead, we will think of the quasi-reversible nodes as "black boxes", which have a given output for a given input, with a corresponding stationary distribution. Rather than employ the Poisson-in, Poisson-out property directly, we will use the definition of quasi-reversibility in (2.45). The external arrival rates α_k^j for the classes at a given node j will be given by the total arrival rate λ_k for the corresponding class in the network. This will be consistent with jobs always leaving the node after being served, without being routed to another class.

The nodes we employ are assumed to have state spaces $S^j,\ j = 1, \ldots, J$, which are the components of the state space S for the network. Therefore, for $x = (x^1, \ldots, x^J) \in S$ with $x^j \in S^j$, $j = 1, \ldots, J$, one can also interpret x^j as the state of the corresponding node. Jobs at a given node will have classes $k \in \mathcal{C}(j)$, which are in one-to-one correspondence with the classes of the correspondingly labelled station in the network. For $x^j \in S^j$ and $k \in \mathcal{C}(j)$, we employ notation introduced earlier in the section for quasi-reversible nodes, with $A_k^j(x^j)$, $E_k^j(x^j)$, and $I^j(x^j)$ denoting those states y^j obtained from x^j by an arrival or exit at class k, or by an internal state change. We let $q^j,\ j = 1, \ldots, J$, denote the transition rates for the Markov processes $X^j(\cdot)$ of the nodes. The nodes are assumed to be quasi-reversible, with (2.45) being satisfied by $\hat{q}^j$, the transition rates of the reversed processes $\hat{X}^j(\cdot)$. As mentioned earlier, the external arrival rates of the nodes are given by $\alpha_k^j = \lambda_k$. As earlier in the section, we will assume that all states of a given node communicate.

We will assume that the transition rates $q(x, y)$ for the queueing network can be written in terms of the rates $q^j(x^j, y^j)$ for the nodes as follows. For $y \in A_k(x)$, we assume that

$$q(x, y) = (\alpha_k/\lambda_k) q^j(x^j, y^j). \tag{2.48}$$

Setting $p_k^j(x^j, y^j) = q^j(x^j, y^j)/\lambda_k$, this can be written as

$$q(x, y) = \alpha_k p_k^j(x^j, y^j), \tag{2.48$'$}$$

where $\sum_{y^j \in A_k^j(x^j)} p_k^j(x^j, y^j) = 1$ holds. (Here and later on, when j and k appear together, we implicitly assume that $k \in \mathcal{C}(j)$.) For $y \in E_k(x)$, we assume that

$$q(x, y) = q^j(x^j, y^j) P_{k,0}, \tag{2.49}$$

where $P_{k,0} \stackrel{\text{def}}{=} 1 - \sum_\ell P_{k,\ell}$. For $y \in R_{k,\ell}(x)$, we require the existence of an "intermediate" state z between x and y, with $z \in E_k(x)$ and $y \in A_\ell(z)$, and such that

$$q(x,y) = q^j(x^j, z^j)q^h(z^h, y^h)(P_{k,\ell}/\lambda_\ell), \tag{2.50}$$

where $h = s(\ell)$. One can also write this as

$$q(x,y) = q^j(x^j, z^j)P_{k,\ell}\, p^h_\ell(z^h, y^h). \tag{2.50'}$$

For $y \in I_j(x)$, we assume that

$$q(x,y) = q^j(x^j, y^j), \tag{2.51}$$

and finally, on the complement of the above sets, we assume that

$$q(x,y) = 0. \tag{2.52}$$

The equations (2.48)–(2.52) have the following interpretation in terms of the transition rates of the queueing network. The J different stations operate independently of one another, except for the movement of jobs between them. So, the transition rates in (2.48′), (2.49), and (2.51) depend on x^j and y^j instead of on the entire states x and y. In (2.50′), after a class k job is served, it moves to class ℓ with probability $P_{k,\ell}$, with the probability of the new state y depending on just z^h and y^h. When $h \neq j$, z is automatically given by $z^j = y^j$, $z^h = x^h$, and $z^{j'} = x^{j'} = y^{j'}$ for other values j'. The transition rates q^j in each display are those of node j, which does not permit returns. This node can be thought of as the one obtained from the corresponding station by replacing transitions to and from each class k by external arrivals and exits at the same rates.

We now employ the above terminology to state Theorem 2.11. When its hypotheses are satisfied, the theorem enables us to write the stationary distribution of a queueing network as the product of the stationary distributions of the corresponding nodes.

Theorem 2.11. *Suppose that the transition rates $q(x,y)$ of a queueing network satisfy (2.48)-(2.52), where the nodes with the transition rates $q^j(x^j, y^j)$ are quasi-reversible with stationary distributions π^j. Then, the queueing network has stationary distribution π given by*

$$\pi(x) = \prod_{j=1}^{J} \pi^j(x^j), \tag{2.53}$$

where $x = (x^1, \ldots, x^J)$. Moreover, the queueing network is itself quasi-reversible.

The proof of Theorem 2.11 will be given in the next subsection. We first note the following consequences of Theorem 2.11 and quasi-reversibility.

As an elementary illustration of Theorem 2.11, we return to the "sequence of $M/M/1$ queues" in (2.42). Equations (2.48)-(2.52) all hold in this setting, if q^j are the transition rates for the $M/M/1$ queues with $\alpha^j = \alpha$ and $m^j = m_j$.

All of these equations are easy to see and are nondegenerate in only a few cases. In (2.48), $q(x,y) \neq 0$ only for $k = j = 1$ and, in (2.49), $q(x,y) \neq 0$ only for $k = K = J$. In (2.50′), with $1 \leq k < K$, one has $P_{k,k+1} = p_{k+1}^{k+1}(z^{k+1}, y^{k+1}) = 1$ if z is chosen by removing a class k job from x; (2.51) is vacuous in this setting. Moreover, since the $M/M/1$ queues are reversible, they are quasi-reversible. These queues are assumed to be subcritical and have stationary distributions given by (2.1). Theorem 2.10 therefore follows as a special case of Theorem 2.11.

Equations (2.48)-(2.52) also hold for the more general Jackson networks (which are, in turn, special cases of FIFO networks of Kelly type). Again, q^j are the transition rates for $M/M/1$ queues, this time with $\alpha^j = \lambda_j$ and $m^j = m_j$. The equations are similar to those for the previous example, except that the mean transition matrix P is general, and so (2.48)-(2.50′) may be nonzero for arbitrary k. The formula for the stationary distribution in (2.44) is consequently an easy application of Theorem 2.11.

We now generalize Theorem 2.3 of Section 2.1. For homogeneous networks of Kelly type and symmetric networks, equations (2.48)-(2.52) all hold if the corresponding nodes are chosen in the natural way. Namely, each such node, for $j = 1, \ldots, J$, is obtained from the corresponding station by replacing transitions involving routing from one class to another by exits from the network, and by increasing the rate of external arrivals at each class k from α_k to λ_k to compensate for this. Then, (2.48) and (2.49) are immediate. In (2.50′), the state z is chosen by removing the served job at k from x. The equality (2.50′) then follows since a transition from x to y, with $y \in R_{k,\ell}(x)$, consists of a service completion at k, followed by the routing of the corresponding job to class l of a station h, with the job then being assigned a position i according to the rule p_ℓ^h. When $S^j = S_0$, the transition $q(x,y)$ in (2.51) does not occur; when, for a symmetric node, $S^j = S_e$, the transition corresponds to the advance of a stage. In either case, (2.51) is clear. Moreover, on account of (2.22) and (2.36) in Sections 2.3 and 2.4,

$$\sum_{y^j \in A_k^j(x^j)} \hat{q}^j(x^j, y^j) = \alpha_k \quad \text{for all } x^j \in S^j, \tag{2.54}$$

and so each such node is quasi-reversible. Note that this characterization continues to hold for networks that are of mixed type, with some stations being homogeneous of Kelly type and others being symmetric.

Recall that in Theorems 2.7 and 2.8, we saw that such nodes themselves have stationary distributions that are of product form, as given in (2.20) and (2.32). Theorem 2.7 was stated, for homogeneous nodes, in the context of service times that are exponentially distributed, and Theorem 2.8 was stated, for symmetric nodes, in the context of mixtures of Erlang distributions. Combining these results with Theorem 2.11, we therefore obtain Theorem 2.12. As in Theorems 2.7 and 2.8, when the node is homogeneous, $x^j \in S_0$ is assumed, whereas when the node is symmetric, $x^j \in S_e$. In either case, we employ the

condition

$$B_j \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} \left(\rho_j^n \Big/ \prod_{i=1}^{n} \phi(i) \right) < \infty, \tag{2.55}$$

with $\rho_j = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k$.

Theorem 2.12. *Suppose that each station j of a queueing network is either homogeneous of Kelly type or is symmetric, and satisfies (2.55). Then, the queueing network has a stationary distribution π that is given by*

$$\pi(x) = \prod_{j=1}^{J} \pi^j(x^j), \tag{2.56}$$

where each π^j is either of the form (2.20) or (2.32), depending on whether the station j is homogeneous of Kelly type or is symmetric, and α_k in these formulas is replaced by λ_k.

Theorem 2.11 shows that the stationary distribution of a queueing network that is composed of stations corresponding to quasi-reversible nodes has the product structure given in (2.53). Nevertheless, the processes that are associated with the stations are not independent. One can see this easily for the example at the beginning of the section consisting of a sequence of $M/M/1$ queues: a departure from one queue coincides with an arrival to the next. Similarly, the combined arrival processes at different classes (i.e., arrivals from other classes as well as external arrivals) are typically not independent, nor are the departure processes.

These processes are not to be confused with the processes of external arrivals or the processes of jobs exiting from the network, which are in either case independent. We note, though, that under the stationary distribution for a queueing network composed of quasi-reversible nodes, the conditional distribution at a station found by an arriving job is the same as the stationary distribution there. This is clearly the case for external arrivals, but is also true for arrivals in general. This is shown on page 70 of [Ke79] by considering the reversed Markov process for the network.

Demonstration of Theorem 2.11

In order to demonstrate Theorem 2.11, we will employ Proposition 2.6. We therefore need a candidate $\hat{q}$ for the transition rates of the reversed process $\hat{X}(\cdot)$ for the network under its stationary distribution. Letting $\hat{q}^j$ denote the reversed transition rates corresponding to q^j, we define $\hat{q}$ using the following analogs of (2.48)-(2.52). We set

$$\hat{q}(x,y) = (\hat{\alpha}_k/\lambda_k)\hat{q}^j(x^j,y^j) = P_{k,0}\hat{q}^j(x^j,y^j) \quad \text{for } y \in A_k(x), \tag{2.57}$$

$$\hat{q}(x,y) = \hat{q}^j(x^j,y^j)\hat{P}_{k,0} = \hat{q}^j(x^j,y^j)(\alpha_k/\lambda_k) \quad \text{for } y \in E_k(x). \tag{2.58}$$

For $y \in R_{k,\ell}(x)$, we set

$$\hat{q}(x,y) = \hat{q}^j(x^j,z^j)\hat{P}_{k,\ell}\frac{\hat{q}^h(z^h,y^h)}{\sum_{w^h\in A^h_\ell(z^h)}\hat{q}^h(z^h,w^h)} \tag{2.59}$$
$$= \hat{q}^j(x^j,z^j)\hat{q}^h(z^h,y^h)(P_{\ell,k}/\lambda_k).$$

We also set

$$\hat{q}(x,y) = \hat{q}^j(x^j,y^j) \quad \text{for } x \in I_j(y), \tag{2.60}$$

$$\hat{q}(x,y) = 0 \quad \text{for other values of } y. \tag{2.61}$$

Here, we are setting

$$\hat{\alpha}_k = \lambda_k P_{k,0}, \quad \hat{P}_{k,\ell} = \lambda_\ell P_{\ell,k}/\lambda_k, \quad \hat{P}_{k,0} = \alpha_k/\lambda_k. \tag{2.62}$$

The term $\lambda_k P_{k,0}$ is the rate at which jobs exit from the original network at class k, and so should be the rate they enter the reversed network at k. The second equality is obtained by reversing the direction of the mean transition matrix P; the third equality is obtained by setting $\hat{P}_{k,0} = 1 - \sum_\ell \hat{P}_{k,\ell}$, and applying the previous equality together with (1.6). We have implicitly set $\hat{\lambda}_k = \lambda_k$ in (2.57). The second equality in (2.59) needs to be justified; it follows from (2.62) together with

$$\sum_{w^h\in A^h_\ell(z^h)} \hat{q}^h(z^h,w^h) = \lambda_\ell. \tag{2.63}$$

Since each node is assumed to be quasi-reversible, (2.63) follows from (2.46) and $\alpha^h_\ell = \lambda_\ell$. In the proof of Theorem 2.11, we will also employ

$$\sum_k \hat{\alpha}_k = \sum_k \alpha_k, \tag{2.64}$$

which follows from the definition of $\hat{\alpha}_k$ and (1.6).

Proof of Theorem 2.11. The quasi-reversibility of the queueing network follows immediately from the first equality in (2.57) and from (2.63), since

$$\sum_{y\in A_k(x)} \hat{q}(x,y) = (\hat{\alpha}_k/\lambda_k) \sum_{y^j\in A^j_k(x^j)} \hat{q}^j(x^j,y^j) = (\hat{\alpha}_k/\lambda_k)\lambda_k = \hat{\alpha}_k,$$

which does not depend on x.

The remainder of the proof is devoted to showing that the distribution π in (2.53) is stationary. We wish to show that π satisfies

$$\pi(x)q(x,y) = \pi(y)\hat{q}(y,x) \quad \text{for all } x,y \in S \tag{2.65}$$

and

$$q(x) = \hat{q}(x) \quad \text{for all } x \in S, \tag{2.66}$$

where $\hat{q}$ is defined in (2.57)-(2.61). These are restatements of (2.11) and (2.18), and together with Proposition 2.6 imply that π is stationary. Since all states are assumed to communicate, this is the unique such distribution.

Demonstration of (2.65). In order to verify (2.65), one needs to check the different cases given by the formulas for q in (2.48)-(2.52). Each is straightforward, with the most involved case being $y \in R_{k,\ell}$. To check (2.65) for $y \in R_{k,\ell}(x)$, note that by (2.50) and the second part of (2.59), (2.65) reduces to

$$\pi^j(x^j)\pi^h(x^h)q^j(x^j,z^j)q^h(z^h,y^h) = \pi^j(y^j)\pi^h(y^h)\hat{q}^j(y^j,z^j)\hat{q}^h(z^h,x^h) \quad (2.67)$$

after cancelling the common terms $\pi^{j'}(x^{j'})$, with $j' \neq j$ and $j' \neq h$, and $P_{k,\ell}/\lambda_\ell$. This equality follows immediately from the definition of $\hat{q}^j$ and $\hat{q}^h$ in (2.11).

For the cases where $y \in A_k(x)$ and $y \in E_k(x)$, (2.65) reduces to analogs of (2.67), which are somewhat simpler since only one node rather than two is involved. The case $y \in I_j(x)$ follows from the definition of $\hat{q}^j$. For pairs x and y not covered in the preceding four cases, $q(x,y) = \hat{q}(y,x) = 0$ by (2.52) and (2.61). So, (2.65) holds in this last case as well.

Demonstration of (2.66). This part requires more work. We will show that

$$q(x) = \sum_k \alpha_k - \sum_k \lambda_k + \sum_j q^j(x^j) \quad (2.68)$$

and

$$\hat{q}(x) = \sum_k \hat{\alpha}_k - \sum_k \lambda_k + \sum_j \hat{q}^j(x^j). \quad (2.69)$$

The first sums in the two equalities are equal by (2.64), and the last sums are equal since (2.18) holds for each node. So, (2.68) and (2.69) together imply (2.66).

We first show (2.68). We rewrite $q(x)$ as

$$q(x) = \left(\sum_k \sum_{y \in A_k(x)} + \sum_k \sum_{y \in E_k(x)} + \sum_{k,\ell} \sum_{y \in R_{k,\ell}(x)} + \sum_j \sum_{y \in I_j(x)} \right) q(x,y), \quad (2.70)$$

and analyze the different parts. By (2.48′), the first double sum on the right equals

$$\sum_k \alpha_k \sum_{y^j \in A_k^j(x^j)} p_k^j(x^j,y^j) = \sum_k \alpha_k. \quad (2.71)$$

By (2.49), the second double sum equals

$$\sum_k \sum_{y^j \in E_k^j(x^j)} q(x^j,y^j) P_{k,0}. \quad (2.72)$$

By (2.50′), the third double sum equals

$$\sum_k \sum_{z^j \in E_k^j(x^j)} q^j(x^j, z^j) \sum_\ell P_{k,\ell} \sum_{y^h \in A_\ell^h(z^h)} p_\ell^h(z^h, y^h) \tag{2.73}$$
$$= \sum_k \sum_{z^j \in E_k^j(x^j)} q^j(x^j, z^j) \sum_\ell P_{k,\ell}.$$

By (2.51), the last double sum equals

$$\sum_j \sum_{y^j \in I^j(x^j)} q^j(x^j, y^j). \tag{2.74}$$

Summation of (2.72)-(2.74) gives

$$\left(\sum_k \sum_{y^j \in E_k^j(x^j)} + \sum_j \sum_{y^j \in I^j(x^j)} \right) q^j(x^j, y^j). \tag{2.75}$$

Also, note that

$$\sum_k \sum_{y^j \in A_k^j(x^j)} q^j(x^j, y^j) = \sum_k \alpha_k^j = \sum_k \lambda_k. \tag{2.76}$$

The sum of (2.75) and the left side of (2.76) is just $\sum_j q^j(x^j)$. On the other hand, the right side of (2.70) is equal to the sum of the left side of (2.71) and (2.75). So, $q(x)$ is equal to the sum of $\sum_k \alpha_k$ and (2.75), whereas $\sum_j q^j(x^j)$ is equal to the sum of $\sum_k \lambda_k$ and (2.75). Solving for this last term implies (2.68).

The argument for (2.69) is similar, and we employ the analog of the decomposition in (2.70), but for $\hat{q}(x)$ instead of $q(x)$. By the first equality in (2.57) and (2.63),

$$\sum_k \sum_{y \in A_k(x)} \hat{q}(x, y) = \sum_k (\hat{\alpha}_k / \lambda_k) \sum_{y^j \in A_k^j(x^j)} \hat{q}^j(x^j, y^j) = \sum_k \hat{\alpha}_k. \tag{2.77}$$

Also, using the first equalities in (2.58) and (2.59), and (2.60), the same reasoning as that leading to (2.75) implies that the sum of the terms corresponding to the last three double sums in (2.70) is

$$\left(\sum_k \sum_{y^j \in E_k^j(x^j)} + \sum_j \sum_{y^j \in I^j(x^j)} \right) \hat{q}^j(x^j, y^j). \tag{2.78}$$

On the other hand, it follows from (2.63) that

$$\sum_k \sum_{y^j \in A_k(x^j)} \hat{q}^j(x^j, y^j) = \sum_k \lambda_k. \tag{2.79}$$

The sum of (2.78) and the left side of (2.79) is just $\sum_j \hat{q}^j(x^j)$. Employing (2.77), (2.78), and (2.79) as we did (2.71), (2.75), and (2.76), the same reasoning as before implies (2.69). ■

Another proof for Theorem 2.11

Another proof for Theorem 2.11, that is more probabilistic, is given in [Wa82] and [Wa83] (see also [Wa88]). The basic idea of the proof is to modify the queueing network by imposing an ϵ delay, with $\epsilon > 0$, on all routing between classes. The corresponding stochastic process will be easier to analyze. It will not be Markov, but will have a distribution that is of product form and is invariant over time, and is the same for all values of ϵ. The limiting process as $\epsilon \downarrow 0$ will be the Markov process for the original queueing network, and its stationary distribution will be this distribution.

We now sketch the argument. Consider the J quasi-reversible nodes that are associated with the queueing network as in (2.48)-(2.52), but which have external arrival rates α_k instead of λ_k. We form a new network from these nodes by assuming that when jobs leave a node j from class k, they are routed back to class ℓ of node h with probability $P_{k,\ell}$, but with a fixed deterministic delay $\epsilon > 0$. During this delay, such jobs are assumed to not affect the transitions within the nodes, which now play the role of individual stations within the network.

One can construct the corresponding stochastic process inductively over time intervals of length ϵ, starting with $[0, \epsilon]$. One argues by first *assuming* that (a) the initial states at the J stations are independent of one another and are given by the stationary distributions of the isolated nodes with external arrival rates λ_ℓ and (b) over the time interval $(0, \epsilon]$, the jobs returning to classes ℓ constitute independent Poisson processes having rates $\lambda_k P_{k,\ell}$, which are independent of the initial states in (a). Jobs from outside the system arrive at class ℓ at rate α_ℓ, and so by (b) and the traffic equations (1.6), the combined arrivals at ℓ from these two sources of jobs are Poisson processes with rates λ_ℓ and are independent of one another. Because of the ϵ delay for returning jobs, jobs departing from nodes over $(0, \epsilon]$ will not return over this period, and so do not affect arrivals at ℓ.

On account of these arrival processes, the processes at the stations will be stationary over $(0, \epsilon]$ and independent of one another. Since the corresponding nodes are quasi-reversible, jobs depart from the classes k according to independent rate-λ_k Poisson processes over this period, which are independent of the states of the stations at time ϵ. Because of the deterministic ϵ delay required for returns, these jobs return to the classes ℓ as Poisson processes with rates $\lambda_k P_{k,\ell}$, over the period $(\epsilon, 2\epsilon]$. Consequently, the analogs of conditions (a) and (b) hold over the time interval $[\epsilon, 2\epsilon]$.

Iteration over the time intervals $(\epsilon, 2\epsilon]$, $(2\epsilon, 3\epsilon], \ldots$ produces a stochastic process on $[0, \infty)$ whose states at different stations at any fixed time are independent of one another, and whose distributions are the same as the

stationary distributions of the corresponding isolated nodes. This process can also be extended to all times $t \in (-\infty, \infty)$.

Letting $\epsilon \downarrow 0$, the sequence of these processes will converge to the Markov process corresponding to the original queueing network. Since each of these processes has the same joint distribution at any given time, this distribution will be stationary for the limiting Markov process. Since this distribution has the desired product form, this reasoning implies (2.53) of Theorem 2.11. By considering the exit processes of the sequence of processes, one can also show that the original queueing network is quasi-reversible.