Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

O.J. Boxma

Models of two queues: a few new views

Department of Operations Research and System Theory Report OS-R8603 January

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Models of Two Queues: a Few New Views

O.J. Boxma
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

This paper presents a review of results for models of two queues, with an emphasis on mathematical analysis techniques. Two classes of models are investigated: (i) two parallel queues attended by a single server, and (ii) two queues in series.

1980 Mathematics Subject Classification: 60K25, 68M20. 69C40
Key Words & Phrases: survey, two parallel queues, two queues in series.
Note: This report will appear in the proceedings of the International Seminar on Teletraffic Analysis and Computer Performance Evaluation.

1. Introduction

In 1969, Cohen published "The Single Server Queue" [24]. In a sense this monumental work marked the end of an era in queueing theory, in which the emphasis in queueing research had been placed on the exact mathematical analysis of models with *one server* and/or *one queue.* Although Burke's Output Theorem [15] and Burke's Sojourn-time Theorem [16] had opened the way to the analysis of queue lengths and sojourn times in more general *networks* of queues, only a very limited number of network results had become available in the sixties.

Around 1970 successful applications of queueing theory to problems of computer system performance started to appear. Rather simple queueing network models turned out to be able to yield quite accurate predictions of the behaviour of complex computer systems [62,19], thus stimulating queueing network research. It soon became clear that queueing theory presented an extremely powerful tool for the prediction and evaluation of performance of computer communication networks. The acceptance of this belief in computer science circles was strongly enhanced by the presentation in Kleinrock [63].

Presently, the interplay between queueing theory and computer system modeling is a very fruitful one, indeed. On the one hand, deep queueing network results have been obtained in the last decennium, and have been made available for engineering purposes by the introduction of efficient numerical algorithms (see ,e.g., Lavenberg [67]); these algorithms are widely being implemented in versatile packages for the exact, numerical and/or simulation analysis of queueing systems - a first step towards the development of decision support systems for computer system performance evaluation. On the other hand, computer performance modeling constantly gives rise to new, "simple", mathematically intriguing and intricate queueing models, for the analysis of which sometimes new mathematical tools must be developed.

In queueing network theory there is a sharp distinction between networks for which "everything" is known (the extensively reviewed product-form networks) and networks for which little - apart from Little - is known. In the latter case, unless one has admission to unlimited computer (simulation) resources, the use of approximation techniques will be inevitable. In this respect decomposition and aggregation procedures are important, and will be even more important in the future. These procedures naturally lead to the analysis of "mini-networks" of only two, or a few, queues. A mathematical study of a mini-network may not only be useful in itself, it usually gives also considerable insight into the possibility of analysing its generalizations, and into techniques which are suitable for such an analysis. Moreover, such a study may naturally suggest sharp approximations.

Report OS-R8603
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The present paper is devoted to a discussion of models with *two queues* , emphasizing some interesting mathematical techniques for their analysis. It reviews the substantial progress recently made in the analysis of a number of practically important two-queue models. We have not strived for anything near completeness. To some extent we have been led by our own research preferences; in fact some of the results presented are new.

The organization of the paper is as follows. In Section 2 models of *two parallel queues and one server* are studied. Subsection 2.1 also contains some results for more general multi-queue, one-server systems with a cyclic-service discipline. Section 3 considers models of *two queues in series.* The open (tandem) case is discussed in Subsection 3.1, and the closed (cyclic) case in Subsection 3.2; Subsection 3.3 is concerned with an interesting mixed model with one external arrival stream and one closed loop. Sections 2 and 3 both end with a short discussion of variants and extensions, along with a few key references.

2. Two parallel queues served by a single server

2.1 Cyclic service models

Some 15 years ago the analysis of polling schemes, employed to multiplex the service requests of several users in computer-terminal communication systems, gave rise to a new class of queueing models: a single server serves a number of queues in some cyclic fashion. Presently, these queueing models are finding a new application in local area networks with a ring or bus topology, employing a medium access control protocol based on token passing. This section is mainly devoted to such single-server multi-queue models. We consider a number of cyclic-service disciplines, emphasizing the methodologies used for their analysis.

Let us first present a more detailed *model description.* A single server S serves N queues $Q_1, \ldots, Q_N$ (with infinite buffer capacities) in a fixed cyclic order: $Q_1, Q_2, \ldots, Q_N, Q_1, \ldots$. The service strategy at each queue will be specified later. The switch-over times of the server between the i-th and (i+1)-th queue are independent, identically distributed stochastic variables with first moment s_i. The mean of the total switch-over time during a cycle of the server, s, is given by

$$s := \sum_{1}^{N} s_i;$$

its second moment is denoted by $s^{(2)}$. Customers arrive at all queues according to independent Poisson processes with rates $\lambda_1, \ldots, \lambda_N$; the total arrival rate is Λ. Customers who arrive at Q_i are called type-i customers. The service times of type-i customers are independent, identically distributed stochastic variables with distribution $B_i(.)$, with first and second moments β_i and $\beta_i^{(2)}$ and LST $\beta_i(.)$; the service process is also independent of the arrival process and of the switch-over process. The utilization at Q_i, ρ_i, is defined as

$$\rho_i := \lambda_i \beta_i, \qquad i = 1, \ldots, N.$$

The total utilization of the server, ρ, is defined as

$$\rho := \sum_{1}^{N} \rho_i.$$

Several cyclic-service disciplines have been considered, which differ in the number of customers who may be served in a queue during a visit of S to that queue. Assume that S visits Q_i. When Q_i is empty, S immediately starts to switch to Q_{i+1} (we disregard variants in which S does not switch if none of the queues contains customers). Otherwise, S acts as follows, depending on the cyclic-service discipline:

I. Exhaustive service (E): S serves type-i customers until Q_i is empty.
II. Gated service (G): S serves exactly those type-i customers present upon his arrival at Q_i (a gate closes upon his arrival).
III. Nonexhaustive service (NE): S serves only one type-i customer (the generalization to "service of at most K customers" has hardly been analysed, and will also not be considered here).
IV. Semi-exhaustive service (SE): S continues serving type-i customers until the number present is one less than the number present upon his arrival.

In all cases, the order of service within each queue is FCFS.
Before restricting ourself to the case of $N=2$ queues (Subsection 2.2), we shortly consider the most important results valid for an arbitrary number of queues. See Takagi and Kleinrock [87] for an extensive discussion of the E, G and NE disciplines. The SE discipline has recently been introduced by Takagi [86], who studies it in the case where all arrival rates, service-time and switch-over time distributions are the same for all queues; this will in the sequel be denoted as the *completely symmetric case.*

Important quantities in these four single-server multi-queue models are the cycle time, $\mathbf{C}_i$, and the intervisit time, $\mathbf{V}_i$, for Q_i. $\mathbf{C}_i$ is the time between two successive arrivals of S at Q_i, and $\mathbf{V}_i$ is the time between a departure of S from Q_i and his next arrival at this queue. It is well-known, and easily seen (cf. Watson [92]), that for any strict cyclic-service discipline the mean cycle time $E\mathbf{C}_i$ is independent of i, and is given by

$$E\mathbf{C} = \frac{s}{1-\rho}. \tag{2.1}$$

The mean number of type-i arrivals during a cycle is $\lambda_i E\mathbf{C}$; a balancing argument now implies that, in the stationary situation, the mean total *visit* time of S at Q_i during a cycle equals $\rho_i E\mathbf{C}$, and hence the mean *intervisit* time $E\mathbf{V}_i$ equals

$$E\mathbf{V}_i = \frac{s(1-\rho_i)}{1-\rho}, \qquad i=1,...,N. \tag{2.2}$$

Clearly, $\rho<1$ is a necessary condition for ergodicity of these cyclic-service systems. For exhaustive and gated service, this condition is also sufficient. For nonexhaustive service, Kühn [65] has shown that

$$\rho < 1 \quad \text{and} \quad \frac{\lambda_i s}{1-\rho} < 1, \quad i=1,...,N, \tag{2.3}$$

are necessary conditions for ergodicity (indeed, the mean number of type-i arrivals during a cycle should be less than one).
For semi-exhaustive service, the mean number of type-i arrivals during an intervisit time $\mathbf{V}_i$ should be less than one, for during visit times the number of type-i customers is at most reduced by one. This leads to the following necessary conditions for ergodicity for the SE case:

$$\rho < 1 \quad \text{and} \quad \frac{\lambda_i s(1-\rho_i)}{1-\rho} < 1, \quad i=1,...,N. \tag{2.4}$$

Note that these conditions are less strict than those for NE, but stricter than those for E and G (which do not depend on the switch-over process). Also note that, contrary to E and G, in SE and NE some queues can be overloaded, without every queue being overloaded. Condition (2.4) corrects a minor error in [86], formulas (27a) and (28). In particular, in the completely symmetric case we have, with $\lambda_i \equiv \lambda$, $\beta_i \equiv \beta$,

$$\lambda s \frac{1-\lambda\beta}{1-N\lambda\beta} < 1,$$

which after some calculations yields (cf. (27a) of [86]):

$$\lambda < \frac{2}{s+N\beta+\sqrt{(s+N\beta)^2-4s\beta}}. \tag{2.5}$$

Formula (2.5) implies in particular for $N=1$ the intuitively obvious condition:

$$\lambda < min(\frac{1}{s},\frac{1}{\beta}).$$

In the special case of zero switch-over times $\rho<1$ is a necessary and sufficient condition for ergodicity, as is clear from a comparison with the related M/G/1 queue resulting when all customers in the system are served in FCFS order.

In the case of zero switch-over times, it is well-known that a *conservation law* holds for the total amount of work in the system. This amount should not depend on the order of service, and should hence equal the amount of work in an M/G/1 queue with arrival rate Λ and service time distribution the mixture $\sum(\lambda_i/\Lambda)B_i(.)$. Let $E\mathbf{x}_i$ denote the number of type-i customers waiting at an arbitrary epoch, and $E\mathbf{w}_i$ the mean waiting time of type-i customers. The foregoing implies [83] that, regardless of the service discipline, the amount of work of the waiting customers equals:

$$\sum_{i=1}^{N} \beta_i E\mathbf{x}_i = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} - \sum_{i=1}^{N}\rho_i\frac{\beta_i^{(2)}}{2\beta_i} = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)}\rho.$$

Application of Little's formula yields the conservation law (cf. Schrage [83], Kleinrock [63]):

$$\sum_{i=1}^{N}\frac{\rho_i}{\rho}E\mathbf{w}_i = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)}. \tag{2.6}$$

Recently this conservation law has been generalized by Watson [92] to the cases E, G and NE *with* switch-over times (see also [46] for the cases E and G). One can derive a similar formula for SE, following Watson's approach for NE (details will be presented in a forthcoming report). Below we state all four (pseudo-)conservation laws, in a form which is slightly different from Watson's.

$$E: \sum_{i=1}^{N}\frac{\rho_i}{\rho}E\mathbf{w}_i = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{s}{2\rho(1-\rho)}[\rho^2-\sum_{i=1}^{N}\rho_i^2]. \tag{2.7}$$

$$G: \sum_{i=1}^{N}\frac{\rho_i}{\rho}E\mathbf{w}_i = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{s}{2\rho(1-\rho)}[\rho^2+\sum_{i=1}^{N}\rho_i^2]. \tag{2.8}$$

$$NE: \sum_{i=1}^{N}\frac{\rho_i}{\rho}[1-\frac{\lambda_i s}{1-\rho}]E\mathbf{w}_i = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{s}{2\rho(1-\rho)}[\rho^2+\sum_{i=1}^{N}\rho_i^2]. \tag{2.9}$$

$$SE: \sum_{i=1}^{N}\frac{\rho_i}{\rho}[1-\frac{\lambda_i s(1-\rho_i)}{1-\rho}]E\mathbf{w}_i = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}(1-\lambda_i s\rho_i/\rho)}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{s}{2\rho(1-\rho)}[\rho^2-\sum_{i=1}^{N}\rho_i^2]. \tag{2.10}$$

Several comments are in order.

1. (2.7)-(2.10) yield, in the completely symmetric case, Takagi's [86] expressions for the mean waiting times W_E, W_G, W_{NE}, W_{SE}, with $W_E \leq W_G$, $W_{SE} \leq W_{NE}$, and with W_G and W_{SE} not strictly ordered. These comparisons are important for studying the trade-off between total amount of waiting

work and "fairness" (in the E discipline a heavily loaded station can monopolize S; for this reason the NE discipline is presently of greater practical importance).
2. The expressions in (2.7)-(2.10) are independent of the order in which S visits the queues, and they do not involve individual switch-over times.
3. The expressions in (2.7)-(2.10) are extremely useful for obtaining (or testing) approximations for individual mean waiting times (e.g., an approximation of Bux and Truong [18] for E satisfies (2.7), and an approximation in [12] for NE was specifically constructed to satisfy (2.9)).
4. In view of their obvious interest, it is important to really understand (2.7)-(2.10). The interpretation of the right-hand sides is not fully clear, and it is not evident that the right-hand sides for G and NE should be equal; is this some form of conservation?

For SE and NE, this is about all that is known for an *arbitrary* number of queues (see below for N=2 queues). Analysis of E and G turns out to be basically simpler. Let us consider the LST of the waiting-time distribution of a customer at Q_i, $\omega_i(s)$, in these two cases. In the E case, $\omega_i(s)$ can be expressed in the LST $v_i(s)$ of the distribution of the intervisit time $\mathbf{V}_i$ (Eisenberg [42]):

$$\omega_i(s) = \frac{1-v_i(s)}{E\mathbf{V}_i} \frac{1-\rho_i}{s-\lambda_i+\lambda_i\beta_i(s)}. \tag{2.11}$$

After differentiation the mean waiting time $E\mathbf{w}_i$ is expressed in $E\mathbf{V}_i^2$. Ferguson and Aminetzah [46] show that all $E\mathbf{V}_i^2$ can be computed by solving N^2 linear equations (thus significantly improving upon earlier results). Hereto they study the *terminal service time* of Q_i, defined as the time between the server's arrivals at Q_{i-1} and Q_i (for G: between the server's arrivals at Q_i and Q_{i+1}).

The analysis and results for G are very similar. Ferguson and Aminetzah [46] express $\omega_i(s)$ in the LST of the distribution of the cycle time $\mathbf{C}_i$; after differentiation $E\mathbf{w}_i$ is expressed in $E\mathbf{C}_i^2$; these second moments can be computed by solving N^2 linear equations (again exploiting properties of the terminal service times).

The approach sketched above breaks down in the case of zero switch-over times: an idle period of the whole system contains infinitely many cycles of zero length, so mean cycle times and mean intervisit times are zero. Cooper and Murray [30] handle this case (for both E and G) by deriving expressions for the generating functions of the joint queue-length distributions at only those embedded epochs at which the server leaves a queue *after having served at least one customer in that queue.* Cooper [31] shows that $\omega_i(s)$ can be expressed in these generating functions (see also Cooper [32] for a correction for G, and see Ch. 7 of [87]).

Both Cooper [31] and (for nonzero switch-over times) Eisenberg [42] show how one can actually calculate $\omega_i(s)$ by iteratively determining the queue-length generating functions in which $\omega_i(s)$ can be expressed.

2.2 Detailed analysis of two-queue one-server models

This section is devoted to an exposition of the mathematical analysis of the E- and NE disciplines in the case of $N=2$ queues. First Takács' approach [85] to the E discipline is outlined, and subsequently the approach of Cohen and Boxma [23,25] to the NE discipline. In order to let the similarities and differences between the two approaches come out as clear as possible, we restrict ourself to the completely symmetric case without switch-over times (i.p., let $\lambda_i \equiv \lambda$ and $\beta_i(.) \equiv \beta(.)$). We consider the ergodic situation, assuming that $\rho<1$.

In both models, let $\Pi_i(z_1,z_2)$ denote the generating function of the joint queue-length distribution in Q_1 and Q_2 immediately after the departure of S from Q_i, $i=1,2$, and define for $|z_1| \leqslant 1$, $|z_2| \leqslant 1$:

$$\beta(z_1,z_2) := \beta(\lambda(1-z_1)+\lambda(1-z_2)).$$

The E discipline

Takács [85] shows for $|z_1| \leq 1$, $|z_2| \leq 1$:

$$\Pi_1(z_1,z_2) = [\Pi_1(0,z_1) - \Pi_1(0,z_2) - \frac{1}{2}\Pi_0(1-z_1)] \frac{\beta(z_1,z_2)}{z_1 - \beta(z_1,z_2)}, \tag{2.12}$$

and symmetrically for $\Pi_2(z_1,z_2)$; here

$$\Pi_0 = \Pi_1(0,0) + \Pi_2(0,0) = 2\Pi_1(0,0).$$

Takács proceeds to determine $\Pi_1(z_1,z_2)$; once this generating function is known, the waiting-time LST is easily obtained. The analysis leading to $\Pi_1(z_1,z_2)$ consists of three steps.

Step 1: The set-up

According to its definition as a generating function, $\Pi_1(z_1,z_2)$ should be regular for $|z_1| < 1$, continuous for $|z_1| \leq 1$, for every fixed z_2 with $|z_2| \leq 1$; and similarly with z_1 and z_2 interchanged. Hence every zero-pair (z_1,z_2) of the denominator of the rhs. of (2.12), the "kernel"

$$K(z_1,z_2) := z_1 - \beta(z_1,z_2),$$

should make the numerator of the rhs. of (2.12) zero.

Step 2: Analysis of the kernel

Using Rouché's theorem one can prove (cf. Cohen [24]) that, for each z_2, $|z_2| \leq 1$, the kernel $K(z_1,z_2)$ has exactly one root $z_1 = \delta(z_2)$, in $|z_1| \leq 1$, with

$$\delta(z_2) = \gamma(\lambda(1-z_2));$$

here $\gamma(.)$ denotes the LST of the busy-period distribution in an M/G/1 queue with arrival rate λ and service-time distribution B(.).

Step 3: Iteration procedure

Steps 1 and 2 imply that for all $z_1 = \delta(z_2)$, $|z_2| \leq 1$,

$$\Pi_1(0,z_2) = \Pi_1(0,\delta(z_2)) - \frac{1}{2}\Pi_0(1-\delta(z_2)). \tag{2.13}$$

Introduce

$$\delta^{(0)}(z) := z,$$

$$\delta^{(n)}(z) := \delta(\delta^{(n-1)}(z)), \quad n = 1,2,\ldots;$$

it can be shown that, for $\rho < 1$, $\lim_{n\to\infty} \delta^{(n)}(z) = 1$, $|z| \leq 1$. Iteration of (2.13) yields:

$$\Pi_1(0,z_2) = -\frac{1}{2}\Pi_0 \sum_{n=1}^{\infty} (1-\delta^{(n)}(z_2)) + \Pi_1(0,1). \tag{2.14}$$

Substitution of $\Pi_1(0,z_2)$ and $\Pi_1(0,z_1)$ in (2.12) gives us an expression for $\Pi_1(z_1,z_2)$, in which only Π_0 is unknown; the normalization condition (or the observation that Π_0 is not influenced by the order of service) implies that $\Pi_0 = 1-\rho$.

Finally $\omega_i(s)$ can easily be expressed in the above generating functions. In the asymmetric case *two* functions $\delta_1(z)$ and $\delta_2(z)$ occur; the iteration procedure is not significantly more complicated.

So far for Takács' analysis of the E discipline. We refer to Hofri [52] for an interesting extension of the present model to the situation that the server, when ready at one queue, only switches to the other queue if the queue length there exceeds a certain threshold value.

The NE discipline

In the NE case we have:

$$\Pi_1(z_1,z_2) = [\Pi_2(z_1,z_2) + \Pi_1(z_1,0) - \Pi_2(0,z_2) + \frac{1}{2}\Pi_0(z_1-1)]\,\frac{\beta(z_1,z_2)}{z_1}. \tag{2.15}$$

Combination of (2.15) with its symmetrical counterpart for $\Pi_2(z_1,z_2)$ yields:

$$\Pi_1(z_1,z_2) = [\sigma_1(z_1) - \sigma_2(z_2) - \Pi_0(1-\frac{1}{2}(z_1+z_2))\frac{z_2+\beta(z_1,z_2)}{z_2-\beta(z_1,z_2)}] \tag{2.16}$$

$$\frac{1}{2}\beta(z_1,z_2)\frac{z_2-\beta(z_1,z_2)}{z_1z_2-\beta^2(z_1,z_2)}, \quad |z_1| \leq 1,\ |z_2| \leq 1,$$

with

$$\sigma_1(z_1) := 2\Pi_1(z_1,0) + \frac{1}{2}\Pi_0(z_1-1), \tag{2.17}$$

and $\sigma_2(z_2)$ being symmetrically defined. In this symmetric case, in fact, $\sigma_1(z_1) = \sigma_2(z_1)$. The further analysis consists of four steps.

Step 1: The set-up

Similarly as step 1 for the E discipline.

Step 2: Analysis of the kernel

The kernel

$$K(z_1,z_2) := z_1z_2 - \beta^2(z_1,z_2), \tag{2.18}$$

is much more complicated than the kernel for the E discipline. It is no longer possible to determine, explicitly, exactly one zero z_1 in $|z_1| \leq 1$ for each z_2 in $|z_2| \leq 1$. Kernels like the one in (2.18) have been studied, in more generality, in Chapter II.4 of [25]. The theory developed there yields in this special case the following approach.

It turns out to be advantageous, in this symmetric case, to search for pairs of zeros of the kernel which are complex conjugates: $(z_1,z_2) = (w,\overline{w})$. These pairs of zeros turn out to supply all the information we need. The following should hold for w:

$$|w|^2 = \beta^2(2\lambda(1-\mathrm{Re}\, w)).$$

Write

$$w = e^{i\phi}\beta(2\lambda(1-Re\ w)),\ \ 0 \leq \phi \leq 2\pi. \tag{2.19}$$

Hence Re w is for each ϕ determined as the unique zero in Re $\delta \leq 1$ of $\delta - \cos(\phi)\ \beta(2\lambda(1-\delta))$. Finally (2.19) implies that

$$\mathrm{Im}\, w = \tan(\phi)\ \mathrm{Re}\, w. \tag{2.20}$$

It is readily seen that, for $\phi \in [0,2\pi]$, $w = w(\phi)$ once encircles a smooth contour F that is contained in the unit circle, having the point $w=1$ in common with this circle; $0 \in F^+$, the interior of F. For every $w \in F$, $(z_1,z_2) = (w,\overline{w})$ forms a pair of zeros of the kernel $K(z_1,z_2)$.

Step 3: Formulation of a boundary value problem

The basic idea of the approach is to transform the analysis of (2.16) into the analysis of a standard boundary value problem from mathematical physics. Steps 1 and 2 imply that, for all $(z_1,z_2) = (w,\overline{w})$, $w \in F$,

$$\sigma_1(w) - \sigma_1(\overline{w}) = \Pi_0(1-\mathrm{Re}\, w)\frac{\overline{w}+\beta(2\lambda(1-\mathrm{Re}\, w))}{\overline{w}-\beta(2\lambda(1-\mathrm{Re}\, w))}. \tag{2.21}$$

Now the advantage of considering complex conjugate zeros becomes apparent:

$$\sigma_1(\overline{w}) = \overline{\sigma_1(w)}. \tag{2.22}$$

It follows, introducing

$$Q(Re\ w) := \frac{\Pi_0(1-\text{Re } w)}{\beta(2\lambda(1-\text{Re } w)) - \text{Re } w},$$

that for all $w \in F$,

$$\text{Im } \sigma_1(w) = \frac{1}{2}\text{Im } w\ Q(Re\ w). \tag{2.23}$$

We can now formulate a Dirichlet boundary value problem for the determination of $\sigma_1(w)$: determine a function $\sigma_1(w)$, $w \in F^+ \cup F$, with
(i) $\sigma_1(w)$ is regular in F^+, continuous in $F^+ \cup F$;
(ii) (2.23) holds for $w \in F$ (the boundary).

Step 4: Solution of the boundary value problem

In the standard Dirichlet problem formulation the contour involved is the unit circle. The conformal mapping $f(.)$ of F^+ onto C^+ (with as inverse the conformal mapping $f_0(.)$) accomplishes transformation into the standard formulation (this hardly presents an additional problem: the conformal mapping $f_0(.)$ is uniquely determined, and easily numerically evaluated, as the continuous solution of the Theodorsen singular integral equation, cf. [25]). It follows (cf. [48]) that for $w \in F^+$:

$$\sigma_1(w) = \sigma_1(0) + \frac{1}{2\pi}\int_{|y|=1}\left(\frac{1}{y-f(w)} - \frac{1}{y}\right)\text{Im } f_0(y)\ Q(Re\ f_0(y))\ dy. \tag{2.24}$$

After application of the Plemelj-Sokhotski formula [48], $\sigma_1(w)$ is also determined for $w \in F$; in particular we now have the important value $\sigma_1(1)$. Analytic continuation subsequently yields $\sigma_1(w)$ for $|w| \leq 1$. Finally $\Pi_1(z_1,z_2)$ follows from (2.16); the normalization condition implies that $\Pi_0 = 1-\rho$. Mean waiting times are easily evaluated. In this symmetric case their expressions are simple, but in the general case contour integrals occur (however, the "conservation law" (2.9) still holds).

We refer to [23,25] for a (more detailed) discussion of the asymmetric case, and to [9] for an analysis of the case *with* switch-over times.

The foregoing discussion reveals a considerable difference in complexity between the E- and NE cases; a difference which becomes more apparent for $N>2$. E.g., in the completely symmetric case we get, for E, N kernels $z_i - \beta(z_1, \ldots, z_N)$ and, for NE, one kernel $z_1 \cdots z_N - \beta^N(z_1, \ldots, z_N)$. The E analysis can be straightforwardly extended, whereas the NE analysis breaks down; it is not clear how the above two-dimensional analysis can be generalized to higher dimensions.

REMARK 2.1

The reader will have observed that there is considerable freedom in the choice of zero-pairs of the kernel in (2.18). Generally speaking, a careful scrutiny of the structure of the kernel and of the right-hand side of the functional equation under investigation (here (2.16)) should suggest a suitable choice.

REMARK 2.2

Cohen [29] has recently investigated the SE discipline in the asymmetric case without switch-over times, following a similar approach as for the NE discipline. The kernel is

$$z_1 z_2 - \gamma_1(\lambda_2(1-z_2))\ \gamma_2(\lambda_1(1-z_1)),$$

with $\gamma_i(.)$ the LST of the busy-period distribution in an M/G/1 queue with arrival rate λ_i and service-time distribution $B_i(.)$. A suitable choice of zero-pairs of this kernel enables transformation into a Riemann boundary value problem.

2.3 VARIANTS

A. Cyclic-service systems and vacation times

The intervisit time at a queue in a multi-queue single-server system with cyclic service can be viewed as a vacation time for the server w.r.t. this queue. The GI/G/1 queue with vacation times has been the subject of several studies. Skinner [84] is an early reference; we refer to Doshi [40] for further references, and for a beautiful geometric proof (based on sample-path arguments) of the following very general result. Assume that the server in a GI/G/1 queue takes a vacation when becoming idle; when on return from vacation the system is still empty, he takes another vacation, and so on. Under fairly general conditions the stationary waiting time is distributed as the sum of two independent stochastic variables: one corresponding to the waiting time in the same system without vacations, and the other to the stationary forward recurrence time of the vacations.

Of particular interest for cyclic-service systems, and throwing new light on the convolution result in (2.11), is Doshi's extension: in the case of a Poisson arrival process, the above-mentioned result even holds when a vacation time is only independent of the interarrival times during and after this vacation, but *not* necessarily before this vacation.

Lee [68] presents an exact analysis of the M/G/1 queue with vacations and with finite waiting room. The assumption of finite waiting space is also realistic in cyclic-service systems; see Tran-Gia and Raith [88] for an approximate analysis of this model under the NE discipline.

Almost all studies of cyclic-service systems have restricted their attention to the case of only one server. An interesting exception is the approximation study of Morris and Wang [72] for multi-queue systems with multiple cyclic servers. They argue that, from the point of view of obtaining simple, reasonably accurate approximations, the fact that there are multiple servers can actually be helpful because of an "averaging out" effect.

B. Non-cyclic service disciplines

It is interesting to compare results for multi-queue single-server systems with a cyclic-service discipline with those with a priority discipline. Priority disciplines without switch-over times have been extensively studied. Murata and Takagi [74] present one of the few priority studies *with* switch-over times.

Cohen [27] investigates a new priority model which is of practical interest, viz., a two-queue one-server model with priority for the longer queue: after each service completion, the server chooses the first customer of the longest queue. He reduces the analysis to the solution of a Riemann boundary value problem. Although the approach is somewhat similar to that described above for the NE-discipline, the "longer-queue priority discipline" poses several mathematically interesting complications.

C. Other models of two parallel queues

We close this section by mentioning a few practically relevant systems of two parallel queues with *two* servers and some form of interaction between the two parts of the system. These models are generally mathematically hard to analyse, but recently several have been successfully tackled by the technique of transformation into a boundary value problem.

Fayolle and Iasnogorodski in their theses (see also [44]) thus solved the "shortest-queue" problem: two servers, two queues and arriving customers choosing the shortest queue. Note that the "shortest-queue" model is in a sense dual to the "longer-queue priority" model.

Fayolle and Iasnogorodski [43] have also studied a model of two coupled processors: two M/M/1 queues with service speeds dependent on whether the other server is busy or idle. They determine the joint queue-length distribution by reducing the problem to a Riemann-Hilbert boundary value problem. The coupled-processor model is contained in a fairly general two-dimensional model investigated in [45]: a two-dimensional birth-and-death process with birth rates and death rates that are state-dependent in a restricted way. It is of interest to remark that this also includes the following model: two queues and two servers, with all arriving customers choosing the first queue when the first server

is idle. Thus an open problem described by Roque [80] can be considered to be solved.

The coupled-processor model with *general* service-time distributions is analysed in Chapter III.3 of [25]. The stochastic process characterized by the workloads of the servers is investigated. Hence Laplace-Stieltjes transforms occur instead of generating functions, and a Wiener-Hopf type of boundary value problem is formulated instead of a Riemann-Hilbert one. It is conjectured that in general a state-space description by means of workloads instead of that by queue lengths, if possible, will lead to a simpler and more general analysis, since there is less need to assume negative exponential service-time distributions.

The technique of transforming two-dimensional queueing problems into boundary value problems is still in its infancy; a large field of further research is lying open. Already several notorious problems have been solved, and recent numerical experience suggests that efficient numerical solution procedures can be developed on the basis of the theory. We refer to Fayolle [44] for a short exposition of the technique developed by him and Iasnogorodski (which is based on an analysis of the branch points of the algebraic curve(s), determined by the zeros of the kernel(s)). Cohen [26] presents a review of the application of the boundary value technique to a large class of two-dimensional random walks. This class contains several of the queueing models mentioned above; the kernel is specified in less detail than the kernels of those queueing models, but still quite general results can be obtained.

3. Two queues in series

3.1 Tandem queues

In 1954, R.R.P. Jackson [55] analysed an extremely simple model of two queues in series: the output of an M/M/1 queue Q_1, with arrival rate λ and mean service time β_1, forms the input to a single-server queue Q_2; the required service times at Q_2 are independent, negative exponentially distributed stochastic variables with mean β_2, independent of the arrival and service processes at Q_1. Cf. Fig. 1.

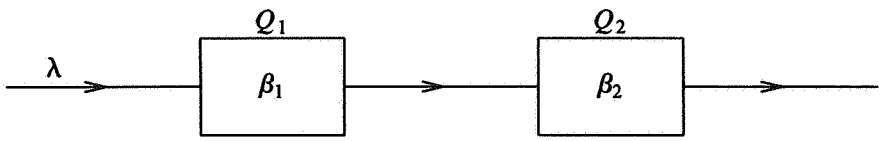

Fig. 1 A tandem model

Jackson obtained the stationary joint distribution of queue lengths x_1 and x_2 at Q_1 and Q_2 by solving the steady-state equations. His surprising result,

$$Pr\{x_1 = n_1, x_2 = n_2\} = (1-\lambda\beta_1)(\lambda\beta_1)^{n_1}\,(1-\lambda\beta_2)(\lambda\beta_2)^{n_2}, \quad n_1,n_2=0,1,..., \tag{3.1}$$

has the following implications:

(i) The queue-length distribution at Q_2 is the same as the queue-length distribution in an M/M/1 queue with arrival rate λ and mean service time β_2;
(ii) The queue lengths at Q_1 and Q_2 are independent.

These results are explained by Burke's [15]

OUTPUT THEOREM

(i) In equilibrium the departure process from an M/M/s queue is a Poisson process, with the same rate as the arrival process;

(ii) The queue length in an M/M/s queue at an arbitrary time t_0 is independent of the departure process prior to t_0.

See Reich [78] for a simple proof based on reversibility, which immediately explains the remarkable second statement.

Reich [78] also studied the joint distribution of a customer's sojourn times (= waiting + service times) at the two queues of Fig. 1; he showed that these sojourn times, too, are independent. This result, in its turn, was explained by Burke's [16]

SOJOURN-TIME THEOREM

The sojourn time of a customer in an M/M/s queue, with service in order of arrival, is independent of the departure process prior to his departure.

The most elegant proof of both the Output Theorem and the Sojourn-time Theorem is based on the reversibility of the state process of the M/M/s queue. A very lucid discussion of the above-mentioned results, their implications and several extensions, is presented in Burke [17]. One extension should be mentioned here: the independence of successive sojourn times remains true in the case of an arbitrary number of stationary M/M/1 queues in series. The first and the last queue may contain multiple servers, but when one or more of the intermediate queues has multiple servers the sojourn time independence is destroyed by the possibility of customers overtaking each other at those queues.

The network extension of the above-mentioned results concerning independence of queue lengths (and the resulting "product form", cf. (3.1)) and concerning independence of sojourn times of a customer at consecutive nodes forms a fundamental chapter in queueing theory. A very short review will be presented in Subsection 3.4. The remainder of the present subsection is devoted to two important new developments concerning two queues in series.

Recently Blanc [4] has studied the *relaxation time* of the model of two M/M/1 queues in series described above. The relaxation time of a system is a measure for the time required to reach ergodicity from an arbitrary initial state. In many practical queueing situations it is crucial to have information concerning the system relaxation time (e.g., for simulation purposes, or in models where underlying parameters do not remain constant over a lengthy period of time). However, the complexity of a time-dependent analysis has caused this to be a neglected subject in queueing theory. Morse [73] seems to have been the first to use the term "relaxation time" in a queueing context; see Cohen [24], Keilson et al. [58] and Blanc and Van Doorn [5] for some exact relaxation-time results for single-server queues, and see Odoni and Roth [76] for an empirical study. Blanc's exact analysis [4] of the tandem model is a very important contribution, as it yields much insight into the relaxation time of a Jackson network (in this respect the interesting study of Massey [70] and a generalization in [89] should also be mentioned).

Blanc determines the transform of the joint time-dependent distribution of the queue lengths at the two M/M/1 queues in series (following the boundary-value approach sketched in Subsection 2.2 above for the NE discipline). First he derives a functional equation for this transform; subsequently he analyses the zeros of the kernel of this equation, using the technique of Fayolle and Iasnogorodski (cf. [44]); the functional equation gives rise to a Riemann-Hilbert boundary value problem, solution of which yields an explicit expression for the sought transform. Thus Blanc finally obtains the following deep results.

Let $\rho_i := \lambda\beta_i$, $T_i := \dfrac{\beta_i}{(1-\sqrt{\rho_i})^2}$, $i=1,2$, and let $p_0(t)$ denote the probability that the whole system is empty at time t, given that it is empty at time 0. Then, as $t \to \infty$,

for $\rho_2<\rho_1<1$,

$$p_0(t) = (1-\rho_1)(1-\rho_2) + \frac{1}{2\sqrt{\pi}} \frac{\rho_1^{\frac{1}{4}}}{(1-\sqrt{\rho_1})^2} \frac{(\sqrt{\rho_1}-\rho_2)^2}{\rho_1-\rho_2} \frac{e^{-t/T_1}}{(t/\beta_1)^{3/2}} [1+O(\frac{1}{t})]; \quad (3.2a)$$

for $\rho_1=\rho_2<1$,

$$p_0(t) = (1-\rho_1)^2 + \frac{1}{2\sqrt{\pi}} \rho_1^{1/4} (1-\sqrt{\rho_1}) \frac{e^{-t/T_1}}{(t/\beta_1)^{1/2}} [1+O(\frac{1}{t})]; \quad (3.2b)$$

for $\rho_1<\rho_2<1$, the expression in (3.2a) should be replaced by one in which all indices 1 and 2 are interchanged. These formulas should be compared with the corresponding one for Q_1, cf. Cohen [24], p. 178. Following Cohen's definition of relaxation time [24], Blanc's results imply that the relaxation time of the tandem model in the ergodic case is:

$$T := max(T_1,T_2), \quad (3.3)$$

i.e., the maximum of the relaxation times of the two M/M/1 systems Q_1 and Q_2. Formula (3.3) leads to the conjecture that the relaxation time of an ergodic system of exponential queues in series is the maximum of the relaxation times of each of the components. Much research will be needed to further investigate this intuitively appealing conjecture and further extensions. The time-dependent behaviour of the tandem system in the *non-ergodic* case is also analysed in [4]; see also Goodman and Massey [50] who, in the non-ergodic case, determine the maximal subnetwork that *does* achieve steady-state. In [3] Blanc discusses the transient behaviour of networks with infinite-server nodes.

Many variants of R.R.P. Jackson's model have been considered: more general arrival process, general service-time distributions at Q_1 and/or Q_2, finite waiting rooms, dependent service times of a customer at both queues. We refer the interested reader to Chapter 7 of Gnedenko and König [49,Vol. II] and Section VI.3 of Disney and König [39]. We restrict ourself here to one recent study that is interesting both from a methodological and a modeling point of view.

Coffman, Fayolle and Mitrani [21] study the total sojourn time of a customer in the simple model of Fig. 1, but with the processor-sharing (PS) discipline at Q_1 and/or Q_2. An interesting aspect of this discipline is that overtaking is possible, so that the remaining sojourn time of a customer can be affected by later arrivals. The reversibility of the system implies that, as far as total sojourn time is concerned, one can restrict oneself to the two cases in which Q_1 has a PS discipline and Q_2 has either a PS- or a FCFS discipline. The analysis in both cases proceeds as follows. Let $\phi_1(n_1,n_2,s)$ be the conditional LST of the remaining sojourn time of a customer in the system, given that the customer is at Q_1 and that there are n_1 other customers at Q_1, and n_2 at Q_2. Introduce the generating function $G(x,y,s)$ of $\phi_1(n_1,n_2,s)$, a function which is regular in the region $|x|<1$, $|y|<1$. Let $\pi(n_1,n_2)$ denote the probability that, upon arrival at the system, a customer finds n_i customers at Q_i, $i=1,2$. It is well-known that $\pi(n_1,n_2)$ is given by (3.1). Defining $\rho_i:=\lambda\beta_i$, $i=1,2$, the LST $\phi(s)$ of the total sojourn time of a customer in the system is given by

$$\phi(s) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \pi(n_1,n_2)\, \phi_1(n_1,n_2,s) = (1-\rho_1)(1-\rho_2)G(\rho_1,\rho_2,s). \quad (3.4)$$

Hence $G(\rho_1,\rho_2,s)$ has to be determined. Unfortunately, for this evaluation it appears to be necessary to determine $G(x,y,s)$ for arbitrary values of x and y, with $|x|<1$, $|y|<1$. $G(x,y,s)$ is shown to satisfy a functional equation, not completely unlike the one found by Blanc [4] in his relaxation-time study. Like Blanc, Coffman et al. carefully study the zeros of the kernel of this functional equation, and use the information provided by these zeros to transform the problem. In the case of one PS and one FCFS server, $G(x,y,s)$ can be obtained by solving (numerically) a Fredholm integral equation of the first type, while solution of two such integral equations is required in the case of two PS servers in tandem.

3.2 Cyclic queues

In this subsection we consider the simplest possible cyclic-queue configuration: N customers cycle through a system of two single-server queues Q_1 and Q_2, requiring services with negative exponentially distributed service times (cf. Fig. 2). Such a system can, e.g., serve as a model of the central part of a multiprogrammed computer system (cf. Allen [1]).

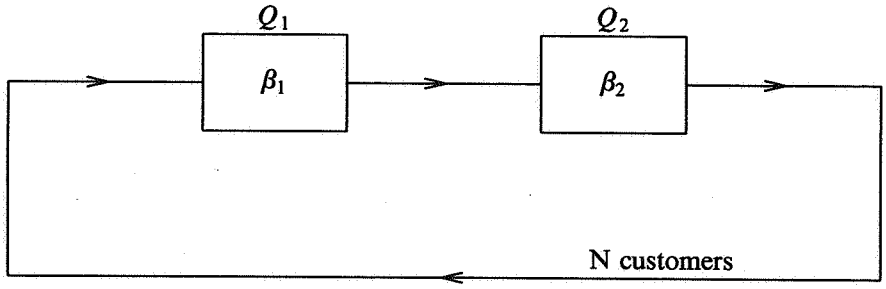

Fig. 2 A cyclic model

It is well-known and easily verified that, if the service discipline in both queues is FCFS, then Q_1 and Q_2 behave like M/M/1 queues with finite capacity N; if Q_1 has a general service-time distribution, then this queue behaves like an M/G/1-N queue. This observation immediately yields the marginal queue-length distribution at Q_1 (both at an arbitrary epoch, a departure epoch and an arrival epoch), and hence - trivially - also the joint queue-length distribution.

Let us turn to the more interesting study of sojourn times. The paper of Chow [20] can be considered as the starting-point for the research of sojourn times in closed queueing networks. Chow studies the model of Fig. 2 with a FCFS discipline at both queues. He obtains, after a lengthy derivation, the distribution of the *cycle time,* i.e., the time between two consecutive departures of a particular customer from Q_2 (or from Q_1, what yields the same result).

Boxma and Donk [6] generalize this result, determining the *joint distribution* of a customer's successive sojourn times at Q_1 and Q_2. They exploit the fact that Q_1 behaves like an M/M/1-N queue. The queue-length process at such a queue is a birth-and-death process and hence a *reversible* process, what leads to the following conclusion: the distribution of a tagged customer's sojourn time at Q_1, conditionally that he has just left k customers *behind,* is equal to the distribution of a tagged customer's sojourn time at Q_1, conditionally that he found k customers present *upon arrival.* One can now easily determine the joint distribution of a tagged customer's successive sojourn times $\mathbf{S}_1$ and $\mathbf{S}_2$ at Q_1 and Q_2, by conditioning on the number of customers he leaves behind in Q_1. In the stationary situation we have, introducing $\rho := \beta_1 / \beta_2$:
for Re $\nu_1, \nu_2 \geq 0$, $\rho \neq 1$,

$$E[exp(-\nu_1 \mathbf{S}_1 - \nu_2 \mathbf{S}_2)] = \sum_{k=0}^{N-1} \frac{1-\rho}{1-\rho^N} \rho^k \left(\frac{1}{1+\beta_1 \nu_1}\right)^{k+1} \left(\frac{1}{1+\beta_2 \nu_2}\right)^{(N-k-1)+1}; \qquad (3.5)$$

(here $\frac{1-\rho}{1-\rho^N}\rho^k$ represents the probability that a departing customer leaves k customers behind in Q_1; for $\rho = 1$, it should be replaced by $1/N$).
Both the derivation and (product-)form of (3.5) are natural analogues of those of the sojourn-time result for the tandem model of Fig. 1. Of course $\mathbf{S}_1$ and $\mathbf{S}_2$ are not independent, as they are in the tandem model, but their dependence has a special structure, inherited from the dependence of the underlying queue lengths. We have, with $\mathbf{Z}$ the number of customers left behind in Q_1 at a departure:

$$cov(\mathbf{S}_1, \mathbf{S}_2) = \beta_1 \beta_2 \, cov(\mathbf{Z}+1, N-\mathbf{Z}) = -\beta_1 \beta_2 Var(\mathbf{Z}) < 0. \qquad (3.6)$$

The result (3.5), revealing a product-form of the LST of successive sojourn times of a customer w.r.t. the underlying queue-length distribution at a departure epoch, has recently been generalized extensively for closed product-form networks (see Subsection 3.4). Just as in the case of sojourn times along a path in an open product-form network, there is one basic restriction: the path should be *overtake-free,* i.e., a customer cannot be overtaken directly by customers behind him, nor indirectly by the influences created by such customers.

Let us, as in Subsection 3.1, turn to some simple two-queue models in which overtaking *is* possible.

I. Queues with processor sharing

Consider the model of Fig. 2, but with processor sharing at both nodes. Now overtaking of customers is possible because of the internal structure of the nodes. Nevertheless, Daduna [37] is able to derive an explicit expression for the LST of the cycle time of a customer. He defines the cycle time of a customer to be the time between two successive departures of this customer from Q_1. He starts (similarly as in several of his other interesting cycle-time studies) by writing down a recursive relation for $f_{2,N,n}(s)$, the LST of the cycle time of a tagged customer, given that this customer finds n customers present at Q_2 upon his arrival. Considering the various jump possibilities, he obtains:

$$f_{2,N,n}(s) = \frac{M(n)}{M(n)+s}\left[\frac{(1-\delta_{n,N-1})/\beta_1}{M(n)} f_{2,N,n+1}(s) + \frac{n}{n+1}\frac{1/\beta_2}{M(n)} f_{2,N,n-1}(s)\right. \tag{3.7}$$

$$\left. + \frac{1}{n+1}\frac{1/\beta_2}{M(n)} f_{1,N,N-1-n}(s)\right], \quad 0 \leqslant n \leqslant N-1,$$

with

$$M(n) := \frac{1}{\beta_2} + (1-\delta_{n,N-1})\frac{1}{\beta_1};$$

$M(n)/(M(n)+s)$ represents the LST of the time until the next jump in the system after the tagged customer's arrival at Q_2, and $f_{1,N,n}(s)$ denotes the LST of the sojourn time of the tagged customer in Q_1, given that he finds n customers present there upon arrival. Daduna first calculates $f_{1,N,n}(s)$ explicitly. Then he shows by induction that

$$f_{2,N,n}(s) = f_{2,N,0}(s)v_n - w_n,$$

where v_n and w_n can now be explicitly determined. $f_{2,N,0}(s)$ can also be determined; finally the LST of the cycle time is a weighted sum of the terms $f_{2,N,n}(s)$.

Although complex, this analysis is considerably simpler than the one in [21] for the *tandem* (open) model of two PS nodes, discussed in Subsection 3.1. It would also be interesting to determine the joint distribution of successive sojourn times, and to investigate the dependence between such sojourn times.

II. Successive cycles

Consider the exponential model of Fig. 2. In some applications it is of importance to have insight into the distribution of the *sum* of a number of consecutive cycle times. In such a situation with more than one cycle, indirect overtaking can occur: if a customer returns to Q_1, he may still find there customers who were waiting behind him at his previous visit to Q_1. Their influence has overtaken him. An explicit expression for the LST of the joint distribution of four consecutive sojourn times of a particular customer in Q_1, Q_2, Q_1,Q_2 has been derived in [8]. The concept of reversibility can again be used, but the analysis is much less straightforward than in [6]. Daduna [35] obtains the LST of the distribution of the sum of r consecutive cycle times. The result is given in a way that makes it suitable for recursive evaluation.

The resulting formulas in [8] and [35] are rather complicated. However, they give rise to some interesting practical observations. Theoretical and numerical results in [8] concerning the dependence

between two successive cycle times suggest that, in most cases, *this dependence is so small that the distribution of the sum of r successive cycle times can be very accurately approximated by the r-fold convolution of the distribution of one cycle time.* In particular it has been shown in [8] that, with $\mathbf{C}_1$ and $\mathbf{C}_2$ two successive cycle times of a tagged customer,

$$\lim_{N\to\infty} correl(\mathbf{C}_1,\mathbf{C}_2) = 0. \tag{3.8}$$

In all cases investigated numerically, $correl(\mathbf{C}_1,\mathbf{C}_2)<0$ (probably a result with quite general validity), and, for fixed N, $correl(\mathbf{C}_1,\mathbf{C}_2)$ takes on its most negative value for $\beta_1=\beta_2$. The extreme value, $correl(\mathbf{C}_1,\mathbf{C}_2)=-0.113$, is found for $N=4$, $\beta_1=\beta_2$. The correlation between cycles which are farther apart can be expected to be even smaller.

The following discussion concerning the (overbearing) influence of the slowest of the two servers on the cycle-time distribution supports the views expressed above (in [11] this discussion is extended to more general networks). Suppose that $\beta_1>\beta_2$; the server at Q_1 is "slower" than his colleague at Q_2. Let $\mathbf{C}$ denote the cycle time of a customer. One can prove, using (3.5), that

$$E[exp(-s\mathbf{C})] = (\frac{1}{1+\beta_1 s})^N [1 + O(\frac{s+1/\beta_1}{s+1/\beta_2})^N], \quad s\geq 0,\ N\to\infty; \tag{3.9}$$

in particular, for $N\to\infty$,

$$E\mathbf{C} = N\beta_1[1 + O((\frac{\beta_2}{\beta_1})^N)],$$

$$E\mathbf{C}^2 = N(N+1)\beta_1^2[1 + O((\frac{\beta_2}{\beta_1})^N)]. \tag{3.10}$$

Numerical results presented in [11] show that replacement of $\mathbf{C}$ by the sum of N service times at Q_1 (as suggested by (3.9)) leads to remarkably small errors, even if β_2 is close to β_1 and N is not very large. The following reasoning yields an intuitive explanation. Let us define the cycle time (from the viewpoint of a customer) as the time between two successive departures of a tagged customer from Q_1; then we can write, considering one cycle from the point of view of Q_1:

$$\mathbf{C} = \mathbf{I}_0+\tau_1+\mathbf{I}_1+\tau_2+\cdots+\mathbf{I}_{N-1}+\tau_N, \tag{3.11}$$

where $\tau_1,\ldots,\tau_N$ and $\mathbf{I}_0,\ldots,\mathbf{I}_{N-1}$ denote the N service times at Q_1 during a cycle, and the N idle periods of the server at Q_1 immediately before those services ($\mathbf{I}_{j-1}>0$ iff Q_1 is empty just before the j-th service). As Q_1 has the slowest server, $Pr\{\mathbf{I}_j=0\}$ will be close to one for all j (in particular if N is large), and $\mathbf{C}$ will be closely approximated by $\tau_1+\cdots+\tau_N$. Returning to the case of several consecutive cycles, we note that the same queue-view argument can be applied here. With an obvious notation,

$$\mathbf{C}_1+\mathbf{C}_2 = \sum_{j=1}^{N}\{\mathbf{I}_{j-1,1}+\tau_{j,1}\} + \sum_{j=1}^{N}\{\mathbf{I}_{j-1,2}+\tau_{j,2}\} \approx \sum_{j=1}^{N}\tau_{j,1} + \sum_{j=1}^{N}\tau_{j,2}, \tag{3.12}$$

unless N is small and $\beta_1\approx\beta_2$; and the two cycle times are almost independent (cf. (3.8)).

REMARK 3.1
The relaxation time in the system of Fig. 2, with FCFS service discipline at both nodes, can be easily obtained using the fact that Q_1 behaves like an M/M/1-N queue. See, e.g., Keilson et al. [58]; their results imply that a sharp peak in the relaxation time occurs for $\beta_1=\beta_2$. This phenomenon, which has also been observed by Tran-Gia (personal communication) is not surprising in view of the above discussion concerning the influence of the slowest server.

REMARK 3.2
In [7] Q_2 is allowed to have a general service-time distribution. An explicit (invertible) expression for

the LST of the joint distribution of the successive sojourn times of a customer in Q_1 and Q_2 has been derived, by again exploiting the connection with the M/G/1-N queue. Subsequently Daduna [34] has obtained a recursive scheme to calculate the LST of the cycle-time distribution; in his model, moreover, the service-time distribution at Q_1 is a mixture of Erlang distributions.

3.3 Finite and infinite source interaction

In Subsections 3.1 and 3.2 the simplest possible open and closed queueing-network configurations have been considered. We now turn to the simplest possible two-queue configuration with one external arrival stream and one closed loop. It is a model of finite and infinite source interaction. Infinite source customers arrive at a single-server queue Q according to a stationary Poisson process with rate λ, requiring service times which are i.i.d. stochastic variables with distribution $B_P(.)$ with LST $\beta_P(.)$ and mean $1/\mu_P$. There is one finite source (fs), having negative exponentially distributed think times with mean $1/\gamma$. The finite-source customer also requires a service at Q; successive service times of the fs customer at Q are i.i.d. stochastic variables with distribution $B_{fs}(.)$ with LST $\beta_{fs}(.)$ and mean $1/\mu_{fs}$, independent of think times and of the Poisson customers. Both customer types thus share Q (cf. Fig. 3). The queueing discipline at Q is FCFS; hence no customer type has priority over the other type.

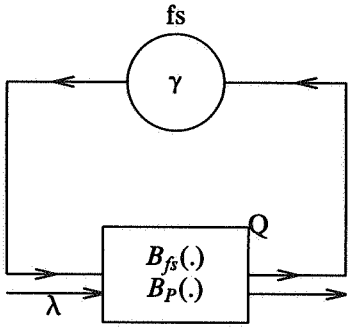

Fig. 3 An interaction model

If $B_P(t) = B_{fs}(t) = 1-e^{-\mu_P t}$, then the network is a simple product-form network. A much more interesting situation arises if the two service-time distributions of the fs customer and the Poisson customers at Q are *not* identical (and negative exponential) - a realistic assumption if the network is used to model, e.g., the interaction of batch traffic and interactive traffic at a CPU (Q). It is now crucial to take the exact position of the fs customer in Q into account, and to know, in particular, whether the customer in service in Q is the fs customer. The above model, with both service-time distributions at Q negative exponential, has been studied by Kaufman [56]. He presents a very accurate approximation method for estimating all mean performance measures of interest, thus obtaining insight in the extent to which the fs customer increases the congestion experienced by the Poisson customers. He allows not just one but N identical finite sources. In [57] he extends his analysis to the case of N heterogeneous finite sources.

Doshi and Wong [41], and, independently, the present author [13] have given an *exact* analysis of the model of Fig. 3. Doshi and Wong allow generalized hyperexponential distributions for the think-time and service-time distribution of the fs customer; the Poisson customers have negative exponentially distributed service requirements. They obtain the generating function of the queue-length distribution at Q *seen by the arriving fs customer.* They also consider the case of LCFS service in Q. In [13] only negative exponential distributions are admitted, but somewhat more general results are

obtained: the joint distribution of queue length at Q and position of the fs customer in the system is determined, leading to exact expressions for various performance measures. The analysis of [13] can be extended to allow completely general service-time distributions for both customer types in Q. The approach for this general case is sketched below; details will be presented in a forthcoming paper.

In the sequel we assume that

$$\rho := \frac{\lambda}{\mu_P} < 1 ; \tag{3.13}$$

(this can be shown to be a necessary and sufficient condition for ergodicity of the system).

Consider successive departure epochs from Q. Let

$$q(x,m) := Pr\{\mathbf{x} = x, \mathbf{n} = x+m\}, \quad x,m = 0,1,..., \tag{3.14}$$

with

$\mathbf{n} :=$ the total number of customers in Q just after a departure from Q,
$\mathbf{x} :=$ the position of the fs customer just after a departure from Q;

here

$\mathbf{x} = 0$ if the fs customer is in its source,
$\mathbf{x} = i$ if the fs customer is in Q in position i (position 1: about to receive service).

Introduce

$$R(w,z) := \sum_{x=0}^{\infty} \sum_{m=0}^{\infty} q(x,m) w^x z^m , \quad |w| \leq 1, \ |z| \leq 1, \tag{3.15}$$

$$S(z) := \sum_{m=0}^{\infty} q(1,m) z^m , \quad |z| \leq 1. \tag{3.16}$$

Our goal is the determination of $R(w,z)$. Recurrence relations for $q(x,m)$ at successive departure epochs lead, in a standard way, to the following formulas: for $|w| \leq 1$, $|z| \leq 1$,

$$R(0,z) = \beta_{fs}(\lambda(1-z))S(z) + \frac{1}{z}\beta_P(\lambda(1-z)+\gamma)[R(0,z)-R(0,0)] \tag{3.17}$$
$$+R(0,0)\frac{\gamma}{\gamma+\lambda}\beta_{fs}(\lambda(1-z))+R(0,0)\frac{\lambda}{\gamma+\lambda}\beta_P(\lambda(1-z)+\gamma),$$

$$[1-\frac{\beta_P(\lambda(1-z))}{w}][R(w,z)-R(0,z)] = -\beta_P(\lambda(1-z))S(z) \tag{3.18}$$
$$+ [R(0,0)\frac{\lambda}{\gamma+\lambda}w+R(0,w)-R(0,0)]\frac{\gamma}{\gamma+\lambda(1-w)-\lambda(1-z)}[\beta_P(\lambda(1-z))-\beta_P(\gamma+\lambda(1-w))].$$

Substitution of (3.17) into (3.18) yields after some calculations:

$$[R(w,z)-R(0,z)] = w[w-\beta_P(\lambda(1-z))]^{-1} \tag{3.19}$$
$$\Big\{[R(0,0)\frac{\lambda}{\gamma+\lambda}w+R(0,w)-R(0,0)]\frac{\gamma}{\gamma+\lambda(1-w)-\lambda(1-z)}[\beta_P(\lambda(1-z))-\beta_P(\gamma+\lambda(1-w))]$$
$$-\frac{\beta_P(\lambda(1-z))}{\beta_{fs}(\lambda(1-z))}R(0,z)[1-\frac{1}{z}\beta_P(\lambda(1-z)+\gamma)]$$
$$-\frac{\beta_P(\lambda(1-z))}{\beta_{fs}(\lambda(1-z))}R(0,0)[(\frac{1}{z}-\frac{\lambda}{\gamma+\lambda})\beta_P(\lambda(1-z)+\gamma) - \frac{\gamma}{\gamma+\lambda}\beta_{fs}(\lambda(1-z))]\Big\}.$$

Once $R(0,z)$ is determined for $|z| \leq 1$, $R(w,z)$ follows from (3.19). Determination of $R(0,z)$ proceeds

as follows (note the close resemblance with the analysis of the E discipline in Subsection 2.2). The denominator of the rhs. of (3.19), the kernel of the functional equation (3.19), becomes zero for

$$w = \delta(z) := \beta_P(\lambda(1-z)). \tag{3.20}$$

For $|z| \leq 1$, clearly $|\delta(z)| \leq 1$. Since the lhs. of (3.19) is regular in $|w| \leq 1$, $|z| \leq 1$, the numerator of the rhs. of (3.19) should be zero for $w=\delta(z)$, $|z| \leq 1$. This condition yields a linear relation between $R(0,z)$, $R(0,0)$ and $R(0,\delta(z))$; for all z with $|z| \leq 1$, we can write with appropriately defined functions $C(z)$ and $D(z)$:

$$R(0,z) = C(z)\, R(0,0) + D(z)\, R(0,\delta(z)). \tag{3.21}$$

$R(0,z)$ has to be determined from the conditions that, for all z with $|z| \leq 1$, (i) it is a regular function of z, and (ii) it satisfies (3.21). An essential role in this analysis is played by $\delta(z)$ and its iterates, defined by:

$$\begin{aligned} &\delta^{(0)}(z) := z\,, \\ &\delta^{(n)}(z) := \delta(\delta^{(n-1)}(z))\,, \quad n=1,2,\ldots\,. \end{aligned} \tag{3.22}$$

One can prove that, for $\rho<1$, $\lim_{n\to\infty} \delta^{(n)}(z)=1$, $|z| \leq 1$. Iteration of relation (3.21) yields (an empty product being one, by definition):

$$R(0,z) = R(0,0) \sum_{i=0}^{\infty} \{C(\delta^{(i)}(z)) \prod_{j=0}^{i-1} D(\delta^{(j)}(z))\} + R(0,1) \prod_{j=0}^{\infty} D(\delta^{(j)}(z))\,, \quad |z| \leq 1. \tag{3.23}$$

The two unknowns $R(0,0)$ and $R(0,1)$ in (3.23) still have to be determined. In this connection note that the functions C(z) and D(z) have exactly one pole z_1, $0<z_1<1$, in $|z| \leq 1$, with

$$z_1 = \beta_P(\lambda(1-z_1)+\gamma). \tag{3.24}$$

One relation between $R(0,0)$ and $R(0,1)$ is obtained by observing that $R(0,z)$ should be regular in $|z| \leq 1$. Using properties of $\delta(z)$ one can show that $\prod_{j=0}^{\infty} D(\delta^{(j)}(z))$ and $\sum_{i=0}^{\infty} \{C(\delta^{(i)}(z)) \prod_{j=0}^{i-1} D(\delta^{(j)}(z))\}$ are well-defined and regular in $|z| \leq 1$, *except* for those z for which a nonnegative integer n exists such that $\delta^{(n)}(z) = z_1$. In particular, $D(z)$, occurring in every term in the right-hand side of (3.23), has a pole at $z=z_1$. Divide both sides of (3.23) by D(z), and subsequently put $z=z_1$; the regularity of $R(0,z)$ in $z=z_1$ now implies that (with appropriate definition of $C(z_1)/D(z_1)$),

$$R(0,0) \left[\sum_{i=1}^{\infty} \{C(\delta^{(i)}(z_1)) \prod_{j=1}^{i-1} D(\delta^{(j)}(z_1))\} + \frac{C(z_1)}{D(z_1)} \right] + R(0,1) \prod_{j=1}^{\infty} D(\delta^{(j)}(z_1)) = 0\,. \tag{3.25}$$

Similarly, divide both sides of (3.23) by $D(z)D(\delta(z))$, and put $z=\delta^{-1}(z_1)$; the analyticity of $R(0,z)$ in $z=\delta^{-1}(z_1)$ again implies relation (3.25). Continuing in the same way, it is seen that Condition (3.25) ensures the analyticity of $R(0,z)$ in all those values of z, for which a positive integer n exists such that $\delta^{(n)}(z)=z_1$ (the fact that one and the same condition takes care of all the singularities is a direct consequence of the structure of (3.21) and of C(.) and D(.)).

Relation (3.25) provides one equation for determining the constants $R(0,0)$ and $R(0,1)$. The second equation is obtained by putting $w=z$ in (3.19), dividing both sides by $z-1$ and subsequently substituting $z=1$. This yields:

$$1-\rho = R(0,1)[1-\rho+\frac{\lambda}{\mu_{fs}}](1-\beta_P(\gamma)) + R(0,0)[\frac{\lambda}{\gamma+\lambda}+\frac{\gamma}{\gamma+\lambda}\beta_P(\gamma)(1-\rho+\frac{\lambda}{\mu_{fs}})]. \tag{3.26}$$

Combination of (3.23), (3.25) and (3.26) yields explicit expressions for R(0,0), R(0,1) and, more generally, R(0,z). Finally, the joint generating function $R(w,z)$ of x and n follows from (3.19).

Numerical evaluation of these expressions requires evaluation of $\sum_{i=0}^{\infty} \{C(\delta^{(i)}(z_1)) \,/ \prod_{j=i}^{\infty} D(\delta^{(j)}(z_1))\}$

and similar terms. It turns out that very accurate results are already obtained when only a few terms of the infinite sum and product are taken - unless ρ is very close to one. The numerical calculations are extremely simple, hardly occupying any computer time.

It is not difficult to evaluate various performance measures, starting from $R(0,z)$. E.g., define λ_{fs}, EL_{fs} and ES_{fs} to be the arrival rate of the fs customer at Q, the mean number of fs customers in Q at an *arbitrary* epoch, and the mean sojourn time of the fs customer in Q, respectively. Clearly

$$\lambda_{fs} = \gamma(1-EL_{fs}). \tag{3.27}$$

Furthermore, the fraction of times that the customer in service in Q is the fs customer equals:

$$\frac{\lambda_{fs}}{\lambda_{fs}+\lambda} = S(1)+R(0,0)\frac{\gamma}{\gamma+\lambda}. \tag{3.28}$$

$S(1)$ can be expressed in $R(0,0)$ and $R(0,1)$ (and hence, using (3.26), only in $R(0,0)$) by putting $z=1$ in (3.17):

$$S(1) = [R(0,1)-\frac{\gamma}{\gamma+\lambda}R(0,0)][1-\beta_P(\gamma)], \tag{3.29}$$

so that λ_{fs}, and hence EL_{fs}, can be expressed in $R(0,0)$. Finally, using Little's formula,

$$ES_{fs} = \frac{EL_{fs}}{\lambda_{fs}} = \frac{EL_{fs}}{\gamma(1-EL_{fs})}. \tag{3.30}$$

One can even obtain the *distribution* of the sojourn time S_{fs}, by relating its LST to the generating function of the distribution of the number of customers left behind by the fs customer in Q. It is also possible to obtain the total mean queue length in Q just after a departure epoch, thus measuring the influence of the finite source on the Poisson customers. Details will be provided elsewhere. See [13] for an extensive analysis of queue lengths at an *arbitrary* epoch in the case of negative exponential distributions of both customer types at Q.

3.4 Extensions

The results of R.R.P. Jackson [55] and of Reich [78] concerning independence of queue lengths and of sojourn times in tandem queues formed the starting-point for one of the most beautiful theories of Queueing. This theory has been exposed in several textbooks and surveys. Therefore we restrict ourself here to a sketch of the main developments. A distinction is made between queue-length results and sojourn-time results.

Queue lengths: product-form results for joint queue-length distributions
J.R. Jackson [53]: open exponential FCFS network, no feedback.
J.R. Jackson [54]: open and closed exponential FCFS networks, state-dependent arrival and service rates.
Gordon and Newell [51]: closed exponential FCFS network.
Baskett et al. [2]: open, closed and mixed networks with exponential FCFS servers, PS nodes, LCFS nodes, IS nodes.
Kelly [59,60]: a very general network structure, which leads to a significant generalization of Jackson's network concept. Kelly fully exploits the concept of time-reversal.
Cohen [22]: a useful generalization of processor sharing. His analysis does not require the often made assumption that the involved distributions have a rational LST.
Surveys can, e.g., be found in Disney [38] (with particular attention for the work of J.R. Jackson and Kelly) and Disney and König [39].

In closed product-form queueing networks with many nodes and/or many customer types, the curse of dimensionality causes the numerical evaluation of performance measures to be a complex problem An excellent discussion of the best available algorithms, and of some sharp approximations based on these algorithms, is given by Lavenberg and Sauer ([67], Chs. 3 and 4). The development of efficient (exact or approximate) algorithms for the analysis of very large product-form networks is an important trend in queueing network theory. Many more interesting results may be expected in the near future.

SOJOURN TIMES: product-form results for the LST of joint sojourn-time distributions
Open networks:
independence of successive sojourn times of a customer along a path has been proved by:
Reich [79]: arbitrary number of M/M/1 queues in series.
Lemoine [69]: tree-like networks of M/M/1 queues.
Walrand and Varaiya [91] (see also Melamed [71]): overtake-free paths in open multi-class Jackson networks. Here, in contrast to earlier papers, the internal customer flows are not Poisson.

Closed networks:
Two parallel developments can be witnessed, starting from [20] and [6], respectively, and discussing cycle-time (and passage-time) distributions and joint sojourn-time distributions, respectively:
Schassberger and Daduna [81] and Boxma, Kelly and Konheim [10], resp.: cyclic system, arbitrary number of queues.
Daduna [33], Kelly and Pollett [61], resp.: overtake-free paths in closed multi-class Jackson networks. These two papers aptly illustrate the strength of the respective techniques of recursive equations for passage-time LST's and of time-reversal.
Finally the most general result is contained in Daduna and Schassberger [36]. Here not only a series of infinite-server nodes is allowed at the beginning and end of a path, but also one multi-server node is allowed between the strings of infinite-server nodes and single-server nodes. Thus a complete analogy between results for open and for closed networks is reached (cf. the discussion of the Sojourn-time theorem in Subsection 3.1).
See Schassberger [82] for a survey, which contains an extensive discussion of the results of [36].

REMARK 3.3
We have so far neglected the important subject of *insensitivity,* i.e., the phenomenon that stationary-state probabilities of some queueing networks are insensitive to the *form* of the service- (and/or arrival-) time distributions, apart from the means of those distributions. A discussion of the insensitivity concept is presented in Disney and König [39], Ch. IV; several references to the important work of the East-German school can here be found. See also the books of Kelly [60] and of Franken et al. [47], which, a.o., contain interesting discussions of necessary and sufficient conditions for insensitivity. Cohen [28] demonstrates a new proof technique for insensitivity. Exploitation of geometric properties of the sample functions of the inherent stochastic processes is shown to lead to an easy proof and a better understanding of the insensitivity of the queue-length distribution in the Engset model (and many other models).

EPILOGUE
In this paper we have discussed some simple-structured queueing network models. Emphasis has been put on new developments (mainly methodological) concerning models of two queues.

Queueing network theory is presently going through a very exciting period. Many interesting developments are taking place, upon most of which we have hardly touched so far. The following techniques and developments, which enhance the possibility of obtaining useful numerical results for large queueing networks, should not be left unmentioned:

- The emergence of standard packages for the analysis of queueing networks by analytic, numerical and simulation methods; a variety of packages is displayed in [77].
- Decomposition and aggregation techniques. Brandwajn [14] casts several approximation techniques based on aggregation and decomposition in a unified framework, and supplies many references.
- Singular perturbation methods. Knessl et al. develop such methods, in [64] and other reports, for the asymptotic analysis of M/G/1-generalizations; their results are superior to those obtained by diffusion approximation.
- The method of first-passage times (cf. Kühn [66]); a promising tool for the numerical study of time-dependent processes (busy periods, sojourn times) in Markovian queueing networks.
- The development of computational probability techniques for the analysis of queueing models (Neuts [75], Tijms [90]).
- The parametric-decomposition approximation method of Whitt, culminating in the software package QNA [93,94]. Whitt characterizes arrival (and service) processes by two or three parameters, and then analyzes individual nodes separately.

References

1. A.O. Allen (1978). *Probability, Statistics and Queueing Theory* (Academic Press, New York).
2. F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios (1975). *Open, closed and mixed networks of queues with different classes of customers,* J. Assoc. Comput. Mach. 22, 248-260.
3. J.P.C. Blanc (1984). *The transient behaviour of networks with infinite server nodes,* In: Performance '84, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 159-174.
4. J.P.C. Blanc (1985). *The relaxation time of two queueing systems in series,* Stochastic Models 1, 1-16.
5. J.P.C. Blanc, E.A. van Doorn (1986). *Relaxation times for queueing systems,* In: Proceedings of the CWI Symposium on Mathematics and Computer Science, eds. J.W. de Bakker, M. Hazewinkel and J.K. Lenstra (CWI Monograph 1, North-Holland Publ. Cy., Amsterdam) pp. 139-162.
6. O.J. Boxma, P. Donk (1982). *On response time and cycle time distributions in a two-stage cyclic queue,* Performance Evaluation 2, 181-194.
7. O.J. Boxma (1983). *The cyclic queue with one general and one exponential server,* Adv. in Appl. Probab. 15, 857-873.
8. O.J. Boxma (1984). *Analysis of successive cycles in a cyclic queue,* In: Performance of Computer-Communication Systems, eds. H. Rudin and W. Bux (North-Holland Publ. Cy., Amsterdam) pp. 293-306.
9. O.J. Boxma (1984). *Two symmetric queues with alternating service and switching times,* In: Performance '84, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 409-431.
10. O.J. Boxma, F.P. Kelly and A.G. Konheim (1984). *The product form for sojourn time distributions in cyclic exponential queues,* J. Assoc. Comput. Mach. 31, 128-133.
11. O.J. Boxma (1985). *Response times in cyclic queues - the influence of the slowest server,* Report 380, Math. Institute, Univ. of Utrecht.
12. O.J. Boxma, B. Meister (1985). *Waiting-time approximations for cyclic-service systems with switch-over times,* Report OS-R8510, Centre for Mathematics and Computer Science.
13. O.J. Boxma (1985). *A queueing model of finite and infinite source interaction,* Report OS-R8511, Centre for Mathematics and Computer Science.
14. A. Brandwajn (1985). *Equivalence and decomposition in queueing systems - a unified approach,* Performance Evaluation 5, 175-186.
15. P.J. Burke (1956). *The output of a queueing system,* Oper. Res. 4, 699-704.
16. P.J. Burke (1968). *The output process of a stationary M/M/s queueing system,* Ann. Math. Statist. 39, 1144-1152.

17. P.J. BURKE (1972). *Output processes and tandem queues,* In: Proc. Symposium on Computer-Communications Networks and Teletraffic, ed. J. Fox (Polytechnic Press, Brooklyn (N.Y.)), pp. 419-428.
18. W. BUX, H.L. TRUONG (1983). *Mean-delay approximation for cyclic-service queueing systems,* Performance Evaluation 3, 187-196.
19. J.P. BUZEN (1971). *Queueing Network Models of Multiprogramming* (Ph.D. Thesis, Harvard Univ., Cambridge (Mass.)).
20. W.-M. CHOW (1980). *The cycle time distribution of exponential cyclic queues,* J. Assoc. Comput. Mach. 27, 281-286.
21. E.G. COFFMAN, JR., G. FAYOLLE AND I. MITRANI (1984). *Sojourn times in a tandem queue with overtaking: reduction to a boundary value problem,* Report AT&T Bell Labs., Murray Hill (N.J.) (to appear in Stochastic Models).
22. J.W. COHEN (1979). *The multiple phase service network with generalized processor sharing,* Acta Informatica 12, 245-284.
23. J.W. COHEN, O.J. BOXMA (1981). *The M/G/1 queue with alternating service formulated as a Riemann-Hilbert problem,* In: Performance '81, ed. F.J. Kylstra (North-Holland Publ. Cy., Amsterdam) pp. 181-199.
24. J.W. COHEN (1982). *The Single Server Queue* (North-Holland Publ. Cy., Amsterdam; 2nd ed.).
25. J.W. COHEN, O.J. BOXMA (1983). *Boundary Value Problems in Queueing System Analysis* (North-Holland Publ. Cy., Amsterdam).
26. J.W. COHEN (1984). *On the analysis of two-dimensional queueing problems,* In: Mathematical Computer Performance and Reliability, eds. G. Iazeolla, P.-J. Courtois and A. Hordijk (North-Holland Publ. Cy., Amsterdam) pp. 17-32.
27. J.W. COHEN (1985). *A two queue, one server model with priority for the longer queue,* Report 379, Math. Institute, Univ. of Utrecht.
28. J.W. COHEN (1985). *On a new derivation of Engset's formula,* Report 394, Math. Institute, Univ. of Utrecht.
29. J.W. COHEN (1985). *A two-queue model with semi-exhaustive alternating service,* Report 395, Math. Institute, Univ. of Utrecht.
30. R.B. COOPER, G. MURRAY (1969). *Queues served in cyclic order,* The Bell System Techn. J. 48, 675-689.
31. R.B. COOPER (1970). *Queues served in cyclic order: waiting times,* The Bell System Techn. J. 49, 399-413.
32. R.B. COOPER (1981). *Introduction to Queueing Theory* (Edward Arnold, London; 2nd ed.).
33. H. DADUNA (1982). *Passage times for overtake-free paths in Gordon-Newell networks,* Adv. in Appl. Probab. 14, 672-686.
34. H. DADUNA (1983). *Two-stage cyclic queues with non-exponential servers: steady-state and cycle time,* Report TU Berlin (to appear in Oper. Res.).
35. H. DADUNA (1983). *Long time cycling in a two-stage cycle: response times in multiprogrammed computer systems with virtual memory,* Report TU Berlin (to appear in Oper. Res.).
36. H. DADUNA, R. SCHASSBERGER (1984). *Sojourn times in queueing networks with multiserver nodes,* Report TU Berlin.
37. H. DADUNA (1985). *The distribution of residence times and cycle times in a closed tandem of processor sharing queues,* In: Messung, Modellierung und Bewertung von Rechensystemen, ed. H. Beilner (Springer-Verlag, Berlin), pp. 127-140.
38. R.L. DISNEY (1981). *Queueing networks,* In: Operations Research: Mathematics and Models, ed. S.I. Gass, AMS Symposia in Applied Mathematics 25 (American Math. Soc., Providence (Rhode Island)), pp. 53-83.
39. R.L. DISNEY, D. KÖNIG (1985). *Queueing Networks: a survey of their random processes,* SIAM Review 27, 335-403.
40. B.T. DOSHI (1985). *A note on stochastic decomposition in a GI/G/1 queue with vacations or set-up*

times, J. Appl. Probab. 22, 419-428.
41. B.T. DOSHI, W.S. WONG (1985). *Exact solution to a simple finite infinite source interaction model,* Report AT&T Bell Labs., Holmdel (N.J.).
42. M. EISENBERG (1972). *Queues with periodic service and changeover times,* Oper. Res. 20, 440-451.
43. G. FAYOLLE, R. IASNOGORODSKI (1979). *Two coupled processors: the reduction to a Riemann-Hilbert problem,* Z. Wahrsch. Verw. Gebiete 47, 325-351.
44. G. FAYOLLE (1984). *On functional equations of one and two complex variables arising in the analysis of stochastic models,* In: Mathematical Computer Performance and Reliability, eds. G. Iazeolla, P.-J. Courtois and A. Hordijk (North-Holland Publ. Cy., Amsterdam) pp. 55-75.
45. G. FAYOLLE, P.J.B. KING AND I. MITRANI (1982). *The solution of certain two-dimensional Markov models,* Adv. in Appl. Probab. 14, 295-308.
46. M.J. FERGUSON, Y.J. AMINETZAH (1985). *Exact results for nonsymmetric token ring systems,* IEEE Trans. on Communications Vol. COM-33, 223-231.
47. P. FRANKEN, D. KÖNIG, U. ARNDT AND V. SCHMIDT (1982). *Queues and Point Processes* (Wiley, New York).
48. F.D. GAKHOV (1966). *Boundary Value Problems* (Pergamon Press, Oxford).
49. B.W. GNEDENKO, D. KÖNIG (1983,1984). *Handbuch der Bedienungstheorie, Vol. I,II* (Akademie-Verlag, Berlin).
50. J.B. GOODMAN, W.A. MASSEY (1984). *The non-ergodic Jackson network,* J. Appl. Probab. 21, 860-869.
51. W.J. GORDON, G.F. NEWELL (1967). *Closed queueing systems with exponential servers,* Oper. Res. 15, 254-265.
52. M. HOFRI (1986). *Two queues and one server with threshold switching,* In these proceedings (North-Holland Publ. Cy., Amsterdam).
53. J.R. JACKSON (1957). *Networks of waiting lines,* Oper. Res. 5, 518-521.
54. J.R. JACKSON (1963). *Jobshop-like queueing systems,* Management Sci. 10, 131-142.
55. R.R.P. JACKSON (1954). *Queuing systems with phase-type service,* Oper. Res. Quart. 5, 109-120.
56. J.S. KAUFMAN (1984). *Finite and infinite source interactions,* In: Performance '84, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 345-359.
57. J.S. KAUFMAN (1985). *A recursive approximation technique for a combined source queueing model,* Report AT&T Bell Labs., Holmdel (N.J.).
58. J. KEILSON, U. SUMITA AND F. MACHIHARA (1985). *The relaxation time of single server queueing systems with Poisson arrivals and hyperexponential/Erlang service times,* In: Proc. ITC 11, ed. M. Akiyama (North-Holland Publ. Cy., Amsterdam).
59. F.P. KELLY (1976). *Networks of queues,* Adv. in Appl. Probab. 8, 416-432.
60. F.P. KELLY (1979). *Reversibility and Stochastic Networks* (Wiley, New York).
61. F.P. KELLY, P.K. POLLETT (1983). *Sojourn times in closed queueing networks,* Adv. in Appl. Probab. 15, 638-658.
62. L. KLEINROCK (1964). *Analysis of a time-shared processor,* Naval Res. Logist. Quart. 11, 59-73.
63. L. KLEINROCK (1975,1976). *Queueing Systems, Vol. I,II* (Wiley, New York).
64. C. KNESSL, B.J. MATKOWSKY, Z. SCHUSS AND C. TIER (1985). *Asymptotic analysis of a state-dependent M/G/1 queueing system,* to appear in SIAM J. Appl. Math.
65. P.J. KUEHN (1979). *Multiqueue systems with nonexhaustive cyclic service,* The Bell System Techn. J. 58, 671-698.
66. P.J. KUEHN (1983). *Analysis of busy periods and response times in queuing networks by the method of first passage times,* In: Performance '83, eds. A.K. Agrawala and S.K. Tripathi (North-Holland Publ. Cy., Amsterdam) pp. 437-455.
67. S.S. LAVENBERG (1983). *Computer Performance Modeling Handbook* (Academic Press, New York).
68. T.T. LEE (1984). *M/G/1/N queue with vacation time and exhaustive service discipline,* Oper. Res. 32, 774-784.

69. A.J. LEMOINE (1977). *Networks of queues - a survey of equilibrium analysis,* Management Sci. 24, 464-481.
70. W.A. MASSEY (1984). *Open networks of queues: their algebraic structure and estimating their transient behavior,* Adv. in Appl. Probab. 16, 176-201.
71. B. MELAMED (1982). *Sojourn times in queueing networks,* Math. Oper. Res. 7, 223-244.
72. R.J.T. MORRIS, Y.T. WANG (1984). *Some results for multi-queue systems with multiple cyclic servers,* In: Performance of Computer-Communication Systems, eds. H. Rudin and W. Bux (North-Holland Publ. Cy., Amsterdam) pp. 245-258.
73. PH.M. MORSE (1955). *Stochastic properties of waiting lines,* Oper. Res. 3, 255-261.
74. M. MURATA, H. TAKAGI (1986). *Mean waiting times in nonpreemptive priority M/G/1 queues with server switchover times,* In these proceedings (North-Holland Publ. Cy., Amsterdam).
75. M.F. NEUTS (1981). *Matrix-Geometric Solutions in Stochastic Models* (The Johns Hopkins Univ. Press, Baltimore).
76. A.R. ODONI, E. ROTH (1983). *An empirical investigation of the transient behavior of stationary queueing systems,* Oper. Res. 31, 432-455.
77. D. POTIER (ed.) (1985). *Modeling Techniques and Tools for Performance Analysis* (North-Holland Publ. Cy., Amsterdam).
78. E. REICH (1957). *Waiting times when queues are in tandem,* Ann. Math. Statist. 28, 768-773.
79. E. REICH (1963). *Note on queues in tandem,* Ann. Math. Statist. 34, 338-341.
80. D.R. ROQUE (1980). *A note on "queueing models with lane selection",* Oper. Res. 28, 419-420.
81. R. SCHASSBERGER, H. DADUNA (1983). *The time for a round-trip in a cycle of exponential queues,* J. Assoc. Comput. Mach. 30, 146-150.
82. R. SCHASSBERGER (1985). *Exact results on response time distributions in networks of queues,* In: Messung, Modellierung und Bewertung von Rechensystemen, ed. H. Beilner (Springer-Verlag, Berlin) pp. 115-126.
83. L. SCHRAGE (1970). *An alternative proof of a conservation law for the queue G/G/1,* Oper. Res. 18, 185-187.
84. C.E. SKINNER (1967). *A priority queuing system with server-walking time,* Oper. Res. 15, 278-285.
85. L. TAKÁCS (1968). *Two queues attended by a single server,* Oper. Res. 16, 639-650.
86. H. TAKAGI (1984). *Mean message waiting time in a symmetric polling system,* In: Performance '84, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 293-302.
87. H. TAKAGI, L. KLEINROCK (1985). *Analysis of polling systems,* Report IBM Japan Science Institute.
88. P. TRAN-GIA, T. RAITH (1985). *Multiqueue systems with finite capacity and nonexhaustive cyclic service,* In: Proceedings of the International Seminar on Computer Networking and Performance Evaluation, eds. M. Akiyama, Y. Takahashi and H. Takagi (North-Holland Publ. Cy., Amsterdam).
89. P. TSOUCAS, J. WALRAND (1984). *A note on stochastic bounds for queueing networks,* Adv. in Appl. Probab. 16, 926-928.
90. H.C. TIJMS (1986). *Stochastic Modelling and Analysis: A Computational Approach* (Wiley, New York).
91. J. WALRAND, P. VARAIYA (1980). *Sojourn times and the overtaking condition in Jacksonian networks,* Adv. in Appl. Probab. 12, 1000-1018.
92. K.S. WATSON (1984). *Performance evaluation of cyclic service strategies - a survey,* In: Performance '84, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 521-533.
93. W. WHITT (1983). *The Queueing Network Analyzer,* The Bell System Techn. J. 62, 2779-2815.
94. W. WHITT (1983). *Performance of the Queueing Network Analyzer,* The Bell System Techn. J. 62, 2817-2843.