Queueing models with multiple waiting lines

I.J.B.F. Adan, O.J. Boxma[1], J.A.C. Resing

Department of Mathematics and Computing Science,
Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

This paper discusses analytic solution methods for queueing models with multiple waiting lines. The methods are briefly illustrated, using key models like the 2 × 2 switch, the shortest queue and the cyclic polling system.

AMS subject classification: 60K25, 90B22.
Keywords and phrases: multiple waiting lines, analytic methods.

1 Introduction

This paper reviews analytic solution methods for queueing models with multiple waiting lines. The models under consideration represent isolated service centers; we do not consider networks of queues. We basically distinguish between two classes of models.
Class I: Customers arrive at a service center with several queues, each with one server. The customers choose a server according to some mechanism (e.g., shortest queue or shortest workload) or divide their work among several servers (e.g., fork-join).
Class II: Customers of several types arrive at a service center with one or more servers. The server or servers choose a customer for service according to some static (e.g., preemptive resume) or dynamic (e.g., polling) priority rule.
While not all queueing models with multiple waiting lines are naturally classified as belonging to Class I (passive servers; customers choose a queue) or Class II (active servers; servers choose a customer), this classification does help to bring some structure in the large family of queueing models with multiple waiting lines.

Our aim in this paper is to give the reader insight into which methods are available for the analysis of queueing models with multiple waiting lines. These queueing models often give rise to Markov processes with an N-dimensional state space that is the set of lattice points with integer-valued, nonnegative coordinates. The methods for solving the equilibrium equations of such Markov processes may also be divided into two groups.
Complex-function methods: These methods aim at solving the functional equation for the generating function of the equilibrium distribution.

[1] also: CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

Direct methods: Their aim is to solve directly (i.e., without resorting to transforms) the equilibrium equations.
We emphasize the main ideas behind the methods, and their strengths and limitations. We also point out some relations between the various methods. While our orientation is methodological, we do pay much attention to a few particular queueing models. These are relatively simple but important models, that are used as vehicle to illustrate and compare various methods.

We already mentioned that not all queueing models with multiple waiting lines will fit into our classification. One of these are the interesting queueing models with simultaneous resource possession. For these models an important new product-form result has been established by Berezner et al. [16] which shows that at least in some simple situations they have an exact explicit solution.

The paper is organized in the following way. In Section 2 we discuss three direct methods (the compensation method, the precedence-relation method and the power-series algorithm) and two methods from complex-function theory (the boundary value method and the uniformization method). They are discussed for two basic models of Class I: the 2×2 switch and the shortest queue. In Section 3 we consider models of Class II. Some of the above-mentioned methods also apply to those models; we also discuss some other methods from complex-function theory. Most analytically tractable queueing models are special examples of particular, often two-dimensional, Markov processes or random walks. Section 4 considers a few classes of two-dimensional random walks that allow an exact analysis.

2 Customers choose a queue

2.1 Introduction

In this section we consider two basic queueing models where arriving customers choose a waiting line, viz. the 2×2 switch and the shortest queue model. We expose and compare several techniques that have been used to analyse these models. We end the section with a brief discussion of the fork-join queue, a queueing model in which each job consists of various subjobs who each choose a different queue.

2.2 The 2×2 switch

Among the queueing models with multiple waiting lines, the 2×2 switch may be the simplest non-trivial model. It is therefore most suitable for exposing and comparing various analytic solution techniques. In the present subsection we shall discuss three such techniques, viz.: (i) the compensation method, (ii) the boundary value method, and (iii) the uniformization method. But first we describe the model.

Switches are important elements of communication networks. A 2×2 clocked buffered switch is a switch with 2 input and 2 output ports. Such a switch is modeled as a discrete-time queueing system with 2 parallel servers and 2 types of arriving jobs (see Figure 1). The number of arriving jobs of type i in a time unit (= clock cycle) is one with probability r_i and zero with probability $1 - r_i$, $i = 1, 2$. Jobs always arrive at the beginning of a time unit and once a job of type i has arrived, it joins the queue at server j with probability $t_{i,j}$, $t_{i,j} > 0$ for $j = 1, 2$ and

$t_{i,1} + t_{i,2} = 1$. Jobs that have arrived at the beginning of a time unit are immediately candidates for service. A server serves exactly one job per time unit, if at least one is present. For each server, the average number of arriving jobs per time unit is assumed to be less than one, i.e.

$$r_1 t_{1,j} + r_2 t_{2,j} < 1, \;\; j = 1, 2. \tag{2.1}$$

This assumption guarantees the ergodicity of the system. By (2.1), the case $r_1 = r_2 = 1$ must be excluded.

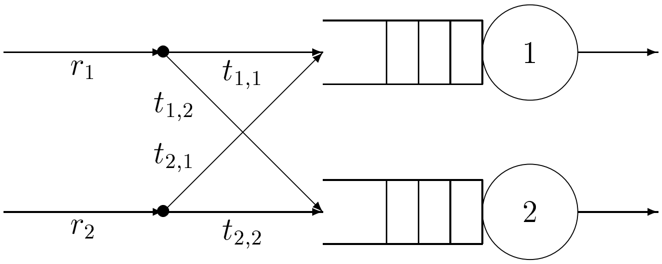

Figure 1: The 2 × 2 switch.

We are interested in the two-dimensional queue-length process, which is a Markov chain. The service discipline is irrelevant; we may assume it to be First-Come-First-Served. The Markov chain has states (m, n), where m and n denote the numbers of waiting jobs at server 1 and server 2 at the beginning of a time unit. For a state (m, n) in the interior of the state space, we only have one-step transitions to the neighbouring states $(m + m', n + n')$ with $m', n' \in \{-1, 0, 1\}$ and $m' + n' \leq 0$. The corresponding one-step transition probabilities $q_{m',n'}$ are equal to:

$$\begin{array}{ll} q_{1,-1} = r_1 r_2 t_{1,1} t_{2,1}, & q_{0,0} = r_1 r_2 (t_{1,1} t_{2,2} + t_{1,2} t_{2,1}), \\ q_{0,-1} = r_1(1 - r_2) t_{1,1} + r_2(1 - r_1) t_{2,1}, & q_{-1,1} = r_1 r_2 t_{1,2} t_{2,2}, \\ q_{-1,0} = r_1(1 - r_2) t_{1,2} + r_2(1 - r_1) t_{2,2}, & q_{-1,-1} = (1 - r_1)(1 - r_2). \end{array}$$

Each transition probability for the states at the boundaries can be written as a sum of the probabilities $q_{m',n'}$. In Figure 2 all one-step transition probabilities, except the ones for the transitions from a state to itself, are illustrated.

The bivariate queue-length distribution $\{p_{m,n}\}$ is characterized as the unique non-negative normalized solution of the equilibrium equations:

$$\begin{aligned} q p_{m,n} &= q_{1,-1} p_{m-1,n+1} + q_{-1,1} p_{m+1,n-1} + q_{0,-1} p_{m,n+1} \\ &+ q_{-1,0} p_{m+1,n} + q_{-1,-1} p_{m+1,n+1}, \quad m > 0, n > 0, \end{aligned} \tag{2.2}$$

$$\begin{aligned} (q - q_{0,-1}) p_{m,0} &= q_{1,-1} p_{m-1,1} + q_{1,-1} p_{m-1,0} + q_{0,-1} p_{m,1} \\ &+ (q_{-1,0} + q_{-1,-1}) p_{m+1,0} + q_{-1,-1} p_{m+1,1}, \quad m > 0, n = 0, \end{aligned} \tag{2.3}$$

$$\begin{aligned} (q - q_{-1,0}) p_{0,n} &= q_{-1,1} p_{1,n-1} + q_{-1,1} p_{0,n-1} + q_{-1,0} p_{1,n} \\ &+ (q_{0,-1} + q_{-1,-1}) p_{0,n+1} + q_{-1,-1} p_{1,n+1}, \quad m = 0, n > 0, \end{aligned} \tag{2.4}$$

$$\begin{aligned} (q_{1,-1} + q_{-1,1}) p_{0,0} &= (q_{-1,0} + q_{-1,-1}) p_{1,0} + (q_{0,-1} + q_{-1,-1}) p_{0,1} \\ &+ q_{-1,-1} p_{1,1}, \quad m = 0, n = 0, \end{aligned} \tag{2.5}$$

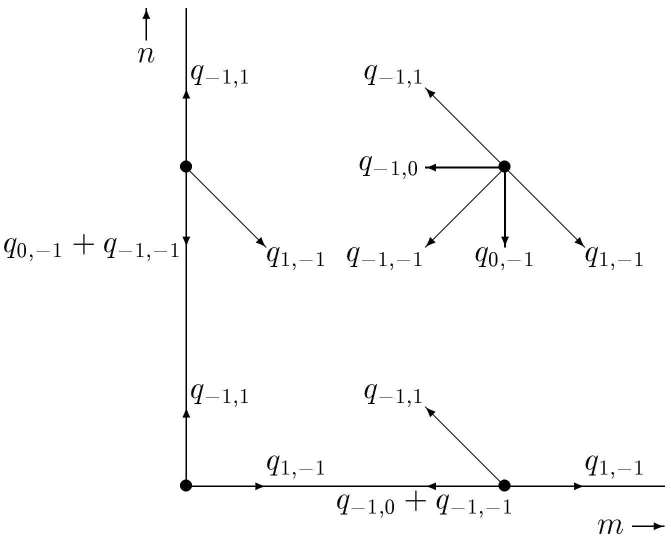

Figure 2: The one-step transition probabilities.

where

$$q := q_{1,-1} + q_{-1,1} + q_{0,-1} + q_{-1,0} + q_{-1,-1}. \tag{2.6}$$

We now describe three methods to solve these equations.

Method 1: The compensation method
This method yields an explicit expression, without transforms, for the equilibrium distribution. It is not only successful for this model, but in [9] it has been shown to work for the class of two-dimensional nearest-neighbor random walks (i.e., with unit steps only) *for which in the interior of the state space no one-step transitions are allowed to the North, North-East and East*, and also for some special cases not contained in this class (see [8, 6]). Clearly, the present random walk is an element in this class (see Figure 2). It is a special one, because the transition probabilities on the boundaries are 'projections' of the transition probabilities in the interior states. This property leads to several simplifications in the analysis. But the essential characteristics in the analysis are still maintained and therefore the problem is most suitable to demonstrate the compensation method.

The compensation method attempts to solve the equilibrium equations by a linear combination of product forms. This is achieved by first characterizing a sufficiently rich basis of product-form solutions satisfying the equilibrium equations in the interior of the state space. Subsequently this basis is used to construct a linear combination that *also* satisfies the equations for the boundary states. A similar method is of course well known in the theory of differential and difference equations (cf. the separation of variables method [73]) where the basis usually contains countably many elements, all of which are used to construct the solution. In the present model, however, the basis contains uncountably many elements. Therefore a procedure is needed to select the appropriate elements. This procedure is based on a compensation argument (which explains the name of the method): after introducing the first term, countably many terms are subsequently added so as to alternatingly compensate for the error on one of the two boundaries. The main steps in the analysis will be briefly outlined below (see [28] for more details on the 2 × 2 switch case).

Step 1: Characterize the set of product forms $\alpha^m \beta^n$ satisfying the equilibrium equations in the interior of the state space, i.e., the equations (2.2). Substitution of the product form into (2.2) and division by common powers yields a quadratic equation in α and β:

$$q\alpha\beta \;=\; q_{1,-1}\beta^2 + q_{-1,1}\alpha^2 + q_{0,-1}\alpha\beta^2 + q_{-1,0}\alpha^2\beta + q_{-1,-1}\alpha^2\beta^2 \,. \tag{2.7}$$

Because later in the analysis the solution has to be normalized, the factors α and β are required to satisfy $0 < |\alpha| < 1$ and $0 < |\beta| < 1$. The points on the curve (2.7) inside this region characterize a continuum of product forms satisfying the inner equations (see Figure 3).

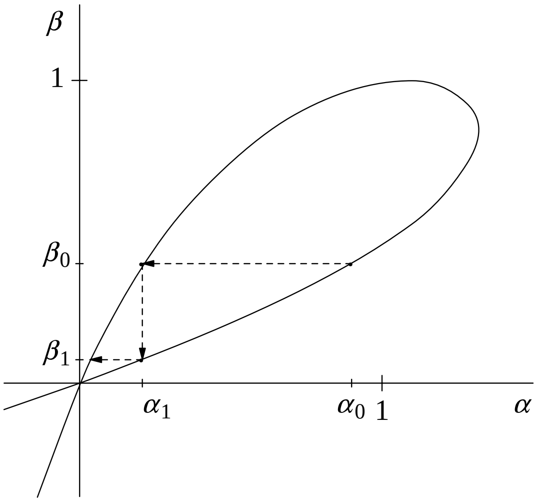

Figure 3: Curve (2.7) in the positive quadrant, generating the sequence $\alpha_0, \beta_0, \alpha_1, \beta_1, \ldots$

Step 2: Construct a linear combination of elements in this rich basis, which is a *formal* solution to the equilibrium equations. Here the word formal is used to indicate that (at this stage) we do not bother about the convergence of the solution. This aspect will be taken care of later. The construction of a linear combination starts with a *suitable initial term*. This is a basis solution $\alpha_0^m \beta_0^n$ that also satisfies the equilibrium equations on one of the two boundaries, i.e., the equations (2.3) or (2.4). Later we will explain why a special starting solution is needed. In the general case in [9] one finds at least one and at most four initial terms. Here there exist two initial terms: one for the horizontal boundary and another for the vertical boundary. Both terms can be explicitly calculated. The first one is given by

$$(\alpha_0, \beta_0) = (\frac{q_{1,-1}}{q_{-1,1} + q_{-1,0} + q_{-1,-1}}, \frac{q_{-1,1}\alpha_0^2}{q_{1,-1} + q_{0,-1}\alpha_0 + q_{-1,-1}\alpha_0^2}) \,; \tag{2.8}$$

the term $(\tilde{\alpha}_0, \tilde{\beta}_0)$ for the vertical boundary is symmetrical. Let us consider the initial term $c_0\alpha_0^m \beta_0^n$ for the horizontal boundary, where c_0 is a (arbitrary) nonnull constant. This term violates the equilibrium equations (2.4) on the vertical boundary. To compensate for this error we add a term $d_1\alpha^m\beta^n$ with α and β satisfying (2.7) such that the sum of the two terms satisfies the equilibrium equations in all states on the vertical boundary. This immediately implies that we are forced to take $\beta = \beta_0$ and thus α has to be the companion root of α_0 of Equation (2.7) with $\beta = \beta_0$. This companion root is denoted by α_1. Substitution of the sum of the two terms into (2.3) and division by common powers then yields a single equation for d_1, the solution of which is given by

$$d_1 = -\frac{1-\alpha_1}{1-\alpha_0}\, c_0. \tag{2.9}$$

The sum $c_0\alpha_0^m\beta_0^n + d_1\alpha_1^m\beta_0^n$ satisfies the equilibrium equations for the interior and the vertical boundary. However, the new term $d_1\alpha_1^m\beta_0^n$ violates the horizontal boundary. To compensate for this error we add again a term $c_1\alpha^m\beta^n$. This step is symmetrical to the one for the vertical boundary: we now have to take $\alpha = \alpha_1$ and $\beta = \beta_1$ where β_1 is the companion root of β_0 of Equation (2.7) with $\alpha = \alpha_1$. For the coefficient c_1 we obtain, similar to (2.9),

$$c_1 = -\frac{1-\beta_1}{1-\beta_0}\, d_1. \tag{2.10}$$

But the third term violates the vertical boundary conditions, so we add again a compensation term and so on. We keep on adding terms so as to alternatingly satisfy the two boundary conditions. This results in the following infinite sum of terms:

$$x_{m,n} = \underbrace{\overbrace{c_0\alpha_0^m\beta_0^n}^{H} + \rlap{$\overbrace{}^{H}$} d_1\alpha_1^m\beta_0^n}_{V} + \underbrace{c_1\alpha_1^m\beta_1^n + \rlap{$\overbrace{}^{H}$} d_2\alpha_2^m\beta_1^n}_{V} + c_2\alpha_2^m\beta_2^n + \cdots \tag{2.11}$$

The construction is such that each term in (2.11) satisfies the equilibrium equations in the interior of the state space, the sum of two terms with the same α satisfies the horizontal boundary conditions (H) and the sum of two terms with the same β satisfies the vertical boundary conditions (V). Hence $x_{m,n}$ satisfies all equilibrium equations. The coefficients c_i and d_i in the sum (2.11) satisfy (cf. (2.9)–(2.10))

$$d_{i+1} = -\frac{1-\alpha_{i+1}}{1-\alpha_i}\, c_i, \qquad c_{i+1} = -\frac{1-\beta_{i+1}}{1-\beta_i}\, d_{i+1}, \qquad i = 0, 1, 2, \ldots,$$

from which we immediately obtain

$$d_{i+1} = -\frac{(1-\alpha_{i+1})(1-\beta_i)}{(1-\alpha_0)(1-\beta_0)}\, c_0, \qquad c_{i+1} = \frac{(1-\beta_{i+1})(1-\alpha_{i+1})}{(1-\beta_0)(1-\alpha_0)}\, c_0.$$

Hence, by choosing $c_0 = (1-\alpha_0)(1-\beta_0)$, we get simple expressions for the coefficients c_i and d_i, resulting in the following elegant expression for $x_{m,n}$ as an alternating sum of two-dimensional geometric distributions:

$$x_{m,n} = \sum_{i=0}^{\infty} (1-\beta_i)\beta_i^n[(1-\alpha_i)\alpha_i^m - (1-\alpha_{i+1})\alpha_{i+1}^m]. \tag{2.12}$$

This completes the construction of the formal solution $x_{m,n}$. In the same way we can construct a formal solution $\tilde{x}_{m,n}$ by starting with the basis solution $(\tilde{\alpha}_0, \tilde{\beta}_0)$ satisfying the vertical boundary conditions. In (2.12) above we then have to replace α_i by $\tilde{\alpha}_i$ and β_i by $\tilde{\beta}_i$; in the completely symmetric case these equal β_i and α_i, respectively.

Step 3: Prove that the two formal solutions converge. In the general case in [9] it is shown that the formal solutions absolutely converge in all states, except in a (possibly empty) neighborhood of the origin. It is important to note that the requirement of convergence of the formal solutions is responsible for the exclusion of one-step transitions to the North, North-East and East and the need of suitable initial terms. In the present model $x(m,n)$ and $\tilde{x}(m,n)$ converge everywhere, except in $(0,0)$. Hence they satisfy the equilibrium equations in all states, except in $(0,0)$, $(1,0)$ and $(0,1)$.

Step 4: Determine the equilibrium probabilities by taking the linear combination $p(m,n) = cx(m,n) + \tilde{c}\tilde{x}(m,n)$ in all states, except in the origin. The unknowns $p(0,0)$, c and $\tilde{c}$ are then determined from the normalization equation and the equilibrium equations in $(1,0)$ and $(0,1)$. Here we may omit the one in $(0,0)$, because the equilibrium equations are dependent. In the present model the coefficients can be solved explicitly, namely $c = \tilde{c} = 1$. So the equilibrium probabilities can be simply expressed as:

$$p_{m,n} = x_{m,n} + \tilde{x}_{m,n}, \qquad m \geq 0, n \geq 0, m+n > 0. \tag{2.13}$$

Clearly, from this result we can derive similar expressions for performance characteristics such as, e.g., the mean queue lengths and the correlation between the queue lengths.

The compensation method clearly has its limitations. The most important one is that transitions to the North, North-East and East are forbidden in the case of two-dimensional nearest-neighbor random walks. The conditions become even more severe for the extension to higher dimensional random walks (see [81]). The strong feature of the method, however, is that it helps in finding such conditions for getting explicit solutions and that it provides constructive methods for obtaining the explicit solutions in case the conditions are satisfied.

Method 2: The boundary value method
Complex-function methods aim at solving the equilibrium equations by introducing the generating function of the equilibrium distribution and studying the functional equations that it should satisfy. Usually, those functional equations present formidable difficulties. For a class of *two-dimensional* random walks and queueing problems, however, techniques have been developed which reduce those functional equations to standard Riemann(-Hilbert) boundary value problems and to singular integral equations for complex-valued functions. The general theory is exposed in the monograph [47] and surveyed in [38]. Jaffe [87] has applied this 'boundary value method' to the (symmetric) 2×2 clocked buffered switch. Below we outline his approach; later in this section we point out some differences and similarities with the compensation method. In the symmetric case we write $r_1 = r_2 = p$, $0 < p < 1$, while $t_{1,1} = t_{1,2} = t_{2,1} = t_{2,2} = \frac{1}{2}$.
Step 1: The set-up. We introduce the generating function

$$f(x,y) := \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} p_{m,n} x^m y^n, \ |x| \leq 1, |y| \leq 1. \tag{2.14}$$

It follows from (2.2)-(2.5) that $f(x,y)$ satisfies the following functional equation: For $|x| \leq 1$,

$|y| \leq 1$,

$$(xy - r(x,y))f(x,y) = (y-1)r(x,0)f(x,0) + (x-1)r(0,y)f(0,y) + (x-1)(y-1)r(0,0)f(0,0), \tag{2.15}$$

where

$$r(x,y) := (1 - p + \frac{p}{2}(x+y))^2. \tag{2.16}$$

The distribution $\{p_{m,n}\}$ is completely determined by (2.15), the normalizing condition $f(1,1) = 1$, and the fact that the probability generating function $f(x,y)$ should be regular for $|x| < 1$, continuous for $|x| \leq 1$ for all $|y| \leq 1$, and similarly with x and y interchanged. In the sequel, we denote the interior of these unit circles by D, and their closure by $\bar{D}$.

Step 2: Analysis of 'the kernel'. The boundedness of $f(x,y)$ implies that the right-hand side of (2.15) should be zero for all those (x,y) on the complex curve $S = \{(x,y) : xy - r(x,y) = 0\}$, that lie in $\bar{D}^2$. These zeros of the *kernel* $xy - r(x,y)$ offer a large choice. It suffices to choose an appropriate set; by analytic continuation, other zero-tuples of the kernel may be constructed. In this symmetric case, it is natural to consider those (x,y) on S with $y = \bar{x}$. This is an *ellipse* $E = \{x : |x|^2 = r(x,\bar{x}),\ x \in \bar{D}\}$.

Step 3: Formulation of a boundary value problem. For the present example this proceeds as follows. Introduce for $|x| \leq 1$, $|x| \neq 1$,

$$g(x) := \frac{r(x,0)f(x,0)}{x-1} + \frac{1}{2}r(0,0)f(0,0) = \frac{r(0,x)f(0,x)}{x-1} + \frac{1}{2}r(0,0)f(0,0). \tag{2.17}$$

Then the boundedness of $f(x,y)$ in $\bar{D}^2$, combined with the fact that $g(\cdot)$ has a simple pole at 1, implies that

$$g(x) + g(\bar{x}) = 2\mathrm{Re}\ g(x) = 0,\ x \in E\backslash\{1\}, \tag{2.18}$$

$$\lim_{x\to 1}(x-1)g(x) = 1 - p. \tag{2.19}$$

We now have a *boundary value problem*: Determine a function $g(x)$ that is analytic inside the ellipse E, that has a simple pole at 1, and that satisfies (2.18) on the ellipse (on the *boundary*). Let ϕ, with inverse ψ, be the conformal mapping of the unit disk onto the region bounded by E, with normalization conditions $\phi(0) = p/(1+p)$, $\phi(1) = 1$. Define $h(w) := g(\phi(w))$. We then obtain a relatively simple 'Riemann-Hilbert boundary value problem with a pole', cf. Section I.3.3 of [47], for $h(.)$ on the unit *circle* D (actually, it is a Dirichlet problem with a pole):

$$\mathrm{Re}\, h(w) = 0,\ w \in D\backslash\{1\}, \tag{2.20}$$

$$\lim_{w\to 1}(w-1)h(w) = \frac{1-p}{\phi'(1)}, \tag{2.21}$$

with $h(.)$ analytic on D, continuous on $\bar{D}\backslash\{1\}$. The solution of this boundary value problem is

$$h(w) = \frac{1}{2}\frac{1-p}{\phi'(1)}\frac{w+1}{w-1},\ w \in D, \tag{2.22}$$

which determines $g(x) = h(\psi(x)) = \frac{1}{2}\frac{1-p}{\phi'(1)}\frac{\psi(x)+1}{\psi(x)-1}$ inside the ellipse E; the conformal mapping $\psi(x)$ is explicitly expressed in the Jacobi elliptic (sin am or sn) function. Substitution in (2.15) and analytic continuation finally yields $f(x, y)$, for $|x| \leq 1$, $|y| \leq 1$:

$$f(x, y) = (1-p)\psi'(1)\frac{(x-1)(y-1)}{(\psi(x)-1)(\psi(y)-1)}\frac{\psi(x)\psi(y)-1}{xy-r(x, y)}. \tag{2.23}$$

Remark 2.1 The boundary value method may also be applied to the *asymmetric* 2 × 2 clocked buffered switch. In fact, it can even be applied to a much more general class of two-dimensional random walks and queueing models; see Subsection 4.3.

Method 3: The uniformization method
The 'uniformization method' is a complex-function method that has been applied by Kingman [91] and Flatto and McKean [68] to the symmetric shortest queue, and by Jaffe [87] to the symmetric 2 × 2 clocked buffered switch. It has been applied by Cohen to both the symmetric [43] and asymmetric [44] 2 × 2 clocked buffered switch and for the symmetric [42] and asymmetric [45] shortest queue. The global idea of the uniformization method is as follows. Firstly, as in the above-described boundary value method, the generating function $f(x, y)$ of the two-dimensional equilibrium distribution is introduced, and the equilibrium equations for the distribution lead to a functional equation for $f(x, y)$. The relatively simple form of the kernel $K(x, y)$ (=$xy - r(x, y)$ in the case of the 2 × 2 switch) allows the following approach. $K(x, y) = 0$ defines an algebraic curve, which can be uniformized in a convenient way. Introduce a uniformizing variable p, writing $x = x(p)$ and $y = y(p)$. Consider $f(x(p), 0)$ and $f(0, y(p))$. They are shown to be analytic functions of p in certain p-regions; in addition, for all p with $|x(p)| \leq 1$, $|y(p)| \leq 1$, for which $K(x(p), y(p)) = 0$, the boundedness of $f(x(p), y(p))$ results in a relation between $f(x(p), 0)$ and $f(0, y(p))$. This relation is used to continue $f(x(p), 0)$ and $f(0, y(p))$ meromorphically into the plane $0 < |p| < \infty$. For the 2×2 switch as well as the shortest queue, the generating functions $f(x, 0)$ and $f(0, y)$ turn out to be *meromorphic*, i.e., all their singularities are isolated poles. Starting with a certain pole, one may find another pole that is linked to it via the fact that together they make the kernel zero, etc., ad infinitum (in much the same way as, in the compensation method, each boundary condition gives rise to yet another product-form term in an infinite sum; the reader will realize that taking generating functions of $x_{m,n}$ in (2.12) results in terms $\frac{1}{1-\alpha_i x}\frac{1}{1-\beta_i y}$, with poles $x = 1/\alpha_i$, $y = 1/\beta_i$).

We discuss the determination of the poles, and also the zeros, in some more detail for the case of the symmetric 2 × 2 clocked buffered switch.

Step 1: The poles. As observed above, for all those zero-tuples (x, y) of the kernel $xy - r(x, y)$ that lie inside $\bar{D}^2$, we should have

$$g(x) + g(y) = \frac{r(x, 0) f(x, 0)}{x-1} + \frac{r(0, y) f(0, y)}{1-y} + r(0, 0) f(0, 0) = 0. \tag{2.24}$$

Since $xy - r(x, y)$ is a biquadratic form in x and y, it is easily seen that it has, for each fixed x, two zeros $y_1(x)$ and $y_2(x)$ with the property that $|y_1(x)| < |x| < |y_2(x)|$ for $|x| \geq 1$, $x \neq 1$; analogously for $x_1(y)$ and $x_2(y)$. Cohen [43] equates $-r(x, 0) f(x, 0)/(x-1)$ to

$r(0, y_2(x)) f(0, y_2(x))/(1-y_2(x))+r(0,0) f(0,0)$, via (2.24), and uses this to continue $f(x,0)$ analytically outside the unit circle; similarly for $f(0,y)$. The only points where these functions are *not* analytic are the simple poles y_n^+ respectively x_n^+, $n = 1, 2, \dots$. With $x_0^+ = y_0^+ := 1$ (remember that $g(x)$ has a simple pole in $x = 1$!), we have $y_1^+ := y_2(x_0^+) = (2-p)^2/p^2$, $y_n^+ := y_2(x_{n-1}^+) = x_n^+ = x_2(y_{n-1}^+)$, $n = 1, 2, \dots$. Note that $y_n^+ = x_n^+$ due to the complete symmetry of the model under consideration.

Step 2: The zeros. After having thus shown that $f(x,0)$ has a unique analytic continuation in $|x| \geq 1$, except for $x = x_n^+$, $n = 1, 2, \dots$, where it has simple poles, Cohen [43] also determines all *zeros* of $f(x,0)$ (this is a rare case in which all zeros can be determined explicitly). He observes that $r(x,0)$ has only one zero $x = x_0^- = 2 - 2/a$, and this zero has multiplicity two. Introducing $y_1^- := y_2(x_0^-)$, $x_n^- = x_2(y_{n-1}^-) = y_n^- = y_2(x_{n-1}^-)$, $n = 1, 2, \dots$, it follows from $g(x) = -g(y_2(x))$ and $g(y) = -g(x_2(y))$ that $f(x,0)$ (and, symmetrically, $f(0,x)$) has zeros of multiplicity two at $x = x_{2n}^-$, $n = 1, 2, \dots$ (and $\frac{r(x,0) f(x,0)}{1-x} + r(0,0) f(0,0)$ has zeros of multiplicity two at $x = x_{2n-1}^-$, $n = 1, 2, \dots$). The zeros and poles do not have a finite accumulation point. It is shown that $f(x,0)$ is a meromorphic function, that is given by:

$$f(x,0) = \frac{1-p}{(1-p/2)^2} \frac{\prod_{n=1}^{\infty}(1-\frac{1}{x_n^+}) \prod_{n=1}^{\infty}(1-\frac{x}{x_{2n}^-})^2}{\prod_{n=1}^{\infty}(1-\frac{x}{x_n^+}) \prod_{n=1}^{\infty}(1-\frac{1}{x_{2n}^-})^2}. \tag{2.25}$$

Finally, an expression for $f(x,y)$ follows from (2.15). Cohen [43] shows that it is the unique solution.

Some relations between the above-mentioned methods are indicated in [28]. Below we summarize them. From an analytic point of view, both function-theoretic methods are for the present model of similar complexity (compared with the shortest queue problem and similar two-dimensional problems, one might say: of similar simplicity). They lead to different representations of the two-dimensional queue-length generating function. From a numerical point of view these representations can be exploited to obtain, e.g., queue length moments; the explicit representations in (2.25) and the one obtained by the compensation method are very suitable for numerical calculations.

It is interesting to compare the compact expression (2.23) for $f(x,y)$ at $y = 0$, as supplied by the boundary value method, with the infinite-product expression in (2.25). In the boundary value method, $g(x)$ has poles at the zeros of $\psi(x) - 1 = 0$. The normalization condition $\phi(1) = 1$ for the conformal mapping implies $\psi(1) = 1$, so that $x_0^+ = 1$ is again found to be a pole of $g(\cdot)$. The periodic nature of the Jacobian elliptic function $\psi(.)$ subsequently leads to (again) the sequence of poles $x_1^+, x_2^+, \dots$.

Let us also compare (2.25) with the expressions (2.12) and (2.13) as obtained by the compensation method. In the completely symmetric case, $\tilde{x}_{m,n}$ is obtained from $x_{m,n}$ by interchanging α_i and β_i. Taking its generating function yields

$$\begin{aligned} f(x,y) &= p_{0,0} + \sum_{i=0}^{\infty} \frac{1-\beta_i}{1-\beta_i y} [\frac{(1-\alpha_i)\alpha_i x}{1-\alpha_i x} - \frac{(1-\alpha_{i+1})\alpha_{i+1} x}{1-\alpha_{i+1} x} - (\alpha_i - \alpha_{i+1})\beta_i y] \\ &+ \sum_{i=0}^{\infty} \frac{1-\beta_i}{1-\beta_i x} [\frac{(1-\alpha_i)\alpha_i y}{1-\alpha_i y} - \frac{(1-\alpha_{i+1})\alpha_{i+1} y}{1-\alpha_{i+1} y} - (\alpha_i - \alpha_{i+1})\beta_i x]. \end{aligned} \tag{2.26}$$

Observe that this generating function is a meromorphic function in x for all y and also in y for all x, with simple poles at $1/\alpha_i$ and at $1/\beta_i$, $i = 0, 1, ...$ It can be verified that the sequence $\{1/\alpha_0, 1/\beta_0, 1/\alpha_1, 1/\beta_1, ...\}$ corresponds to the sequence $\{x_1^+, x_2^+, \ldots\}$. Finding, in the compensation method, pairs (α_i, β_i) that satisfy (2.2) (i.e., finding terms $\alpha_i^m \beta_i^n$ that satisfy the equilibrium equations in the interior of the state space) corresponds to finding pairs (x, y) that satisfy $xy - r(x, y) = 0$; and for such pairs, we have $(x, y) = (\frac{1}{\alpha_i}, \frac{1}{\beta_i})$. Similarly, trying to satisfy the relation $g(x) + g(y(x)) = 0$ for zero-tuples of the kernel corresponds to trying to satisfy equilibrium equations on the boundaries as well as in the interior. In the compensation method each time a new term is added, to compensate an error on one of the boundaries; in the function-theoretic method based on meromorphic functions, this is translated into adding a new pole x_n^+ to compensate a pole x_{n-1}^+ in $g(x) + g(y(x))$ (or $g(x(y)) + g(y)$).

2.3 The shortest queue

One of the most important multiserver models is the shortest queue. In this model there are two or more parallel servers, each with their own queue. Customers arrive according to a Poisson process. An arriving customer chooses the shortest queue. The service times are exponentially distributed. Many solution techniques have been developed and applied to this problem. In fact, the three methods discussed in the previous subsection also work for the shortest queue with two servers. For an application of the first method we refer to [7, 8]; for the second method to Fayolle and Iasnogorodski [62, 86] (who, in their respective PhD theses, reduce the problem to a generalized Riemann-Hilbert problem) and to [47]; and for the third one to [91, 68, 42, 45]. In the last two papers it is shown that the bivariate queue-length generating function is meromorphic in each of its variables, and the zeros and poles of the bivariate queue-length generating function (GF) are determined, yielding a relatively simple expression for this GF. But these methods only work for two queues. In this section we discuss two other methods, viz.: (i) the precedence relation method and (ii) the power-series algorithm (PSA). The strength of these methods is that they are not restricted to two queues, but apply equally well to more than two queues. A weak point, however, is that method (i) produces no exact results, but bounds, and that the theoretical foundation of method (ii) is still incomplete.

Method 1: The precedence relation method
This method has been developed in [82, 83]. It is a systematic approach to the construction of bounds for the average cost in Markov chains. In particular, the method is useful for generating bounds for relevant performance characteristics in Markovian systems. Usually, these characteristics can be expressed as the average cost for some appropriately chosen cost function. Below we explain how the method leads to bounds for the mean sojourn time in the shortest queue problem. To keep the presentation simple we consider the case of two identical servers.

The shortest queue model can be represented by a continuous-time Markov process with states $\bar{m} = (m_1, m_2)$ where m_1 and m_2 are the length of the shortest and longest queue, respectively (so $m_1 \leq m_2$). If we take as cost function the total number of customers in the system, so $c(\bar{m}) = m_1 + m_2$, then the average cost g yields the mean number of customers in the system, and by Little's law, the mean sojourn time.

Step 1: Convert the continuous-time Markov process into a discrete-time Markov chain.

This proceeds straight-forwardly by uniformizing the transition rates. Let λ and μ denote the arrival rate and service rate, respectively. In the states $(0, m_2)$ with $m_2 > 0$ and $(0, 0)$ we introduce fictitious one-step transitions from the state to itself with rate μ and 2μ, respectively (see Figure 4). This implies that in each state the transition rates add up to $\lambda + 2\mu$, where without loss of generality we may take $\lambda + 2\mu = 1$. Hence we can interpret λ and μ as transition probabilities and the corresponding Markov chain has the same distribution as the original Markov process. If we take $c(\bar{m})$ as cost per period, then it also has the same average cost. From here on we consider the Markov chain.

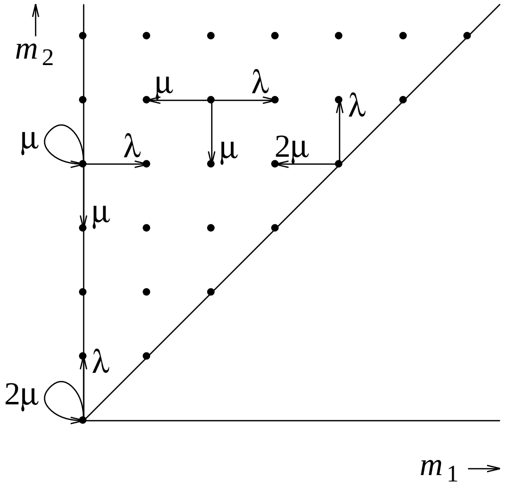

Figure 4: One-step transition rate diagram for the shortest queue system.

Step 2: Identify precedence relations. The precedence relation method systematically constructs modifications of the Markov chain, which produce bounds for the average cost. First one should try to identify *precedences* between states. We say that $\bar{m}$ has precedence over $\bar{n}$, or is more attractive than $\bar{n}$, if $v_t(\bar{m}) \leq v_t(\bar{n})$ for all $t \geq 0$. Here $v_t(\bar{m})$ denotes the expected cost in the first t periods when starting in state $\bar{m}$. Based on these precedences it appears to be easy to construct modifications yielding bounds. Of course one should aim for bounds that are easy to compute.

For the present model it is intuitively clear that state $\bar{m} = (m_1, m_2)$ is more attractive than its neighbours $(m_1 + 1, m_2)$, $(m_1, m_2 + 1)$ and $(m_1 - 1, m_2 + 1)$, provided these neighbours are in the state space. This means that it is preferable to have less customers in the system and to have more balance in queue lengths. The precedences are shown in Figure 5. They can be proved by induction over t. It is important to note that (i) the proof of the induction step can be translated to a standard *transportation problem* (which may be useful in more complex models), and (ii) the precedences depend on the choice of the cost function c.

Step 3: Construction of bounds. We now consider a modification of the original Markov chain by *redirecting* in some states $\bar{m} = (m_1, m_2)$ one or more transitions. Redirecting a transition from (m_1, m_2) to (n_1, n_2) to another state $(\tilde{n}_1, \tilde{n}_2)$, say, means that the new transition probability to (n_1, n_2) is zero and the one to $(\tilde{n}_1, \tilde{n}_2)$ is increased by the original transition probability from (m_1, m_2) to (n_1, n_2). We assume that the modified chain has only one recurrent subchain and that its equilibrium distribution exists. The cost per period is not altered and the average cost is denoted by $\tilde{g}$. The following result is crucial for the construction of bounds. If the modified

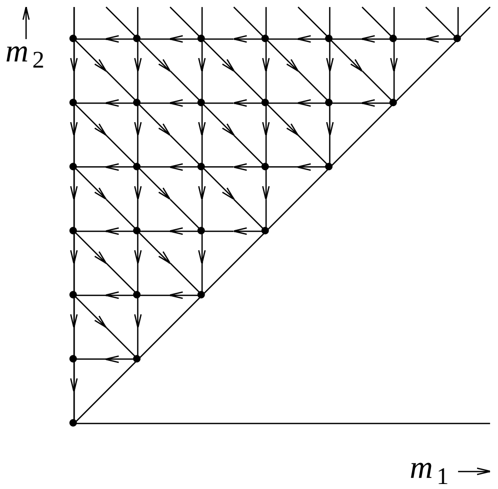

Figure 5: Precedence relations for the shortest queue model with one-period cost $c(\bar{m}) = m_1 + m_2$. Each arrow points to a more attractive state.

chain is obtained by redirecting to more (less) attractive states, then it holds for all t that the expected t-period cost vector in the modified chain is less (greater) than the one in the original chain. This can be proved by induction and it directly implies that $\tilde{g} \leq (\geq)\ g$.

Hence, from the precedences in Figure 5 it is easy to construct models producing upper or lower bounds for g. Two such modifications are shown in Figure 6. In model (a) transitions are redirected to more attractive states, so this model gives a lower bound for g. The interpretation of the redirections is that one customer is allowed to jockey from the longest to the shortest queue as soon as the difference between the queue lengths exceeds a threshold T. Figure 6 shows model (a) for $T = 3$. In model (b) transitions are redirected to less attractive states, so this model yields an upper bound. Its interpretation is that when the departure of a customer would lead to a difference in queue lengths greater than T, then its departure is blocked and the customer is taken into service once more. In both models the state space is a semi-infinite strip, the size of which is controlled by a threshold parameter T. They can be solved very efficiently by using the matrix geometric method of Neuts [115], see also Section 4.2. In fact, the rate matrix can be obtained explicitly in the two models (cf. [118]). It will be clear that the larger T the more accurate the bounds will be, but also the more effort it takes to compute them. Note that the two models exploit the property that most of the probability mass in the shortest queue model is concentrated around the diagonal of the state space. So the bounds will already be tight for small values of T. Based on the precedences it is possible to construct a number of other models producing bounds. They cover the ones studied in [10, 48, 74, 119, 134].

The idea to construct bounds on the basis of precedences has been used for a number of specific models, see e.g. [3, 84, 2, 55, 56, 57, 58, 59]. The precedence relation method can be seen as an attempt to unify the approaches in these references.

Method 2: The power series algorithm

The power-series algorithm (PSA) is a numerical-analytic method for analyzing certain Markov processes. The main idea of the PSA is to consider the steady-state distribution of the Markov process as a function of a system parameter. In the shortest queue model, this system parameter

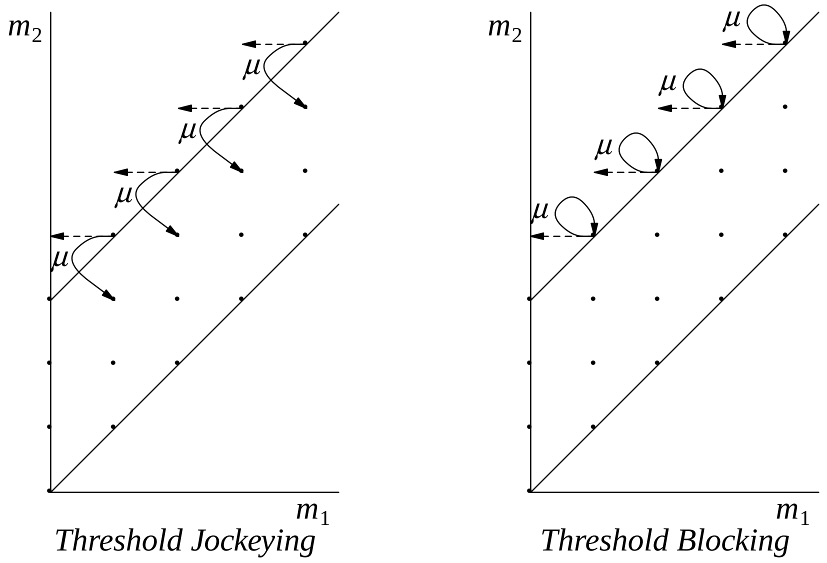

Figure 6: Two modifications for the shortest queue model. The dashed arrows denote the original transitions and the solid arrows the redirected ones.

could, e.g., be the traffic load ρ, normalized such that the system is ergodic for $\rho \in [0, 1)$. The steady-state queue length probabilities are clearly functions of ρ. *Assume* that these probabilities are *analytic* functions of ρ when traffic is light, so that they are determined by the coefficients of the power-series expansion in ρ around $\rho = 0$. Substitute these power-series expansions in the balance and normalization equations of the Markov process. This results in equalities of analytic functions. Analytic functions are only equal if all coefficients of their power-series expansions are identical. We thus obtain equations that enable the recursive calculation of these coefficients.

Let us briefly outline this procedure for the shortest queue model of $s \geq 2$ identical servers; for details regarding the application of the PSA to the shortest queue, see [20, 22]. Denote the steady-state distribution of the queue-length vector $\overline{N}$ at the s queues by $p(\rho; \overline{n})$, with $\rho \in [0, 1)$ the work load per server. Let $\overline{e}_j$ be the vector with zero entries except for a one at the j-th position, $j = 1, \ldots, s$, and let $I(\cdot)$ denote an indicator function. Furthermore, $P(C = j|\overline{N})$ denotes the probability that an arbitrary arriving customer chooses the j-th server, given the queue-length vector; if several queues are shortest, then that customer will choose among those queues according to the weight function $P(C = j|\overline{N})$. The balance equations for the state probabilities are:

$$[\rho + \sum_{j=1}^{s} \frac{I(n_j > 0)}{s}]p(\rho; \overline{n}) = \sum_{j=1}^{s} \frac{1}{s} p(\rho; \overline{n} + \overline{e}_j) +$$

$$+\rho \sum_{j=1}^{s} I(n_j > 0) p(\rho; \overline{n} - \overline{e}_j) P(C = j|\overline{N} = \overline{n} - \overline{e}_j). \tag{2.27}$$

Now introduce

$$p(\rho; \overline{n}) = \rho^{n_1 + \cdots + n_s} \sum_{k=0}^{\infty} \rho^k a(k; \overline{n}). \tag{2.28}$$

The idea is, that the $p(\rho; \bar{n})$ are analytic functions of the traffic load ρ for small values of ρ, and that in fact the limit for $\rho \downarrow 0$ of $\rho^{-(n_1+\cdots+n_s)}p(\rho; \bar{n})$ will exist. Substituting (2.28) into (2.27) and equating the coefficients of corresponding powers of ρ in the resulting equations yields the following equations: For $k = 0, 1, \ldots,$

$$\begin{aligned}\sum_{j=1}^{s} \frac{I(n_j > 0)}{s} a(k; \bar{n}) = &- \sum_{j=1}^{s} \frac{I(n_j > 0)}{s} I(k > 0) a(k-1; \bar{n}) \\ &+ \sum_{j=1}^{s} \frac{1}{s} I(k > 0) a(k-1; \bar{n} + \bar{e}_j) \\ &+ \sum_{j=1}^{s} I(n_j > 0) a(k; \bar{n} - \bar{e}_j) P(C = j | \bar{N} = \bar{n} - \bar{e}_j). \qquad (2.29)\end{aligned}$$

The left-hand side of (2.29) vanishes when $\bar{n} = \bar{0}$. In addition we have from the normalization equation:

$$a(k; \bar{0}) = -\sum \cdots \sum\nolimits_{0<n_1+\cdots+n_s\leq k} a(k - n_1 - \cdots - n_s; \bar{n}), \quad k = 1, 2, \ldots. \qquad (2.30)$$

To obtain the coefficients of $p(\rho; \bar{n})$ up to ρ^M, proceed from $k = 0$ to M: Start from $a(0; \bar{0}) = 1$; determine $a(k; \bar{0})$ from (2.30), and then calculate $a(k; \bar{n})$ recursively from (2.29) for increasing values of $\sum n_i$ up to $\sum n_i = M - k$ (cf. [20]).

However, the power-series expansion of $p(\rho; \bar{n})$ as functions of ρ is not convergent on the whole interval $0 < \rho < 1$ for this model. To enlarge the convergence radius of the power-series expansion we can use the conformal mapping $\theta = \frac{1+G}{1+G\rho}\rho$ (with $G \geq 0$) that maps (0, 1) onto itself and introduce

$$p(\rho(\theta); \bar{n}) = \theta^{n_1+\cdots+n_s} \sum_{k=0}^{\infty} \theta^k b(k; \bar{n}). \qquad (2.31)$$

Substitution of (2.31) into (2.27) and equating the coefficients of corresponding powers of θ again yields a set of equations from which the coefficients can be solved recursively. The power series (2.31) now has a larger radius of convergence than the original one (2.28), i.e., it converges for values of $\rho(\theta)$ for which (2.28) diverges.

The principal idea of the PSA is originally due to Beneš̃ [15], p. 295 ff. It was independently rediscovered by Hooghiemstra, Keane and Van de Ree [78], who applied it to the model of coupled processors with exponential service times (see also Section 3.5). In [14], the coefficients of the expansion are obtained explicitly for the symmetric two-queue coupled processor case. In a series of papers, Blanc and his colleagues greatly extended the applicability of the PSA (see the survey [23]). One of the key contributions of Blanc [21] was the introduction of the epsilon algorithm which strongly improved the convergence properties of the PSA. Koole [98] showed that the PSA can be applied to *any* Markov process with a single recurrent class. Van den Hout and Blanc [80, 79] extended the applicability of the PSA to networks of queues with a multi-queue Markovian arrival process, Markovian service process and Markovian routing. In his PhD thesis, Van den Hout [79] takes the viewpoint that the PSA not only can be viewed as a light-traffic method, but also as a homotopy method: He transforms particular, well-chosen, transition rates of the original Markov process with a parameter γ, such that for $\gamma = 1$ the original process reappears and the transformed process for γ near 0 is easy to analyze. Knowledge about the transformed process near 0 may then be used to solve the problem at $\gamma = 1$. He also strengthened the theoretical foundation of the PSA by proving for a wide class of models that the steady-state probabilities of the transformed process are, indeed, analytic functions of the transformation parameter γ at $\gamma = 0$. Furthermore, he developed some ideas on using the PSA to calculate the transient distribution of a continuous-time Markov process.

The major advantage of the PSA is its flexibility: apart from the Markovian assumption, it does not require much structure. It also does not require sophisticated numerical methods, apart from extrapolation methods like the epsilon algorithm to make the PSA applicable for intermediate and even heavy traffic. Disadvantages of the PSA are: (i) it suffers from the curse of dimensionality (because it directly depends on the balance equations); (ii) no useful error bounds exist sofar; (iii) it is sensitive to extreme parameter values.

Remark 2.2 The MacLaurin approach of Gong and Hu [75] for the $GI/G/1$ queue and of Zhu and Li [135] for the Markov-modulated $G/G/1$ queue is another light-traffic approach, that is somewhat similar to the PSA. These authors derive power-series expansions of the moments of the sojourn and waiting time from the Lindley recursion equation. In this respect, a useful result of Hu [85] should be mentioned; he has proven certain performance measures of $GI/G/1$ queues to be analytic at zero as a function of the arrival rate, when the interarrival time distribution can be expanded as a MacLaurin series over $[0, \infty)$. Blaszczyszyn, Frey and Schmidt [24] and Baccelli and Schmidt [13] also use a light-traffic approach to analyze Markov-modulated multiserver queues and Poisson-driven (max,+)-linear structures, respectively. The latter systems can model *non*-Markovian stochastic Petri nets in the class of event graphs. Examples include fork-join networks and synchronized queueing networks, and Kanban manufacturing networks.

Other approaches to the shortest-queue problem

Knessl et al. [95] develop a scheme to obtain approximations for the joint queue length distribution, valid when one of the queue lengths is large. Foschini and Salz [69] use a heavy traffic diffusion approximation. Gertsbakh [74] and Rao and Posner [119] apply the matrix-geometric method, and Zhao and Grassmann [133] propose an algorithm based on the results of [68]. Halfin [77] obtains bounds by using linear programming techniques, and Nelson and Philips [112] present mean response time approximations for the case of K queues and general interarrival and service time distributions, assuming in their approximation method that the various queue lengths can differ by at most one (see also [111]). Lui et al. [103, 104] develop approximations for a generalization of the shortest queue model, viz. shortest expected delay routing of customers to servers with different working speeds.

Remark 2.3 The ordinary $M/G/2$ queue basically is a 2-queue model where the customers are assigned to the queue with the smallest *workload* instead of the shortest queue. Using a Wiener-Hopf decomposition, Cohen [33] presents an exact analysis of the two-dimensional workload process of this M/G/2 queue, for the case in which the service times have a rational LST. Knessl et al. [96] present an asymptotic analysis for general service times. Formal asymptotic approximations are constructed for the two-dimensional workload process, treating separately the asymptotic limits of heavy traffic, light traffic and large buffer contents.

2.4 The fork-join queue

The fork-join queue is a simple model for a parallel processing system. It consists of c parallel servers, each with their own queue. Each arriving job consists of c subjobs, who each join a different queue (the fork node). A job is completed when all its subjobs have completed service (the join node). Clearly the queues in this model are dependent due to the coupled arrivals. This model frequently arises in the context of computer systems, production systems and maintenance systems.

An exact analysis is possible for $c = 2$ and Poisson arrivals. The model with heterogeneous exponential servers has been studied by Flatto and Hahn [67]. The generating function for the queue lengths is found by using a uniformization technique (expressing two complex variables as analytic functions of one and the same variable) and from this asymptotic expressions for the equilibrium probabilities are obtained. Their generating function is shown to be a meromorphic function on a 2-sheeted Riemann surface. Wright [130] generalizes the result of [67] by also allowing jobs consisting of one subjob to join the system. A completely different approach to obtain asymptotic results has been used by Shwartz and Weiss [122] (see also the afterword in [130]). It is based on large deviations and time reversibility. Baccelli [11] solves the model with general, but exchangeable, service times by using complex-function theory methods. In his PhD thesis, De Klein [92] analyses the model with general service times. He applies and compares two methods, viz. the boundary value method (see Subsection 2.2) and the singular integral equation technique. The latter technique has been initiated by Eisenberg [60] and it concerns the reduction of the functional equation for the generating function to a Fredholm integral equation. For the model with more than two servers no exact analytical results are available in the literature. In this case, bounds and approximations have been developed, see e.g. [12, 113, 114].

3 Servers choose a customer

3.1 Introduction

In this section we consider queues with multiple waiting lines, where the server or servers choose a customer for service according to some priority rule. In polling systems the priority is changing dynamically; these systems are considered in Subsection 3.2. Systems with fixed (static) priority are studied in Subsection 3.3. The 'priority for the longer queue' discipline is briefly considered in Subsection 3.4. An important mechanism to introduce priorities in integrated-services networks with heterogeneous Quality-of-Service requirements is Generalized Processor Sharing; it is the topic of Subsection 3.5.

While some of these models may, in particular cases, be analyzed by direct methods like the power-series algorithm, the methodological emphasis in this chapter is on complex-function methods.

3.2 Polling systems

The performance analysis of computer-communication and production systems often gives rise to single-server multi-queue models. The characteristic feature of these so-called *polling* models is that a single server is moving between a number of queues (which possibly requires some switchover time), implying that the priority of the queues is dynamically (e.g., cyclically) changing. Many applications of polling models can be found in [76, 102, 127]. In [126, 128] extensive and updated bibliographies of polling studies are given.

In a cyclic polling model, the joint queue length process at polling instants (i.e. time points at which the server starts a visit at a queue) can under some conditions on the service disciplines at the queues be represented by a multi-type branching process with immigration [120]. In the case of polling models with switchover times, it turns out that we are dealing with a multi-type branching process with immigration *in each state*, whereas in the case of polling models without switchover times we are dealing with a multi-type branching process with immigration *only in state zero*. The theory of such branching processes then provides necessary and sufficient ergodicity conditions, and, using an iterative method, it gives an explicit expression for the generating function of the joint queue length process at polling instants.

Let us describe the polling model, the required conditions on the service disciplines at the queues and the iterative method in some more detail. A single server S visits n infinite-buffer queues $Q_1, \dots, Q_n$ in cyclic order. When S moves from Q_i to Q_{i+1} this requires a switchover time with distribution function $S_i(\cdot)$ with mean σ_i and Laplace-Stieltjes transform $\sigma_i(\cdot)$. Customers arrive at Q_i according to a Poisson process with rate λ_i. The service times of customers at Q_i have distribution function $B_i(\cdot)$ with mean β_i and Laplace-Stieltjes transform $\beta_i(\cdot)$. The service discipline at each queue should satisfy the following 'branching' property:

Property 1 *If S arrives at Q_i to find k_i customers there, then during the course of the server's visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function $h_i(z_1, \dots, z_n)$, which can be any n-dimensional probability generating function.*

For example, in the case of the well-known gated and exhaustive service disciplines, Property 1 is satisfied with $h_i(z_1, \dots, z_n) = \beta_i(\sum_{j=1}^{n} \lambda_j(1-z_j))$ and $h_i(z_1, \dots, z_n) = \eta_i(\sum_{j \neq i} \lambda_j(1-z_j))$, respectively, where $\eta_i(\cdot)$ is the Laplace-Stieltjes transform of the length of the busy period of the 'corresponding' isolated $M/G/1$ queue of Q_i. Unfortunately, the branching property does not hold for several important service disciplines, like those that put a limit on the time of a server visit or the number of services during a server visit.

From now on we use the notation $z = (z_1, \dots, z_n)$ and we denote by $V(z)$ the generating function of the joint queue length vector at polling instants of a fixed queue, say Q_1. Under the branching property we can prove that

$$V(z) = V(f(z))g(z). \tag{3.1}$$

Here, the function $f(z)$, defined by

$$f(z) := (f_1(z), \dots, f_n(z))$$

with

$$f_i(z) := h_i(z_1, \dots, z_i, f_{i+1}(z), \dots, f_n(z)),$$

represents the *offspring* generating function. The function $g(z)$, defined by

$$g(z) = \prod_{i=1}^{n} \sigma_i(\sum_{j=1}^{i} \lambda_j(1-z_j) + \sum_{j=i+1}^{N} \lambda_j(1-f_j(z))),$$

represents the *immigration* generating function in the branching process.

Iteration of equation (3.1) yields

$$V(z) = \prod_{k=0}^{\infty} g(f^{(k)}(z)), \tag{3.2}$$

where the iterates $f^{(k)}(z)$ are defined by

$$\begin{aligned} f^{(0)}(z) &= z, \\ f^{(k)}(z) &= f(f^{(k-1)}(z)). \end{aligned}$$

The infinite product in (3.2) is convergent when the ergodicity condition is fulfilled.

In the case of zero switchover times we have, instead of Equation (3.1), that

$$V(z) = V(f(z)) - \pi_0(1 - g(z)). \tag{3.3}$$

Here the offspring generating function $f(\cdot)$ is defined as before, π_0 is the probability of an empty system and the immigration generating function $g(\cdot)$ is now given by

$$g(z) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} z_j, \quad \text{or} \quad g(z) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} f_j(z),$$

depending on the behaviour of the server when the system becomes empty. The first expression corresponds to the situation where, when the system becomes empty, the server makes a full cycle and subsequently stops right before Q_1 (see [25]). The second one corresponds to the situation where in an empty system the server immediately stops right behind Q_1 (see [120]). Iteration of Equation (3.3) yields

$$V(z) = 1 - \pi_0 \sum_{k=0}^{\infty} [1 - g(f^{(k)}(z))]. \tag{3.4}$$

The infinite sum in (3.4) is convergent when the ergodicity condition is fulfilled.

As mentioned before, the branching property does not hold for several important service disciplines, like those that put a limit on the time of a server visit or the number of services during one server visit. In exceptional two-queue cases of the latter class, the joint queue length distribution can be determined by using the theory of Riemann-Hilbert boundary value problems [27, 46, 47, 100]. A pioneering paper in this area was that of Eisenberg [60], transforming a two-queue polling problem into a Fredholm integral equation.

Numerical approaches

Leung [101] has developed a numerical procedure, based on the fast Fourier transform, that enables one in principle to determine polling performance measures with any required accuracy. The power series algorithm is also applicable to a large class of polling models (see Blanc [21]). An essential difficulty of both numerical techniques is their large computational complexity. Choudhury and Whitt [30] have shown that probability distributions and moments of performance measures for many polling models can be effectively computed by numerically inverting generating functions and Laplace-Stieltjes transforms. The computational complexity of their approach is much lower than that of the approaches of Leung and Blanc. However, the method of Choudhury and Whitt does not readily extend to models that do not satisfy the branching property.

If switchover times are zero, work conservation leads to a conservation law, an exact expression for a particular weighted sum of mean waiting times [93]. If switchover times are non-zero, a principle of work decomposition [26, 27] gives rise to a pseudo-conservation law which is again an exact expression for a particular weighted sum of mean waiting times. This can provide a useful check for the accuracy of simulations, numerical calculations and approximations, as well as providing the basis for further approximations.

3.3 Fixed priorities

An important queueing model is the single server queue with k classes of customers, that arrive according to independent Poisson processes and have fixed priorities. The non-preemptive and preemptive-resume priority disciplines have in particular been analyzed in considerable detail. A classical reference is the book of Jaiswal [89]; see also the discussions in [34, 94, 129]. Jaiswal discusses, a.o., generating functions of joint queue length distributions. He heavily uses the concept of *completion time*, viz., the time from the start of a service of a particular class until the server is available to serve the next customer of that same class. The completion time concept allows a unified study of several priority disciplines. Another important observation in single server priority queues is that it often suffices to study $k = 2$ classes, temporarily aggregating all lower (higher) classes into a single class.
In the present study we focus on the challenging problem of *multiserver* priority queues. Both interarrival and service times are now assumed to be exponentially distributed. In most studies that have considered multiserver queues with *nonpreemptive priority*, the restrictive assumption has been made that all mean service times are equal. Cohen [32] derived the joint queue length distribution under this assumption, for the case $k = 2$. Cobham [31] derived the mean waiting times under the same assumption, for general k, while Davis [52] and Kella and Yechiali [90] obtained the waiting time transform of each class. Gail et al. [70] discarded the assumption of equal service means. They obtained, for $k = 2$ classes and an arbitrary number m of servers, the steady-state bivariate generating function of a convenient two-dimensional Markov process that is immediately related to the queue length process. We sketch their approach and refer to [70] for the (very involved) analytic details. The generating function equations corresponding to the equilibrium equations take the form

$$A(z, w)G(z, w) = B(z, w)F(w) + C(z, w)K. \tag{3.5}$$

G, F are unknown vector valued analytic functions, K is a vector of unknown constants, and A, B, C are known $(m+1) \times (m+1)$ matrices with polynomial entries. Gail et al. then identify the $m + 1$ zeros $z_i(w)$ of $\det A(z, w)$ for given $|w| \leq 1$, with the property $|z_i(w)| \leq 1$. Since the GF $G(z, w)$ should be bounded in $|z| \leq 1$, $|w| \leq 1$, this yields a matrix equation

$$M(w)F(w) = N(w)K, \tag{3.6}$$

where $M(w)$ is known. This is a set of $(m + 1)^2$ equations, at most $m + 1$ of which are independent. A careful study of the zeros of $\det M(w)$ inside the unit circle, instigated by the

boundedness of the GF $F(w)$ inside the unit circle, yields a set of equations from which K can be determined. $F(w)$ then follows from (3.6) and finally $G(z, w)$ from (3.5).
Gail et al. [71] have used a similar approach, with a similar state-space representation, to determine the bivariate queue-length GF in the *preemptive resume* case (again there are m servers and 2 classes).

3.4 Priority for the longest queue

The queueing model with priority for the longest queue is in a sense dual to the shortest queue model. It is also similar, in the sense that both systems employ a mechanism that tends to equalize the queue lengths. Cohen [37] solves the 2-queue model with server priority for the longest queue. He uses a translation into a Riemann boundary value problem of a type that was not studied earlier in a queueing context. In Van Houtum et al. [84] the precedence relation method (see Subsection 2.3) is used to obtain lower and upper bounds for the model with exponential service times and an arbitrary number of queues. The exponential longest queue model with two queues has also been studied by Flatto [66] in the case that the longest queue policy is applied preemptively (i.e. service to a customer in a given queue is interrupted whenever the other queue becomes larger). He uses a generating-function approach. He argues that the longest queue model is easier than the shortest queue model, in the sense that in the former one $f(0, y)$ does not appear in the basic functional equation, and that it is a rational function of y.

3.5 Generalized Processor Sharing

In the design of high-speed networks, an increasingly important issue is the provision of Quality-of-Service (QoS) guarantees. A key problem is the study of scheduling disciplines to be employed at network switches. Ideally, these scheduling disciplines should protect sessions from the possibly bad behavior of other sessions, but also exploit statistical multiplexing gain. One design paradigm for such scheduling disciplines is the Generalized Processor Sharing (GPS) discipline. GPS-based scheduling algorithms like Weighted Fair Queueing have emerged as an important mechanism for accommodating heterogeneous QoS requirements in integrated-services networks.
GPS operates as follows. Consider N sources (sessions) sharing a link of unit rate. There are weights $\phi_1, \dots, \phi_N$ associated with each of these sources, with $\sum_{i=1}^{N} \phi_i = 1$. If all the sources are backlogged at time t, i.e., the buffer content of each source is positive at time t, then source i is served at rate ϕ_i, $i = 1, \dots, N$. But if some of the sources are not backlogged, then the excess capacity is redistributed among the backlogged sources according to their respective weights. GPS has the following two properties: (i) it is work-conserving, serving at the full link rate whenever at least one source is backlogged; (ii) it guarantees minimum rate ϕ_i to source i whenever it is backlogged.
Generally speaking, the stochastic analysis of the GPS discipline is prohibitively difficult. However, the case $N = 2$ allows, for several variants, a quite detailed analysis. Konheim, Meilijson and Melkman [97] determine the joint queue length distribution in the completely symmetric case of independent Poisson arrival processes and exponential service times (with the same rates at each queue) and $\phi_1 = \phi_2 = 1/2$. They employ a uniformization technique. The coupled processor model, as it was called by the above-mentioned authors, has been a most fruitful

model from the viewpoint of methodology. In a pioneering paper, Fayolle and Iasnogorodski [63] have shown that the functional equation for the two-dimensional generating function of the joint queue length distribution in the asymmetrical coupled processors may be transformed into a Riemann-Hilbert boundary value problem. This allowed the application of powerful results from the theory of boundary value problems. Cohen and Boxma [47] have presented a systematic and detailed study of that new technique. In many cases, their approach allowed *general* service time distributions. One of these cases is the coupled processor model; in [47] it is exactly analyzed for the case of generally distributed service times. The service speed of server i is r_i when the other server is also busy, and r_i^* when the other server is idle ($r_1^* = r_2^* = r_1 + r_2$ gives GPS). The topic of interest in Chapter III.3 of [47] is the LST $\phi(s_1, s_2)$ of the joint workload distribution at both servers. A parametrisation technique yields a suitable set of zeros $(s_1, s_2) = (\delta_1(w), \delta_2(w))$ of the kernel, leading to a Wiener-Hopf problem with boundary (line of discontinuity) Re $w = 0$.

The question naturally arises whether the boundary value technique can be extended to queueing models with a higher-dimensional state space. Cohen [35, 36] obtains partial results for 3 coupled processors, while he also derives a large collection of interesting results for entrance times and entrance points of homogeneous N-dimensional random walks [39]. We are not aware of other such higher-dimensional results. For the same coupled-processor model (not necessarily completely symmetric, and with an arbitrary number $N \geq 2$ of queues), Hooghiemstra, Keane and Van de Ree [78] have developed the Power Series Algorithm (PSA; see Section 2.3).

For the general GPS scheduling discipline with an arbitrary number of sources, there seems to be little hope of obtaining explicit expressions for queue length or workload distributions. It is already quite an accomplishment to derive useful bounds on backlog (workload) and delay. Parekh and Gallager [116, 117] study GPS in which the arriving traffic conforms to Cruz's *linear bounded arrival processes* [49, 50]. They obtain *worst-case* deterministic upper bounds on backlog and delay for each session. Thus hard guarantees can be given for networks employing GPS scheduling. Zhang et al. [132] model the source session traffic as an *exponentially bounded burstiness process* (a notion introduced in [131]), and derive *statistical* bounds on backlog and delay for each session. Other important contributions to the performance analysis of GPS are based on a large deviations approach: see Bertsimas et al. [18], Dupuis and Ramanan [54], and Massoulié [106].

4 Various two-dimensional random walks

4.1 Introduction

In this section we discuss two classes of Markov processes that naturally arise in queueing models with multiple waiting lines, and for which a detailed analysis is possible. Subsection 4.2 considers Markov processes on a semi-infinite strip. Three methods for determining the equilibrium probabilities are briefly described. Subsection 4.3 is concerned with tractable two-dimensional random walks on the lattice in the first quadrant. Some methods that were discussed in the previous sections also apply to these Markov processes. As indicated before, the power-series algorithm can be applied to any Markov process with a single recurrent class but has some restrictions regarding its theoretical foundation and suffers from the curse of dimensionality; the compensation method applies under restrictions on the one-step transition rates.

4.2 Markov processes on a semi-infinite strip

Many queueing systems can be modelled by a Markov process, the state space of which is given by a semi-infinite strip of states (m, n) where m ranges from 0 to s and n from 0 to ∞. In these systems, typically m denotes the state of the service facility or of the arrival process, and n denotes the number of jobs waiting in the system. m and n could also denote the number of type-1 and type-2 customers, the waiting room for type-1 customers being finite. Often, there is a threshold N, such that the transition rates out of state (m, n) do not depend on n when $n \geq N$.

The equilibrium probabilities $p_{m,n}$ can be determined using three alternative methods. They are briefly described below. A comparison of the three methods can be found in the survey paper by Mitrani [107].

Method 1: The matrix-geometric method
In this approach the row vectors of equilibrium probabilities $\bar{p}_n = (p_{0,n}, p_{1,n}, \ldots, p_{s,n})$ are expressed as

$$\bar{p}_n = \bar{p}_N R^{n-N}, \qquad n \geq N, \tag{4.1}$$

where the so-called rate matrix R is the minimal nonnegative solution of a nonlinear matrix equation; see Neuts' book [115]. For the special class of *quasi-birth-and-death processes*, i.e., processes where for each state (m, n) outgoing transtions are restricted to states (k, l) with $|l - n| \leq 1$, Ramaswami and Latouche [99] have developed a highly efficient algorithm to solve the matrix equation for the rate matrix R.

Method 2: The generating function method
This approach uses generating functions to solve the set of equilibrium equations. By introducing the vector generating function $\bar{g}(z) = (g_0(z), g_1(z), \ldots, g_s(z))$ where

$$g_m(z) = \sum_{n=N}^{\infty} p(m, n) z^n,$$

the equilibrium equations can be transformed to a matrix equation for $\bar{g}(z)$ of the form

$$\bar{g}(z) A(z) = \bar{b}(z).$$

The vector $\bar{b}(z)$ involves a number of unknown boundary probabilities. These probabilities are determined by exploiting the zeros of the determinant of the matrix $A(z)$ inside or on the unit circle. The generating function approach is, for example, used by Mitrani and Avi-Itzhak [108] to analyse the $M/M/s$ queue with service interruptions.

Method 3: The spectral expansion method
This method is based on reducing the equilibrium equations to a vector difference equation with constant coefficients, the solution of which can be expressed in terms of eigenvalues and

eigenvectors of the associated characteristic polynomial; see Mitrani and Mitra [109]. This gives

$$\bar{p}_n = \sum_{i=0}^{s} C_i \bar{y}_i x_i^{n-N}, \quad n \geq N, \tag{4.2}$$

where the geometric factors x_i are the $s+1$ eigenvalues inside the unit circle, the vectors $\bar{y}_i$ are the corresponding eigenvectors and the coefficients C_i follow from the boundary equations and the normalization equation. The difficult step in this approach is the computation of the eigenvalues, which are the $s+1$ roots inside the unit circle of a determinantal equation. A direct approach of finding these roots is inefficient for large s and therefore not recommended. However, in special cases, particular properties may be exploited to simplify the eigenvalue problem, see e.g. Elwalid et al. [61]. For a class of systems with rates *linear in m or $s-m$*, Adan and Resing [5] showed that the roots can be determined very efficiently. They use a generating function technique to reduce the single equation for the $s+1$ roots inside the unit circle to $s+1$ equations for a single root in the interval $(-1,1)$. This considerably simplifies the determination of the roots, also because the computations can now be restricted to the real domain. Queueing models included in this class of systems are, e.g., the multi-server queue with Poisson arrivals and E_2, H_2 or C_2 distributed service times (see [121, 123, 124, 17]), the $M/M/s$ queue with service interruptions (see [108]), the $\sum IPP/M/1$ queue (see [61]) and a multi-server queue with locking (see [4]).

Remark 4.1 Result (4.2) may be linked to the (modified) matrix-geometric representation (4.1) of the equilibrium distribution. Clearly, when R is diagonalizable, the factors x_m in the form (4.2) are the eigenvalues of R and the row vectors y_m are the associated row eigenvectors (cf. Daigle and Lucantoni [51]).

Remark 4.2 De Smit [125] surveys the application of *matrix* Wiener-Hopf factorizations to the analysis of *waiting times* in multidimensional queues. The class of models that this method can handle shows considerable overlap with the class which can be solved by matrix-geometric methods. One of De Smit's key examples is a semi-Markov queue. Dukhovny [53] also considers semi-Markov queues. He surveys the application of *vector* Riemann-Hilbert boundary value problems to the analysis of *queue lengths* in multidimensional queues.

4.3 Markov processes on the lattice in the first quadrant

An interesting class of two-dimensional random walks on the lattice in the first quadrant has been studied by Fayolle, King and Mitrani [65]. They assume that there exist positive integers N_1 and N_2, such that the transition rates out of state (n_1, n_2) do not depend on n_1 for $n_1 \geq N_1$ and not on n_2 for $n_2 \geq N_2$. One example is an $M/M/1$ queue with two classes of customers and a restricted processor sharing discipline: Up to N_1 jobs of class 1 and up to N_2 jobs of class 2 are allowed to share the processor at any time, and the remaining jobs must wait. Fayolle et al. [65] reduce the determination of the bivariate generating function to the problem of solving a Riemann-Hilbert boundary value problem on a circle.

A quite general class of two-dimensional random walks on the lattice in the first quadrant has been analysed in [47], where the solution is again obtained via transformation to a boundary value problem. One-step transitions to the West, South-West and South can only go to the nearest neighbour, but one-step transitions in other directions may be more general. The functional equation for the unknown generating function $f(x, y)$ of its equilibrium solution is of the following type: For $|x| \leq 1$, $|y| \leq 1$,

$$K(x, y) f(x, y) = A_{10}(x, y) f(x, 0) + A_{01}(x, y) f(0, y) + A_{00}(x, y) f(0, 0) + B(x, y). \quad (4.3)$$

The kernel $K(x, y)$ contains all the information concerning the structure of the random walk in the interior of its state space. The boundedness of the probability generating function $f(x, y)$ for $|x| \leq 1$, $|y| \leq 1$ again leads to an inspection of the zeros of the kernel, $K(x, y)$, in this product of unit circles. For each of those zeros, the righthand side of (4.3) should be zero. Furthermore, $f(x, 0)$ should be analytic in x for $|x| < 1$ and continuous in x for $|x| \leq 1$, similarly for $f(0, y)$. The structure of the problem of determining $f(x, 0)$ and $f(0, y)$ that satisfy these conditions resembles that of a Riemann-type boundary value problem: The determination of analytic functions in prescribed domains, these functions moreover satisfying a linear relation. In [47] the above problem is indeed transformed into a Riemann-type boundary value problem. Using the extensive theory of Riemann-type boundary value problems [72, 110], the above-described two-dimensional random walk is in principle solved. The solution does involve the determination of some conformal mappings, that can be accomplished via the solution of singular integral equations. In most cases, this requires numerical analysis. The above-sketched approach is surveyed in much more detail in [38]. From a numerical point of view, an interesting approach is also the transformation of the functional equation into a Fredholm integral equation [38]; standard techniques are available to solve such an integral equation numerically (cf. [92]).

Various matters simplify when one restricts oneself, within the class of random walks on the lattice in the first quadrant, to the subclass of nearest-neighbour random walks (only transitions to immediate neighbours may occur). In that case, the kernel $K(x, y)$ is a biquadratic function of x and y. A pioneering study of such random walks is the one of Malyshev [105]. Together with Fayolle and Iasnogorodski, [64], he has recently developed a new analytic approach for nearest-neighbour random walks in the quarter plane. Like in the above approach, they consider the (elliptic) curve $K(x, y) = 0$. A key idea of them is to use Galois automorphisms on this algebraic curve. They prove that the unknown functions $f(x, 0)$ and $f(0, y)$, while in general not being meromorphic functions, can be ‘lifted’ as meromorphic functions onto the ‘universal covering’ of some Riemann surface that corresponds to the algebraic curve $K(x, y) = 0$ (see also [67, 130]). Cohen has made several important contributions to the theory of nearest-neighbour random walks. In the monograph [39], he extensively discusses ergodicity conditions, entrance times into the boundaries, and entrance points. He relates the zero-tuples of the kernel to the distributions of the entrance times into the boundary points of the state space, by a very elegant identity. When only a few one-step transitions are allowed in a nearest-neighbour random walk, further simplifications may occur.

Cohen [40] considers the semi-homogeneous nearest-neighbour random walk without transitions to the East, North-East and East. This is the class of random walks studied in [1] via the compensation method. Cohen proves that the bivariate generating function of the stationary distribution of such two-dimensional random walks in the first quadrant can be represented by meromorphic functions. Subsequently he exposes the construction of those meromorphic functions; this construction is based on the iterative calculation of poles and their residues.

Examples are discussed in [42, 43, 44, 45], cf. Subsection 2.2. Cohen [40] observes that the bivariate generating function for this class of random walks may also be obtained using the boundary value method, even when one-step transitions to the North, East and North-East *are* allowed; but he remarks that when such transitions are excluded, then the construction of the meromorphic function via the iterative calculation of poles and residues is simpler because it avoids the explicit calculation of a conformal mapping.

In [41] Cohen considers a nearest-neighbour random walk with transitions to the South-West, North and East. The generating functions $f(x, 0)$ and $f(0, y)$ are here not meromorphic functions. However, the bivariate generating function $f(x, y)$ is shown to be a fairly simple algebraic function that can be explicitly determined. Cohen uses the uniformization technique of Flatto and Hahn [67] and Wright [130], who study a fork-join queue, cf. Subsection 2.4. The underlying model of this random walk study actually is a queueing model with one server, two Poisson classes of customers and 'paired services': As soon as a service has been completed, a new service is started if there are customers present. In general, a couple of customers of different type is simultaneously served (after which they leave simultaneously). If only customers of one type are present after a service completion, then one customer of that type is served. If a service leaves the system empty, then the server starts serving as soon as a customer has arrived. Cohen [41] assumed exponential service time distributions. Blanc [19] allows a general service time distribution, that is the same in each of the above cases. He obtains several performance measures, including the joint queue-length distribution at service completion epochs and at an arbitrary time. He accomplishes this by formulating the problem as a Riemann-Hilbert or a Hilbert boundary value problem. Its solution requires a conformal mapping that in general cannot be obtained explicitly and must be determined numerically.

Acknowledgment. The authors are indebted to Professor J.W. Cohen for some very interesting discussions. The survey paper [29] has provided some useful input for the present paper. It focuses less on the analytic methods, but it considers a broader range of models that are relevant for the performance analysis of parallel and distributed systems; in addition, it discusses optimization aspects.

References

[1] I.J.B.F. Adan. *A Compensation Approach for Queueing Problems*. PhD thesis, Eindhoven University of Technology, 1991.

[2] I.J.B.F. Adan and G. Hooghiemstra. The $M/M/c$ with critical jobs. *Mathematics of Operations Research*, 47:341–353, 1998.

[3] I.J.B.F. Adan, G.J.J.A.N. van Houtum and J. van der Wal. Upper and lower bounds for the waiting time in the symmetric shortest queue system. *Annals of Operations Research*, 48:197–217, 1994.

[4] I.J.B.F. Adan, A.G. de Kok and J.A.C. Resing. A multi-server queueing model with locking. *European Journal of Operational Research*, 116:16–26, 1999.

[5] I.J.B.F. Adan and J.A.C. Resing. A class of Markov processes on a semi-infinite strip. In B. Plateau, W.J. Stewart and M. Silva, editors, *Numerical Solution of Markov Chains (NSMC '99)*, Prensas Universitarias de Zaragoza, Zaragoza, 1999, pages 41–57.

[6] I.J.B.F. Adan and J. Wessels. Shortest expected delay routing for Erlang servers. *Queueing Systems*, 23:77–105, 1996.

[7] I.J.B.F. Adan, J. Wessels and W.H.M. Zijm. Analysis of the symmetric shortest queue problem. *Stochastic Models*, 6:691–713, 1990.

[8] I.J.B.F. Adan, J. Wessels, and W.H.M. Zijm. Analysis of the asymmetric shortest queue problem. *Queueing Systems*, 8:1–58, 1991.

[9] I.J.B.F. Adan, J. Wessels, and W.H.M. Zijm. A compensation approach for two-dimensional Markov processes. *Advances in Applied Probability*, 25:783–817, 1993.

[10] I.J.B.F. Adan, J. Wessels and W.H.M. Zijm. Matrix-geometric analysis of the symmetric shortest queue problem with threshold jockeying, *Operations Research Letters*, 13:107–112, 1993.

[11] F. Baccelli. Two parallel queues created by arrivals with two demands: The $M/G/2$ symmetrical case. Technical report 426, INRIA-Rocquencourt, 1985.

[12] F. Baccelli, A.M. Makowski, and A. Shwartz. The fork-join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds. *Advances in Applied Probability*, 21:629–660, 1989.

[13] F. Baccelli and V. Schmidt. Taylor series expansions for Poisson driven (max,+)-linear systems. *Annals of Applied Probability*, 6:138–185, 1996.

[14] H. Bavinck, G. Hooghiemstra and E. de Waard. An application of Gegenbauer polynomials in queueing theory. *J. Comput. Appl. Math.*, 49:1-10, 1993.

[15] V.E. Beneš̃. *Mathematical Theory of Connecting Networks and Telephone Traffic*. Academic Press, New York, 1965.

[16] S.A. Berezner, C.F. Kriel and A.E. Krzesinski. Quasi-reversible multiclass queues with order independent departure rates. *Queueing Systems*, 19:345–359, 1995.

[17] D. Bertsimas and X.A. Papaconstantinou. Analysis of the stationary $E_k/C_2/s$ queueing system. *European Journal of Operational Research*, 37:272–287, 1988.

[18] D. Bertsimas, I.Ch. Paschalidis and J.N. Tsitsiklis. Large deviations analysis of the generalized processor sharing policy. Report Boston University, 1997.

[19] J.P.C. Blanc. *Application of the Theory of Boundary Value Problems in the Analysis of a Queueing Model with Paired Services.* PhD Thesis, University of Utrecht, 1982.

[20] J.P.C. Blanc. A note on waiting times in systems with queues in parallel. *Journal of Applied Probability*, 24:540–546, 1987.

[21] J.P.C. Blanc. A numerical approach to cyclic-service queueing models. *Queueing Systems*, 6:173–188, 1990.

[22] J.P.C. Blanc. The power-series algorithm applied to the shortest-queue model. *Operations Research*, 40:157–167, 1992.

[23] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In L. Donatiello and R.D. Nelson, editors, *Performance Evaluation of Computer and Communication Systems*, North-Holland, Amsterdam, 1993, pages 53–80.

[24] B. Blaszczyszyn, A. Frey and V. Schmidt. Light traffic aproximations for Markov-modulated multi-server queues. *Stochastic Models* 11:423–445, 1995.

[25] S.C. Borst and O.J. Boxma. Polling systems with and without switchover times. *Operations Research*, 45:536–543, 1997.

[26] O.J. Boxma. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems*, 5:185–214, 1989.

[27] O.J. Boxma and W.P. Groenendijk. Two queues with alternating service and switching times. In O.J. Boxma and R. Syski, editors, *Queueing Theory and its Applications*, North-Holland, Amsterdam, 1988, pages 261–288.

[28] O.J. Boxma and G.J. van Houtum. The compensation approach applied to a 2×2 switch. *Probability in the Engineering and Informational Sciences*, 7:171–193, 1993.

[29] O.J. Boxma, G.M. Koole and Z. Liu. Queueing-theoretic solution methods for models of parallel and distributed systems. In O.J. Boxma and G.M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems – Solution Methods*, CWI Tract 105, CWI, Amsterdam, 1994, pages 1–24.

[30] G.L. Choudhury and W. Whitt. Computing distributions and moments in polling models by numerical transform inversion. *Performance Evaluation*, 25:267–292, 1996.

[31] A. Cobham. Priority assignment in waiting line problems. *Operations Research*, 2:70–76, 1954.

[32] J.W. Cohen. Certain delay problems for a full availability trunk group loaded by two traffic sources. *Communication News*, 16:105–113, 1956.

[33] J.W. Cohen. On the M/G/2 queueing model. *Stochastic Processes and their Applications*, 12:231–248, 1982.

[34] J.W. Cohen. *The Single Server Queue*. North-Holland, Amsterdam, 1982.

[35] J.W. Cohen. On a functional relation in three complex variables; three coupled processors. Report 359, Mathematical Institute, University of Utrecht, 1984.

[36] J.W. Cohen. On the analysis of parallel, independent processors. Report 374, Mathematical Institute, University of Utrecht, 1985.

[37] J.W. Cohen. A two-queue, one-server model with priority for the longer queue. *Queueing Systems*, 2:261–283, 1987.

[38] J.W. Cohen. Boundary value problems in queueing theory. *Queueing Systems*, 3:97–128, 1988.

[39] J.W. Cohen. *Analysis of Random Walks*. IOS Press, Amsterdam, 1992.

[40] J.W. Cohen. On a class of two-dimensional nearest neighbouring random walks. In J. Gani and J. Galambos, editors, *Studies in Applied Probability Theory*, Special Volume of *Journal of Applied Probability*, 31A:207–237, 1994.

[41] J.W. Cohen. Analysis of a two-dimensional algebraic nearest-neighbour random walk (queue with paired services). Technical report BS-R9437, CWI, Amsterdam, 1994.

[42] J.W. Cohen. Two-dimensional nearest-neighbour queueing models, a review and an example. In F. Baccelli, A. Jean-Marie and I. Mitrani, editors, *Quantitative Methods in Parallel Systems*, Springer, Berlin, 1995, pages 141–152.

[43] J.W. Cohen. On the determination of the stationary distribution of a symmetric clocked buffered switch. In V. Ramaswami and P.E. Wirth, editors, *Teletraffic Contributions for the Information Age, Proc. ITC-15*, North-Holland, Amsterdam, 1997, pages 297–307.

[44] J.W. Cohen. On the asymmetric clocked buffered switch. *Queueing Systems*, 30:385–404, 1998.

[45] J.W. Cohen. Analysis of the asymmetrical shortest two-server queueing model. *J. Applied Math. and Stoch. Analysis*, 11:115–162, 1998.

[46] J.W. Cohen and O.J. Boxma. The M/G/1 queue with alternating service formulated as a Riemann-Hilbert problem. In F.J. Kylstra, editor, *Performance '81*, North-Holland, Amsterdam, 1981, pages 181–199.

[47] J.W. Cohen and O.J. Boxma. *Boundary Value Problems in Queueing System Analysis*. North-Holland, Amsterdam, 1983.

[48] B.W. Conolly. The autostrada queueing problem. *J. Appl. Prob.*, 21:394–403, 1984.

[49] R.L. Cruz. A calculus for network delay, Part I: Network elements in isolation. *IEEE Transactions on Information Theory*, 37:114–131, 1991.

[50] R.L. Cruz. A calculus for network delay, Part II: Network analysis. *IEEE Transactions on Information Theory*, 37:132–141, 1991.

[51] J.N. Daigle and D.M. Lucantoni. Queueing systems having phase-dependent arrival and service rates. In W.J. Stewart, editor, *Numerical Solution of Markov Chains*, Marcel Dekker, New York, 1991, pages 161–202.

[52] R.H. Davis. Waiting-time distribution of a multi-server, priority queuing system. *Operations Research*, 14:133–136, 1966.

[53] A. Dukhovny. Applications of vector Riemann boundary value problems to analysis of queueing systems. In J.H. Dshalalow, editor, *Advances in Queueing: Theory, Methods and Open Problems*, CRC Press, Boca Raton, 1995, pages 353–376.

[54] P. Dupuis and K. Ramanan. A Skorokhod problem formulation and large deviation analysis of a processor sharing model. *Queueing Systems*, 28:109–124, 1998.

[55] N.M. van Dijk. A formal proof for the insensitivity of simple bounds for finite multi-server non-exponential tandem queues. *Stochastic Processes*, 27:261–277, 1988.

[56] N.M. van Dijk. Simple bounds for queueing systems with breakdowns. *Performance Evaluation*, 8:117–128, 1988.

[57] N.M. van Dijk. *Queueing Networks and Product Forms: A Systems Approach*. Wiley, Chichester, 1993.

[58] N.M. van Dijk and B.F. Lamond. Simple bounds for finite single-server exponential tandem queues. *Operations Research*, 36:470–477, 1988.

[59] N.M. van Dijk and J. van der Wal. Simple bounds and monotonicity results for finite multi-server exponential tandem queues. *Queueing Systems*, 4:1–16, 1989.

[60] M. Eisenberg. Two queues with alternating service. *SIAM Journal on Applied Mathematics*, 36:287–303, 1979.

[61] A.I. Elwalid, D. Mitra and T.E. Stern. A theory of statistical multiplexing of Markovian sources: spectral expansions and algorithms. In W.J. Stewart, editor, *Numerical Solution of Markov Chains*, Marcel Dekker, New York, 1991, pages 223–238.

[62] G. Fayolle. *Methodes Analytiques pour les Files d'Attente Couplees*. PhD thesis, Université de Paris VI, Paris, 1979.

[63] G. Fayolle and R. Iasnogorodski. Two coupled processors: The reduction to a Riemann-Hilbert problem. *Z. Wahrsch. Verw. Gebiete*, 47:325–351, 1979.

[64] G. Fayolle, R. Iasnogorodski and V. Malyshev. *Random Walks in the Quarter-Plane*. Springer, Berlin, 1999.

[65] G. Fayolle, P.J. King and I. Mitrani. The solution of certain two-dimensional Markov models. *Advances in Applied Probability*, 14:295–308, 1982.

[66] L. Flatto. The longer queue model. *Prob. Engineer. Inform. Sci.*, 3:537–559, 1989.

[67] L. Flatto and S. Hahn. Two parallel queues created by arrivals with two demands. *SIAM Journal on Applied Mathematics*, 44:1041–1053, 1984.

[68] L. Flatto and H.P. McKean. Two queues in parallel. *Comm. Pure Appl. Math.*, 30:255–263, 1977.

[69] G.J. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, 26:320–327, 1978.

[70] H.R. Gail, S.L. Hantler, and B.A. Taylor. Analysis of a non-preemptive priority multi-server queue. *Advances in Applied Probability*, 20:852–879, 1988.

[71] H.R. Gail, S.L. Hantler, and B.A. Taylor. On a preemptive Markovian queue with multiple servers and two priority classes. *Mathematics of Operations Research*, 17:365–391, 1992.

[72] F.D. Gakhov. *Boundary Value Problems*. Pergamon Press, Oxford, 1966.

[73] P.R. Garabedian. *Partial Differential Equations*. Wiley, Chichester, UK, 1967.

[74] I. Gertsbakh. The shorter queue problem: A numerical study using the matrix geometric solution. *European Journal of Operational Research*, 15:374–381, 1984.

[75] W.B. Gong and J.Q. Hu. The MacLaurin series for the $GI/G/1$ queue. *Journal of Applied Probability*, 29:176–184, 1992.

[76] D. Grillo. Polling mechanism models in communication systems – some application examples. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, North-Holland, Amsterdam, 1990, pages 659–698.

[77] S. Halfin. The shortest queue problem. *Journal of Applied Probability*, 22:865–878, 1985.

[78] G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM Journal on Applied Mathematics*, 48:1159–1166, 1988.

[79] W.B. van den Hout. *The Power-series Algorithm*. Ph.D. Thesis, Tilburg University, 1996.

[80] W.B. van den Hout and J.P.C. Blanc. The power-series algorithm for Markovian queueing networks. In W.J. Stewart, editor, *Computations with Markov Chains*, Kluwer Academic Publishers, Boston, 1995, pages 321–338.

[81] G.J.J.A.N. van Houtum, I.J.B.F. Adan, J. Wessels, and W.H.M. Zijm. The compensation approach for three and more dimensional random walks. Technical Report COSOR 92-39, Eindhoven University of Technology, Eindhoven, 1992.

[82] G.J.J.A.N. van Houtum. *New approaches for multi-dimensional queueing systems*. Ph.D. Thesis, Eindhoven University of Technology, 1995.

[83] G.J.J.A.N. van Houtum, W.H.M. Zijm, I.J.B.F. Adan and J. Wessels. Bounds for performance characteristics: a systematic approach via cost structures. *Stochastic Models*, 14:205–224, 1998.

[84] G.J.J.A.N. van Houtum, I.J.B.F. Adan and J. van der Wal. The symmetric longest queue system. *Stochastic Models*, 13:105–120, 1997.

[85] J.Q. Hu. Analyticity of single-server queues in light traffic. *Queueing Systems*, 19:63–80, 1995.

[86] R. Iasnogorodski. *Problèmes-Frontières dans les Files d'Attente*. PhD thesis, Université de Paris VI, Paris, 1979.

[87] S. Jaffe. The equilibrium distribution for a clocked buffered switch. *Probability in the Engineering and Informational Sciences*, 6:425–438, 1992.

[88] S. Jaffe. Equilibrium results for a pair of coupled discrete-time queues. Ultracomputer Note, NYA Ultracomputer Research Lab, Courant Institute of Mathematical Sciences, New York, 1989.

[89] N.K. Jaiswal. *Priority Queues*. Academic Press, New York, 1968.

[90] O. Kella and U. Yechiali. Waiting times in the non-preemptive priority $M/M/c$ queue. *Stochastic Models*, 1:257–262, 1985.

[91] J.F.C. Kingman. Two similar queues in parallel. *Annals of Mathematical Statistics*, 32:1314–1323, 1961.

[92] S.J. de Klein. *Fredholm Integral Equations in Queueing Analysis*. PhD thesis, University of Utrecht, 1988.

[93] L. Kleinrock. *Communication Nets; Stochastic Message Flow and Delay*. McGraw-Hill, London, 1964.

[94] L. Kleinrock. *Queueing Systems, Vol. 2: Computer Applications*. Wiley, New York, 1976.

[95] C. Knessl, B.J. Matkowsky, Z. Schuss, and C. Tier. Two parallel queues with dynamic routing. *IEEE Transactions on Communications*, 34:1170–1176, 1986.

[96] C. Knessl, B.J. Matkowsky, Z. Schuss, and C. Tier. Two parallel $M/G/1$ queues where arrivals join the system with the smaller buffer content. *IEEE Transactions on Communications*, 35:1153–1158, 1987.

[97] A.G. Konheim, I. Meilijson, A. Melkman. Processor-sharing of two parallel lines. *Journal of Applied Probability*, 18:952–956, 1981.

[98] G.M. Koole. On the power series algorithm. In O.J. Boxma and G.M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*. CWI, Amsterdam, 1994, pages 139–155. CWI Tract 105 & 106.

[99] G. Latouche and V. Ramaswami. A logarithmic reduction algorithm for quasi-birth-and-deatch processes. *Journal of Applied Probability*, 30:650–674, 1993.

[100] D.-S. Lee. Analysis of a two-queue model with Bernoulli schedules. *Journal of Applied Probability*, 34:176–191, 1997.

[101] K.K. Leung. Waiting time distributions for token-passing systems with limited-k service via discrete Fourier transforms. In P.J.B. King, I. Mitrani, and R.J. Pooley, editors, *Performance '90*, North-Holland, Amsterdam, 1990, pages 333–347.

[102] H. Levy and M. Sidi. Polling systems: Applications, modeling and optimization. *IEEE Transactions on Communications*, 38:1750–1760, 1990.

[103] J.C.S. Lui, R.R. Muntz and R. Richard. Algorithmic approach to bounding the mean response time of a minimum expected delay routing system. *Performance Evaluation Review*, 20:140–151, 1992.

[104] J.C.S. Lui, R.R. Muntz and D. Towsley. Bounding the mean response time of a minimum expected delay routing system: an algorithmic approach. CMPSCI Technical report 93-68, University of Massachusetts, 1993.

[105] V. Malyshev. An analytic method in the theory of two-dimensional random walks. *Sibirski Math. Zh.*, 13:1314–1329, 1972.

[106] L. Massoulié. Large deviations for polling and weighted fair queueing service systems. Report France Télécom-CNET, 1998.

[107] I. Mitrani. The spectral expansion solution method for Markov processes on lattice strips. In J.H. Dshalalow, editor, *Advances in Queueing: Theory, Methods and Open Problems*, CRC Press, Boca Raton, 1995, pages 337–352.

[108] I. Mitrani and B. Avi-Itzhak. A many server queue with service interruptions. *Operations Research*, 16:628–638, 1968.

[109] I. Mitrani and D. Mitra. A spectral expansion method for random walks on semi-infinite strips. In R. Beauwens and P. de Groen, editors, *Iterative Methods in Linear Algebra*. North-Holland, Amsterdam, 1992, pages 141–149.

[110] N.I. Mushkelishvili. *Singular Integral Equations*. Noordhoff, Groningen, 1953.

[111] R.D. Nelson and T.K. Philips. An approximation to the mean response time for shortest queue routing. *Performance Evaluation Review*, 7:181–189, 1989.

[112] R.D. Nelson and T.K. Philips. An approximation for the mean response time for shortest queue routing with general interarrival and service times. *Performance Evaluation*, 17:123–139, 1993.

[113] R. Nelson and A.N. Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 37:739–743, 1988.

[114] R. Nelson and A.N. Tantawi. Approximating task response times in fork/join queues. In E. Gelenbe, editor, *High Performance Computer Systems*, Elsevier Science Publishers B.V., Amsterdam, 1988, pages 157–167.

[115] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins, Baltimore, 1981.

[116] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Transactions on Networking*, 1:344–357, 1993.

[117] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Transactions on Networking*, 2:137–150, 1994.

[118] V. Ramaswami and G. Latouche. A general class of Markov processes with explicit matrix-geometric solutions. *OR Spektrum*, 8:209–218, 1986.

[119] B.M. Rao and M.J.M. Posner. Algorithmic and approximation analysis of the shorter queue model. *Naval Res. Log.*, 34:381–398, 1987.

[120] J.A.C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13:409–426, 1993.

[121] S. Shapiro. The M-server queue with Poisson input and Gamma-distributed service of order two. *Operations Research*, 14:685–694, 1966.

[122] A. Shwartz and A. Weiss. Induced rare events: Analysis via large deviations and time reversal. *Advances in Applied Probability*, 25:667–689, 1993.

[123] J.H.A. De Smit. The queue $GI/M/s$ with customers of different types or the queue $GI/H_m/s$. *Advances in Applied Probability*, 15:392–419, 1983.

[124] J.H.A. De Smit. A numerical solution for the multi-server queue with hyper-exponential service times. *Operations Research Letters*, 2:217–224, 1983.

[125] J.H.A. de Smit. Explicit Wiener-Hopf factorizations for the analysis of multidimensional queues. In J.H. Dshalalow, editor, *Advances in Queueing: Theory, Methods and Open Problems*, CRC Press, Boca Raton, 1995, pages 293–311.

[126] H. Takagi. Queueing analysis of polling models: An update. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, North-Holland, Amsterdam, 1990, pages 267–318.

[127] H. Takagi. Applications of polling models to computer networks. *Computer Networks and ISDN Systems*, 22:193–211, 1991.

[128] H. Takagi. Queueing analysis of polling models: progress in 1990-1994. In J.H. Dshalalow, editor, *Frontiers in Queueing : Models and Applications in Science and Engineering*, CRC-Press, Boca Raton, 1997, pages 119–146.

[129] R.W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, 1989.

[130] P.E. Wright. Two parallel processors with coupled inputs. *Advances in Applied Probability*, 24:986–1007, 1992.

[131] O. Yaron and M. Sidi. Performance and stability of communication networks via robust exponential bounds. *IEEE/ACM Transactions on Networking*, 1:372–385, 1993.

[132] Z.-L. Zhang, D. Towsley and J. Kurose. Statistical analysis of the generalized processor sharing scheduling discipline. *IEEE J. Sel. Areas Comm.*, 13:1071–1080, 1995.

[133] Y. Zhao and W.K. Grassmann. A numerically stable algorithm for two server queue models. *Queueing Systems*, 8:59–79, 1991.

[134] Y. Zhao and W.K. Grassmann. Queueing analysis of a jockeying model. *Operations Research*, 43:520–529, 1995.

[135] Y. Zhu and H. Li. The MacLaurin expansion for a $G/G/1$ queue with Markov-modulated arrivals and services. *Queueing Systems*, 14:125–134, 1993.