Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

O.J. Boxma, H. Daduna

Sojourn times in queueing networks

Department of Operations Research, Statistics, and System Theory   Report BS-R8916   August

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 Research (N.W.O.).

Sojourn Times in Queueing Networks

O.J. Boxma
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

H. Daduna
Institute of Mathematical Stochastics
Department of Mathematics, University of Hamburg
Bundesstrasse 55
2000 Hamburg 13, B.R.D.

From a customer's point of view, the most important performance measure in a queueing network is his sojourn time. This paper presents a survey of sojourn time results in queueing networks. Particular attention is paid to product-form networks, for which exact expressions for the joint distribution of a customer's successive sojourn times along a quasi overtake-free path are discussed. When the non-overtaking condition is violated, or when no product form exists, few analytical results are known. The paper mentions some of these results, as well as approximation techniques for product-form networks. For networks without product form, the emphasis is put on computational techniques and on some quite general approximation methods which have recently been adapted to the sojourn time problem.

1980 Mathematics Subject Classification: 60K25, 68M20, 90B22.
Key Words & Phrases: Sojourn time distribution, product-form network, overtaking.
To appear in: H. Takagi (ed.), Stochastic Analysis of Computer and Communication Systems.

I. Introduction

The emergence of computer and communication networks in modern technological systems has had a major impact on the development of queueing network theory. In the early seventies rather simple queueing network models turned out to be able to give good predictions of the behaviour of complex computer systems. This led to an extremely fruitful interplay between queueing theory and computer-communication system modelling, as a result of which queueing theory is now well established as a powerful tool for the prediction and evaluation of the performance of computer and communication networks. One of the fruits of this interplay for queueing network research has been the theory of so-called product-form networks (sometimes called separable networks). Extending the early work of, in particular, J.R. Jackson, W.J. Gordon and G.F. Newell, a few classes of queueing networks were demarcated for which the joint queue length distribution is of a product form. For these classes closed form expressions for throughput and queue length distributions are known, along with algorithms for the evaluation of these quantities (see Kelly [1979], Lavenberg [1983]).

From a customer's point of view, the probably most interesting performance measure of a queueing network is his sojourn time, the time he spends in the system (or in a part of it). Mean sojourn times can be easily related to mean queue lengths, but often the whole sojourn time *distribution* (quantiles, tail behaviour) is of importance. Generally, determining the sojourn time distribution of a customer poses very complicated problems. An important exception to this rule is provided by the sojourn time distribution of a customer along a path in a product-form network, when this path satisfies certain overtake-free conditions. The Laplace-Stieltjes transform of the joint distribution of the successive sojourn times of a customer along a 'quasi overtake-free' path appears to obey a product form with respect to the underlying queue length product form. The simple and elegant structure of this

Report BS-R8916
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

result should be contrasted with the almost complete lack of knowledge about sojourn time distributions when the non-overtaking condition is violated: The possibility of customers - or their influences - overtaking each other leads to dependencies that usually destroy any hope for an analytic solution. Even worse is the situation for non product-form networks - where almost no explicit results are at hand. However, some quite general approximation techniques are now being adapted to the sojourn time problem in such networks.

This paper presents a survey of sojourn time results in queueing networks. Sojourn times in isolated service stations are not studied, with one or two exceptions (a discussion of the single server queue with feedback is included because a customer can experience several successive sojourns in the system).

The paper is organized in the following way. Chapter II presents an extensive survey of sojourn time analysis in product-form queueing networks. Sections II.1 - II.3 are devoted to a discussion of the above-mentioned sojourn time product-form expressions. The most general result is formulated in Theorem 2.4 in Section II.3. The crucial overtaking condition is discussed at length in Sections II.2, II.3 and II.5. Section II.4 considers the numerical evaluation of the sojourn time formulas. Results on mean (conditional) sojourn times are mentioned in Section II.6. Some general approximation approaches are discussed in Section II.7. The need for such approximations in product-form networks arises if overtaking prevents an exact sojourn time analysis, or if the outcome of an exact analysis is not in a form that is readily accessible for numerical computations. Section II.8 pays attention to a few specially structured models.

Approximation techniques occupy a more prominent place in Chapter III, which is devoted to sojourn times in networks *without* product form. Section III.1 presents a few techniques that are based on the replacement of a non product-form network by one with product form. Other classes of approximations are discussed in Sections III.2 and III.3. Computational methods for the calculation of passage time moments and distributions are the subject of investigation in Section III.4. In particular, the uniformization procedure receives some attention here. Sections III.5 and III.6 are devoted to queueing networks with particular structures of the service times that frequently arise in actual communication networks (identical service times of a customer at successive queues, or deterministic service times); in some cases, this structure allows an exact sojourn time analysis. The chapter is closed with a discussion of some miscellaneous results.

If analytical results are not available, simulation is an important method for obtaining insight in (mean) sojourn times in queueing networks. However, we have decided not to cover that area in this survey. Instead, we refer to the monograph of IGLEHART and SHEDLER [1980].

To provide some additional motivation for, and insight in, the modelling and sojourn time analysis of communication networks, we end this introduction with the description of two (by now classical) examples of the construction of queueing networks for the performance evaluation of data communication systems. In both examples customer delay is the performance measure of main interest.

Example A. A message switching communication network

The following discussion is based on the exposition of KLEINROCK [1964], who has made an extensive and fundamental investigation of the analysis and design of message switching communication networks. A communication network is a network of communication centres connected by channels. A message switching communication network (MSCN) is a network through which messages flow in such a manner that a message is not transmitted out of a centre before the complete message has entered this centre. The finite capacities of the channels (bits per second) and the stochastic nature of the message flow lead to queueing of the messages in the centres, i.e., storage of the messages in a memory. As the queueing of messages is an essential feature of the MSCN, it is natural to think of the MSCN as a network of queues. In such a queueing network, the channels are the single server service facilities and the messages are the customers, service time being equal to message length divided

by channel capacity. KLEINROCK considered an MSCN at which the messages arrive according to a Poisson process, with exponentially distributed message lengths, where the messages are transmitted in the order of their arrival. Still, as he observed, the resulting queueing model is not a JACKSON network, for the following reason: Messages maintain their lengths as they pass through the net, and therefore service times at successive service facilities are strongly related, thus causing in the model the interarrival times and service times at one and the same service station to be correlated. The complexity of this queueing network has led KLEINROCK [1964] to the very important and useful *Independence Assumption:*
'Each time a message is completely received at a centre, a new length is sampled for this message according to a negative exponential distribution with the same mean'.
Now, the resulting queueing model *is* a JACKSON network.
As KLEINROCK [1964], p. 9, remarks, 'the single most significant measure of performance for these nets is the average time for a message to make its way from the origin through the net to its destination' (the average message delay). A large part of his monograph is devoted to the minimization of this performance measure for a variety of communication nets. Delay *distributions* are not considered in depth for the MSCN. Chapter II of the present study pays much attention to this problem for the simplified JACKSON network with independent service times (and for more general product-form networks).

A remarkable result that follows from the Independence Assumption is obtained if one investigates a series connection of J channels, neglecting the interplay with the rest of the network. Assuming the message arrival stream to be Poisson(λ) and the channels to be single server queues with exponential service time distributions with mean $1/\mu_j$, $j=1,...,J$, it follows from Theorem 2.1 below that the total message delay T has as distribution the convolution

$$Pr\{T<t\} = (1-e^{-(\mu_1-\lambda)t})* \cdots *(1-e^{-(\mu_J-\lambda)t}), \quad t\geq 0. \qquad (1.1)$$

For the case of all channels having identical capacities this reduces to the Erlang-J distribution.

It is clear that the Independence Assumption does not correspond to the actual situation in any practical communication net, but, as observed by KLEINROCK, in networks with considerable mixing of the traffic streams, the correlation between service times and interarrival times at one and the same service station almost disappears. His simulation results seem to justify the introduction of the Independence Assumption for this kind of network. On the other hand, application of the Independence Assumption in tandem queues or in networks with a strong 'tandem queue character' may lead to large errors. We return to this model in Section III.5, where we discuss the few exact results that are available for networks with dependent service times of a customer at successive queues.

Example B. Window flow control in communication networks
Again consider a communication network of switching centres which are connected by uni-directional channels. Once more, make the exponentiality and independence assumptions. If too many messages lay a claim on the available resources in the network, *flow control procedures* are needed to prevent the system from becoming overloaded. We restrict ourselves here to *global flow control,* i.e., to procedures that regulate the externally offered traffic. The principle of global flow control is to shift congestion from the interior of the network to the points of traffic admittance. *Window flow control* is a form of global flow control that is being exercised on so-called virtual circuits of the network (a virtual circuit is a fixed route, along which messages are transmitted between a particular sender and a particular receiver). The *sliding-window* protocol on a virtual route operates in the following way (REISER [1979,1982]).

(i) For each message dispatched on the virtual route, a counter n (initially set to N, called the window size), is decremented by one.
(ii) If the counter is zero, no new message is admitted to the virtual route.
(iii) Each properly received message at the destination is individually acknowledged. As

acknowledgments return, the sender's counter is incremented.

As observed by REISER [1979], the sliding-window protocol keeps the total number of (messages in transit) + (acknowledgments in transit) + (virtual sender counter) equal to the window size N. The essential observation is that the flow control protocol transforms an open queueing system into a closed cyclic one; cf. Fig. 1.1, where node 1, with number of customers equal to the counter n, represents the source with arrival rate μ_1.

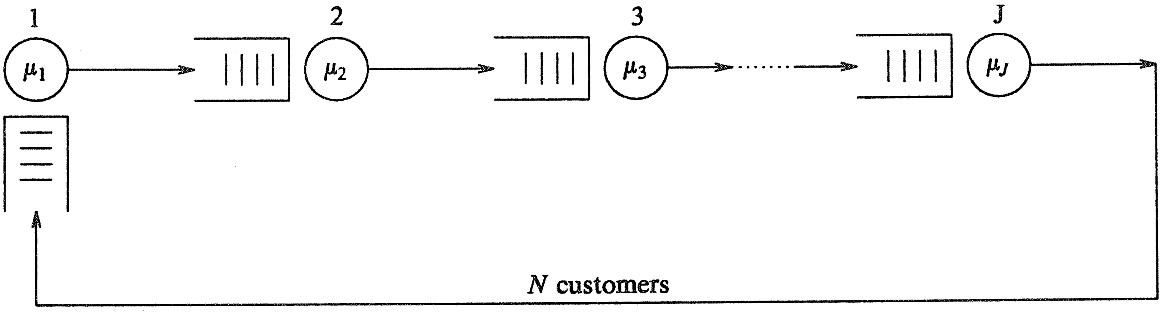

Fig. 1.1

To be more precise, the use of sliding-window protocols on the virtual routes gives rise to a closed queueing network consisting of multiple cyclic routing chains (a product-form GORDON-NEWELL network). Therefore, study of the passage time distribution of a customer (message) along a route is important for the performance evaluation of control procedures.

In Chapter II the passage time distribution of a customer along a path in a closed network is extensively studied. Just as for Example A, we present a simple closed form passage time expression, this time in the form of the Laplace-Stieltjes transform of the distribution of the time T it takes a customer to complete one cycle in the model of Fig. 1.1, consisting of J single-server queues with exponential service time distributions (cf. Theorem 2.3):

$$E[\exp(-\Theta T)] = \sum_{n_1+\cdots+n_J=N-1} \pi(n_1,\ldots,n_J) \prod_{j=1}^{J} \left(\frac{\mu_j}{\mu_j+\Theta}\right)^{n_j+1}, \quad \Theta \geqslant 0, \tag{1.2}$$

in which μ_j is the service rate at the jth queue and $\pi(n_1,\ldots,n_J)$ is the (product-form) queue length distribution of the other customers as seen by a customer at the moment of his arrival at a queue. Formula (1.2) can be inverted to give an explicit expression for the cycle time density (cf. Corollary 2.1 in Section II.4). Mean transit delays can be easily obtained from the above formulas, or via application of the Mean Value Analysis algorithm.

REMARK 1.1.

For simplicity, in (1.2) we have not only assumed that the virtual route is closed for a particular class of customers, but also that there is no interference of other routes (other cyclic routing chains) passing through one of the nodes of that particular route. Using Theorem 2.4 of Section II.3 one can show the following (cf. DADUNA [1989a]). If channel j also has to handle messages from other chains, arriving at node j with rate γ_j, then (1.2) still holds with μ_j replaced by $\mu_j-\gamma_j$. For *queue lengths* in product-form networks this property of 'adjusted service rates' is well known, cf. REISER [1982], pp. 189-190.

REMARK 1.2.

Let us mention here that a similar principle of congestion control is used in the area of computer architecture: Using a maximal level of multiprogramming in a computer system with virtual memory and paging is a window flow control mechanism. For details see, e.g., REGE and SENGUPTA [1988] and SALZA and LAVENBERG [1981].

II. Product-form networks

II.1. Classical results for open tandem systems

We consider an open J-stage tandem of exponential single server nodes (e.g., communication channels) with first-come-first-served (FCFS) queueing discipline. Customers arrive at node 1 in a Poisson(λ) stream, after being served there they proceed directly (without time lag) to node 2, and so on. After being served at node J they leave the system. The service times at node j are exponentially distributed with mean μ_j^{-1}, $j=1,\ldots,J$, to be denoted by exp(μ_j), and all service times constitute a family of independent random variables being independent of the arrival process (cf. Example A in Chapter I).

Let $X_j(t)$ denote the total number of customers present at node j at time t, $j=1,\ldots,J$, and $X(t)=(X_1(t),\ldots,X_J(t))$, $t\in\mathbb{R}$. Then $X=\{X(t),\ t\in\mathbb{R}\}$ is a strong Markov process. To guarantee ergodicity of X we assume $\lambda<\mu_j$, $j=1,\ldots,J$. The unique equilibrium distribution of X is

$$\pi(n_1,\ldots,n_J) = \prod_{j=1}^{J}\left[1-\frac{\lambda}{\mu_j}\right]\left[\frac{\lambda}{\mu_j}\right]^{n_j},\ (n_1,\ldots,n_J)\in\mathbb{N}^J\,. \tag{2.1}$$

Reich [1957, 1963] proved the following theorem on sojourn time distributions in queueing networks:

Theorem 2.1.
If the tandem system is in equilibrium, then the successive sojourn times of a customer at the nodes of the system are independent and exponentially distributed, with mean $1/(\mu_j-\lambda)$ *at node* j, $j=1,\ldots,J$.

In this survey we generally state the results without proof. Below we make an exception, as Reich's proof considerably adds to the insight into sojourn time results in product-form networks.

Proof
In his proof Reich utilizes that the input stream of the system is Poissonian, that $X_1=\{X_1(t),\ t\in\mathbb{R}\}$ is a birth-death process, and that customers pass through the system in a fixed order (the latter property means: customers can not overtake one another during their passage through the tandem).

For simplicity let us assume $J=2$. Observing that a birth-death process in equilibrium is reversible, Reich concluded:

Let $\bar{X}_1$ be the process obtained from X_1 by reversing the time scale. $\bar{X}_1$ describes an $M/M/1/\infty$ queueing system in equilibrium. From reversibility it follows that X_1 and $\bar{X}_1$ are stochastically identical and so the arrival process of the $\bar{X}_1$-system is a Poisson(λ) process such that the following holds: if t_0 is an arrival instant then the arrival process *after* t_0 is independent of $\bar{X}_1(t_0)$. But the arrivals of $\bar{X}_1$ are the departures of the original X_1-system which are therefore Poisson(λ), and the arrivals of $\bar{X}_1$ *after* t_0 are the departures of X_1 *before* $-t_0$, and so these departures are independent of $X_1(-t_0) = \bar{X}_1(t_0)$. (These properties usually are summarized as "Output Theorem", which was first proved by Burke [1956] using different methods.) Consider some arbitrary customer C; denote by T_1, T_2 his sojourn times at nodes 1,2 and let t_0 be the instant of his departure from node 1. Then T_2 depends only on $X_2(t_0+)$. But $X_2(t_0+)$ depends only on the service process at node 2 and the departure process from node 1 until t_0 which by the output theorem is independent of $X_1(t_0)$. We have

$$\Pr\{X_1(t_0)=n_1 \mid T_1=s,\ X_2(t_0)=n_2\} = e^{-\lambda s}\,\frac{(\lambda s)^{n_1}}{n_1!}\,,$$

and therefore

$$E\left[z^{X_1(t_0)} \mid X_2(t_0)=n_2\right] = \int_{[0,\infty)} e^{-\lambda s(1-z)}d\Pr\{T_1\leq s \mid X_2(t_0)=n_2\},\ \ |z|\leq 1\,. \tag{2.2}$$

From the output theorem the LHS does not depend on $\{X_2(t_0)=n_2\}$, so the same holds for the RHS, which is therefore the Laplace-Stieltjes transform (LST) of the distribution of T_1. So T_1 is uniquely

defined by $X_1(t_0)$ and is therefore independent of T_2. That $T_1 \sim \exp(\mu_1-\lambda)$ holds is obvious, and $T_2 \sim \exp(\mu_2-\lambda)$ follows directly from the output theorem.
The case $J>2$ follows by similar arguments and an induction. This completes the proof.

An interesting comment on Reich's sojourn time theorem is the following observation of Burke [1964]: in equilibrium the successive *waiting times* of a customer at the nodes are *dependent!*

Burke [1968] did the next step: Starting from his and Reich's observation that the output theorem holds for the first node of the tandem even if this would be a multiserver node having m_1 service channels, by some involved computations he proved that the successive sojourn times of a customer at the nodes in equilibrium are independent (Burke's *sojourn time theorem).* He further gave a second proof of this theorem by utilizing again reversibility of the queue length process of node 1.

Because obviously the last node of the tandem is allowed to be another multiserver queueing system (even with non-exponential service time distribution), at that time the sojourn time problem for the following "network" in equilibrium was solved:

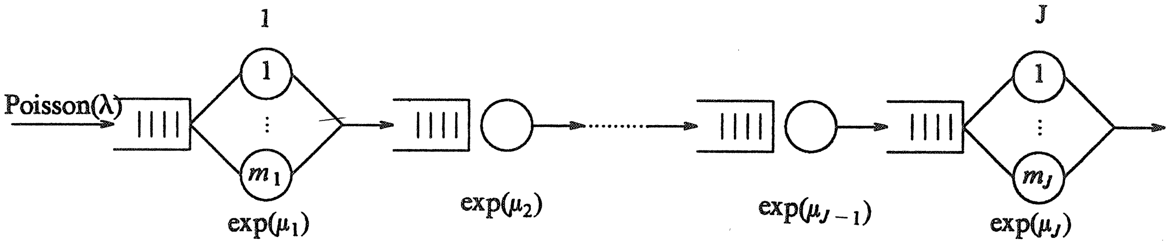

Fig. 2.1.

Additionally Burke [1969] showed that the requirement of having single server nodes in stages $2, 3, \ldots, J-1$ is *necessary* to obtain independence of the successive sojourn times of a customer. Krämer [1973] investigates the dependence of the successive waiting times of a customer in a system consisting of an $M/M/1$ queue in series with an $M/M/s$ queue. In particular, he calculates the distribution of the total time a customer spends waiting in the two queues.

The sojourn time theorem for the tandem model of Fig. 2.1. remained the state-of-the-art until 1979/80; for an intermediate review see Burke [1972]. The situation concerning *closed* queueing networks, especially closed cycles, was even more unsatisfactory, as is reflected by the almost complete lack of references on sojourn times for such networks in the survey paper of Koenigsberg [1982] on cyclic queues (and generalisations of it).

II.2. The non-overtaking condition

At the end of the seventies the discussion about analytical results on sojourn times in more general networks was re-opened by Lemoine [1977] (see also Mitrani [1979]) and Wong [1978]. They pointed out that the essentials used in the proofs of Burke and Reich seemed to be the Poissonian nature of the traffic processes in the network and (following from the output theorem) the independence of sojourn time in a node and queue length at the subsequent nodes at the end of that sojourn time. But surprisingly enough it turned out that these ideas could only be applied to the case of networks having a tree structure (Lemoine [1979]).

The research that followed, and that will be described in the next section, suggests that the main ingredients of the sojourn time theorem are:

— the above-mentioned possibility to compute the joint distribution of (i) a customer's sojourn time in the first node of his itinerary and (ii) the joint queue length processes at his departure instant from this node; but also

— the overtake-free property of the paths the customers have to traverse - with respect to both the topological structure of the network and the single-server FCFS structure of the nodes in the inner part of the path.

Let us now discuss the overtake-free condition which was introduced independently by Walrand and Varaiya [1980] and Melamed [1982]. This property - as they defined it - is concerned only with the topological structure of the network and the possible routing of customers, which may depend on the types of the customers. We shall sketch the main idea in the context of the paper of Walrand and Varaiya [1980].

They consider an open multiclass Jackson network of $\underset{\sim}{M}/\underset{\sim}{M}/1/\infty$ nodes with FCFS queueing discipline. There are nodes $\underset{\sim}{J}=\{1,\ldots,J\}$ and customer types $\underset{\sim}{L}=\{1,\ldots,L\}$. Customers of type $l\in\underset{\sim}{L}$ arrive at node $i\in\underset{\sim}{J}$ in a Poisson stream of intensity $\gamma_i^l\geq 0$; the arrival streams are independent. The routing of the customers is Markovian: a customer of type l leaving node i becomes a type-k customer and either proceeds to node j with probability $r(l,\,i;\,k,\,j)\geq 0$, or leaves the network with probability $r(l,\,i;\,k,\,0)\geq 0$ (it is assumed that $r(l,\,i;\,k,\,i)=0$). The service times at node $i\in\underset{\sim}{J}$ are type-independent $\exp(\mu_i)$ distributed; all service times form an independent family, independent of anything else.
We have the following

Definition 2.1. (Walrand and Varaiya [1980])

(i) For $i,\,j\in\underset{\sim}{J}$ write $i\rightarrow j$ if $r(l,\,i;\,k,\,j)>0$ for some $l,k\in\underset{\sim}{L}$.

(ii) For $i,\,j,\,k\in\underset{\sim}{J}$ let
$P(i,\,j)=\{(i,\,i_1,\ldots,i_m,\,j):\,i_1,\ldots,i_m\in\underset{\sim}{J},\ i\rightarrow i_1,\ i_1\rightarrow i_2,\ldots,\ i_{m-1}\rightarrow i_m,\ i_m\rightarrow j\}$ be the set of "paths" from i to j, and let
$P(i,\,k,\,j)=\{(i,\,i_1,\ldots,i_m,\,k,\,i_{m+1},\ldots,i_n,\,j):$
$i_1,\ldots,i_n\in\underset{\sim}{J},\ i\rightarrow i_1,\ i_1\rightarrow i_2,\ldots,i_m\rightarrow k,\ k\rightarrow i_{m+1},\ldots,i_{n-1}\rightarrow i_n,\ i_n\rightarrow j\}$
be the set of "paths" from i to j going through k.

(iii) The path $(i_1,\ldots,i_m)\in P(i_1,\,i_m)$ *permits no overtaking* (is overtake-free) if
$P(i_u,\,i_v)\subseteq P(i_u,\,i_{u+1},\,i_v)$ for all $1\leq u<v\leq m$.

Let us point out that a "path" from i to j is not necessarily any customer route. What happens on an overtake-free path is: Neither any customer B behind some customer A on the path can overtake A physically, nor can influences generated by B on a place behind A on the overtake-free path overtake A in a relay-race manner.

Examples:

(A) Only one customer type in the network;

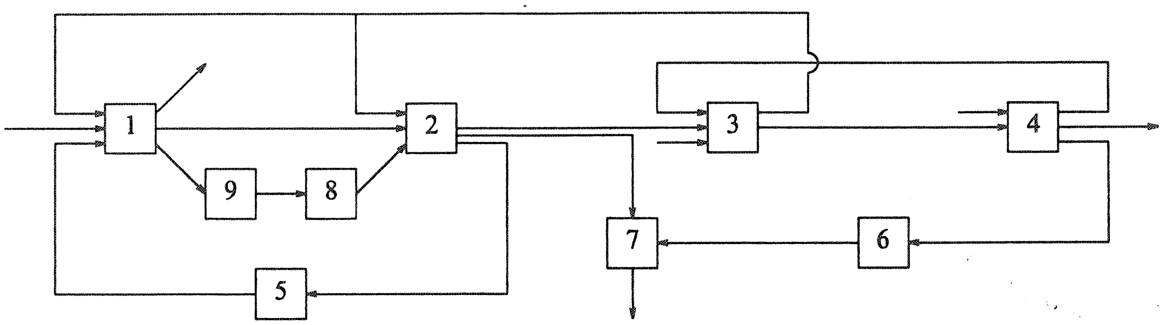

Fig. 2.2

(1, 2, 3, 4) is overtake-free. Generally in single-type networks with FCFS single server nodes we have: A path is overtake-free if and only if no physical overtaking can happen. So (1, 2, 5) is overtake-free, (1, 9, 8, 2) is not overtake-free.

(B) Only one customer type in the network;

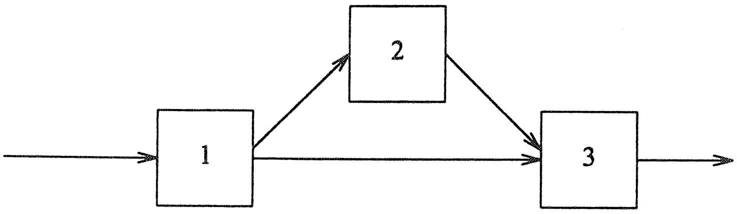

Fig. 2.3.

Path (1, 2, 3) is not overtake-free, while paths (1, 2), (1, 3), (2, 3) are overtake-free. Generally we have: Any path consisting of only two nodes which are visited subsequently is overtake-free. This network is known as SIMON-FOLEY network, see SIMON and FOLEY [1979].

(C) Customer types: 1 ______ ; 2 - - - - - - - ; 3 .... ; 4 —.—.—.—.

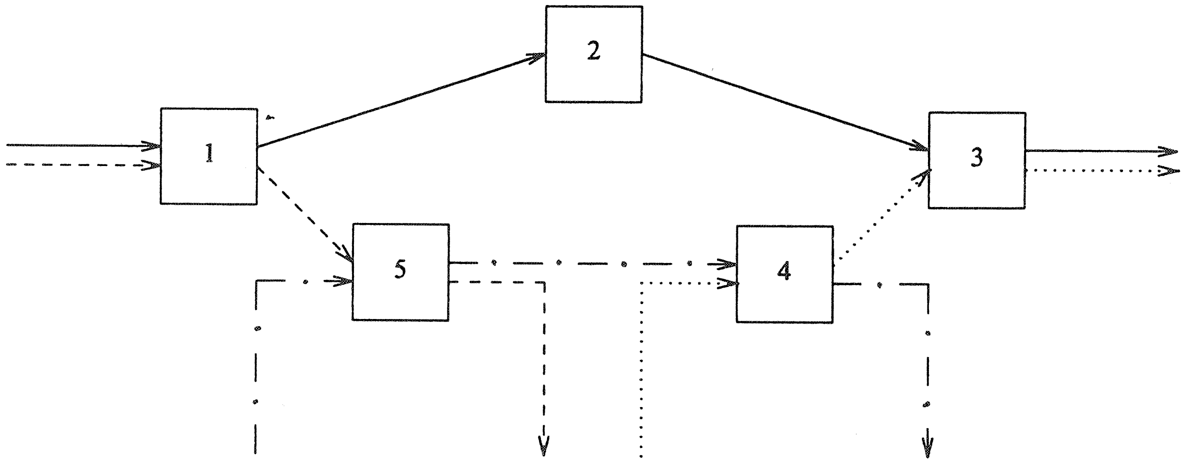

Fig. 2.4.

Path (1, 2, 3) is not overtake-free although customers on this path cannot be physically overtaken.

Now the important result of WALRAND and VARAIYA [1980] for the open multi-class Jackson network described above is:

THEOREM 2.2.
Let $(i_1, i_2, \ldots, i_n)$ *be an overtake-free path in the network, and C a customer who traverses this path. If the network is in equilibrium then the successive sojourn times of C at the nodes* $i_1, i_2, \ldots, i_n$ *are independent and* $\exp(\mu_{i_m} - \lambda_{i_m})$ *distributed, where* λ_{i_m} *is the total arrival rate at node* i_m, $m = 1, \ldots, n$.

In the light of the above-mentioned results of REICH, BURKE and LEMOINE this theorem is surprising because the customer streams on the overtake-free paths may be strongly non-Poissonian (MELAMED [1979]). Additional proof that Poissonian traffic flows do not form the characteristic feature of the sojourn time results is given by SIMON and FOLEY [1979] and WALRAND and VARAIYA [1980]; they proved that the sojourn times on the path (1, 2, 3) in the network of example (B) are not independent although all customer streams in the network are Poissonian.

In a certain sense the nice and simple-to-state results of Burke, Reich, Lemoine, Walrand and Varaiya seemed to discourage any successful approach to the sojourn time problem in *closed* queueing networks. Indeed, each time when an explicit result was found it requested independence of the sojourn times for computing the sojourn time distribution. But due to the finite fixed number of customers in the closed network even the queue lengths are not independent - and obviously dependencies hold for successive sojourn times. The nice result of Chow [1980] on a two node closed exponential cycle seemed to support this suggestion. He proved that the cycle time distribution is a mixture of two Erlang distributions. The resulting expression bore no resemblance to the open network result.

II.3. The general product-form result

As it turned out, the paper of Chow [1980] was the starting point for a sequence of papers which developed step by step, and using different methods, the analytical theory for computing sojourn time distributions in closed queueing networks: Boxma and Donk [1982] computed the joint sojourn time distribution in Chow's model, Schassberger and Daduna [1983] derived the cycle time distribution for a J-node tandem, $J \geqslant 2$, Boxma, Kelly and Konheim [1984] the joint sojourn time distribution for the successive sojourns of a customer during one cycle. The cycle time distribution in a tree-like Gordon-Newell network of single server queues was given by Harrison [1984a]. The passage time distribution for an overtake-free path in a single server Gordon-Newell network was found by Daduna [1982] and the joint distribution for a customer's sojourn times at the nodes of an overtake-free path in such networks was computed by Kelly and Pollett [1983] - thus arriving at a stage where open and closed network theory were developed almost to a similar extent.
To demonstrate some of the central ideas of the development and the fundamental difference with the proofs for the open tandem system we sketch the ideas in the paper of Boxma, Kelly and Konheim [1984].
The model is as follows: N identical customers are cycling in a closed cycle of J exponential single server FCFS queues (with infinite waiting room), i.e., every customer successively proceeds through nodes $1,2,\ldots,J$, $1,2,\ldots$. The service times at node j are $\exp(\mu_j)$ distributed, $j=1,\ldots,J$, and all service times are independent. We assume the system to be in equilibrium.
For some customer C let $T_1,T_2,\ldots,T_J$ denote the sequence of successive sojourn times at nodes $1,2,\ldots,J$ during one cycle. In this cycle we concentrate on the instant when C departs from node 1 proceeding to node 2. From the equilibrium assumption it follows that in this instant C finds exactly n_j other customers at node j, $j=1,\ldots,J$, $n_1+\ldots+n_J=N-1$, with probability

$$\pi^{(N-1)}(n_1,\ldots,n_J) = G(N-1,J)^{-1}\prod_{j=1}^{J}\mu_j^{-n_j}\ ;$$

indeed, the arrival theorem of Sevcik and Mitrani [1981] and of Lavenberg and Reiser [1980] implies that the steady state arrival distribution of the other customers which C sees when he leaves node 1 equals the system's steady state distribution for population size $N-1$; $G(N-1,J)$ is the normalizing constant for the steady state distribution of the system if there are $N-1$ customers cycling.
Hence the distribution of Y_1, the number of other customers left behind at node 1 by C on his departure to node 2, is known, and we can condition the Laplace-Stieltjes transform of the joint distribution of $T_1,\ldots,T_J$ on Y_1 in the following way:

$$E\Big[\exp\left(-\Theta_1 T_1-\cdots-\Theta_J T_J\right)\Big]=$$

$$\sum_{n_1=0}^{N-1} P(Y_1=n_1)\, E\Big[\exp\left(-\Theta_2 T_2-\cdots-\Theta_J T_J\right)\mid Y_1=n_1\Big]\,. \tag{2.3}$$

$$E\Big[\exp\left(-\Theta_1 T_1\right)\mid Y_1=n_1,\,T_2,\ldots,T_J\Big]\,,\quad \Theta_j\geqslant 0,\ j=1,\ldots,J\,.$$

To prove

$$E\left[\exp(-\Theta_1 T_1) \mid Y_1=n_1, T_2, \ldots, T_J\right] = \left[\frac{\mu_1}{\mu_1+\Theta_1}\right]^{n_1+1}, \tag{2.4}$$

we consider the joint queue length process $X=\{(X_1(t), \ldots, X_J(t)),\ t\in\mathbf{R}\}$ of the system in equilibrium and compare it with the process $\bar{X}=\{(X_1(-t), \ldots, X_J(-t)),\ t\in\mathbf{R}\}$ obtained from X by time reversal. Although X is *not reversible*, $\bar{X}$ is again Markov and stationary, with the same equilibrium distribution, describing the same cyclic queue with only one difference: The customers move in the opposite direction through the cycle, $J, J-1, \ldots, 2, 1, J, J-1, \ldots$.
Therefore the operation of time reversal transforms our situation where C jumps from node 1 to node 2 into a situation where C jumps from node 2 to node 1, and the distribution of T_1 given $\{Y_1=n_1, T_2, \ldots, T_J\}$ in the original system becomes, in the time reversed system, just the distribution of the next sojourn time of C in node 1, finding there n_1 customers before him and having experienced sojourn times $T_J, \ldots, T_2$ at nodes $J, \ldots, 2$ just before entering node 1. But given n_1 the distribution of the next sojourn time at node 1 does not depend on the previous sojourn times; this proves (2.4) because sojourn times in the original system and the system under time reversal are identically distributed.
To determine

$$E\left[\exp(-\Theta_2 T_2 - \cdots - \Theta_J T_J) \mid Y_1=n_1\right]$$

we apply the overtake-free property of the path $(1, 2, \ldots, J)$ which C has to pass. The n_1 customers left behind at node 1 cannot influence C's sojourn times at nodes $2, \ldots, J$. So the joint sojourn time distribution of $T_2, \ldots, T_J$ can be determined as the joint distribution of the successive sojourn times of C in a cycle consisting pf nodes $2, \ldots, J$ with $N-n_1$ customers cycling.
Having at hand the result of Boxma and Donk [1982] for a two-queue cycle we therefore can proceed by induction to obtain the following theorem:

Theorem 2.3. (Boxma, Kelly and Konheim [1984])
The LST of the equilibrium joint sojourn time distribution for the successive sojourn times of a customer at the J $\exp(\mu_j)$ single server FCFS queues of a closed cycle with N customers is:

$$E\left[\exp(-\Theta_1 T_1 - \cdots - \Theta_J T_J)\right] =$$

$$G(N-1,J)^{-1} \sum_{n_1+\ldots+n_J=N-1} \left[\prod_{j=1}^{J} \mu_j^{-n_j}\right] \left[\prod_{j=1}^{J} \left[\frac{\mu_j}{\mu_j+\Theta_j}\right]^{n_j+1}\right], \tag{2.5}$$
$$\Theta_j \geq 0,\ j=1, \ldots, J .$$

Before we proceed, some comments are in order.

Remark 2.1.
(1) The queue length distribution at arrival instants, $\pi^{(N-1)}(\cdot)$, as well as the LST $E\left[\exp(-\Theta_1 T_1 - \ldots - \Theta_J T_J)\right]$ appearing in the theorem, are called to be of "product form" (the latter with respect to the underlying queue length product form).
(2) In the proof of the theorem we have applied the principle of observing a 'tagged' customer's passage through a prescribed path (here the complete cycle). In the present case this principle can be sketched as follows. Because all customers are identical, and because in equilibrium their behaviour is stochastically indistinguishable, we can observe an arbitrary one of them in detail - but we must fix the one who is chosen. This is done by colouring the 'tagged' or 'test' customer, C, in such a way that the system's behaviour is not perturbed. From the celebrated *Arrival*

Theorem of SEVCIK and MITRANI [1981] and of LAVENBERG and REISER [1980] it is known how the other customers are distributed when C enters a cycle (or, in more general problems, a prescribed path). For open networks the analogous problem is solved by the PASTA theorem ('Poisson Arrivals See Time Averages') of WOLFF [1982]. The principle of using test customers has found wide application in the design of simulation experiments for obtaining passage times, cf. IGLEHART and SHEDLER [1980] for a further discussion of the principle. In the following we shall apply this useful principle freely.

(3) Note the symmetry of (2.5) in the service rates $\mu_1,...,\mu_J$, which reveals the fact that the joint sojourn time distribution does not depend on the order of the queues.

(4) $T_1, \ldots, T_J$ are dependent; for a discussion see KELLY [1984], BOXMA and DONK [1982], BOXMA, KELLY and KONHEIM [1984]. Formula (2.5) reveals that the dependence between T_i and T_j has the same structure as the dependence between queue lengths at stations i and j at the jump epoch of an arbitrary customer.

The connection between the last theorem and the independence result of REICH [1963] (cf. Theorem 2.1) is made by the following observation:
Denote by $T_1, \ldots, T_J$ C's successive sojourn times at the nodes $1, \ldots, J$ of an open tandem of $\exp(\mu_j)$ single server FCFS queues with Poisson(λ) input process at the first node. The independence of $T_1, \ldots, T_J$ implies that the LST of the joint distribution of $(T_1, \ldots, T_J)$ is

$$E\left[\exp\left(-\Theta_1 T_1 - \cdots - \Theta_J T_J\right)\right] = \prod_{j=1}^{J} \frac{\mu_j-\lambda}{\mu_j-\lambda+\Theta_j} =$$
$$\sum_{(m_1,\ldots,m_J)\in\mathbb{N}^J} \left[\prod_{j=1}^{J}\left[1-\frac{\lambda}{\mu_j}\right]\left[\frac{\lambda}{\mu_j}\right]^{m_j}\right]\left[\prod_{j=1}^{J}\left[\frac{\mu_j}{\mu_j+\Theta_j}\right]^{m_j+1}\right], \tag{2.6}$$
$$\Theta_j \geq 0,\ j=1,\ldots,J .$$

Observing that

$$\pi(m_1,\ldots,m_J)=\prod_{j=1}^{J}\left[1-\frac{\lambda}{\mu_j}\right]\left[\frac{\lambda}{\mu_j}\right]^{m_j},\ (m_1,\ldots,m_J)\in\mathbb{N}^J ,$$

is the steady state distribution for the number of other customers found by C at the nodes on entering the tandem, connects (2.6) with the closed-cycle result (2.5).

Later on in this section we state the general Theorem 2.4 in which the above results for open and closed systems are included in a unified formulation. Before going into the details of a general system description let us point out that, similar to our last discussion, the results of KELLY and POLLETT [1983] and WALRAND and VARAIYA [1980] on the joint distribution for sojourn times on overtake-free paths of a closed, respectively open, network can both be expressed as "product-form" LST results. Compared with BURKE's 1968 paper, these results still lacked one feature: the allowance of multiserver queues at the first station of the tandem sequence in BURKE's sojourn time theorem (allowing multiserver queues at the *last* station of an overtake-free path poses no serious problem). Including an extension of BURKE's sojourn time theorem into a general theorem will be the final step of our development of closed form analytical results for sojourn time distributions. We shall describe this in detail to obtain a unified formulation for almost all the known explicit results.

Network description
We consider a network of nodes $\tilde{J}=\{1,\ldots,J\}$ with an internal population $\tilde{M}=\{1,\ldots,M\}$ and an unlimited external population. ($M=0$, i.e. no internal customers, is allowed and leads to the case of an open network; similarly, a closed network can be obtained as another special case.) At any time the internal customer m is of some type $t\in T(m)$, $m\in\tilde{M}$, while each external customer present is of some type $t\in T(0)$, where $T(0), T(1), \ldots, T(M)$ is a collection of pairwise disjoint countable sets. With each type $t\in\bigcup_{m=0}^{M} T(m)$ is associated a finite route $t=[r(t,1),\ r(t,2), \ldots, r(t,S(t))]$, $1\leq S(t)<\infty$.

External customers of type $t \in T(0)$ arrive in a Poisson stream of intensity $\nu(t)$ at the network, pass through the associated route t, and depart eventually from the network. The internal customer $m \in \tilde{M}$, being of type $t \in T(m)$, travels along route t, thereafter changes to type $t' \in T(m)$ with probability $\nu(t')$, travels along route t', and so on ($\sum_{t' \in T(m)} \nu(t')=1,\ m \in \tilde{M}$).

We assume that all nodes of the network operate in such a way that the network is of a so-called product form. This includes, e.g., the node structures of BCMP networks (BASKETT et al. [1975]), KELLY'S symmetric servers (KELLY [1979]), and any non-priority server with type-independent exponentially distributed service times. For details about such nodes see Section II.6. For simplicity of the presentation we shall assume that we only have exponential multiserver nodes:

Node $j \in \tilde{J}$ has $m(j)$ service positions, $1 \leqslant m(j) \leqslant \infty$, an infinite waiting room with FCFS regime, and the service time for each customer at node j is drawn according to an exponential distribution with mean $\mu(j)^{-1}$.

Finally we make the usual overall independence assumption: the set of all service times, all arrival times, and all type decisions is an independent family.

A Markovian description $X=\{X(t);\ t \in \mathbb{R}\}$ of the network evolution over time is constructed in the usual way (for details see KELLY [1979]), recording for any node: (i) the types of the customers (if any) present at the service and waiting places and (ii) the actual stage of their routes.

We assume X to be ergodic on the set of feasible states of the state space $S(M,\ J)$. Then the unique equilibrium distribution of X is of product form; in detail:

For $x=(x_1, \ldots, x_J) \in S(M,\ J)$ let $n_j = n_j(x)$ be the number of customers present at node $j \in \tilde{J}$, and $c_j(k)=(t_j(k),\ s_j(k))$, with $t_j(k)$ the type of the customer on place $k \in \{\underset{\sim}{1}, \ldots, n_j\}$ and $s_j(k) \in \{1, \ldots, S(t_j(k))\}$ the actual stage of his route $t_j(k)$; then $x_j=(c_j(1), \ldots, c_j(n_j))$, $j \in J$.

The stationary distribution of X is given by

$$\overline{\pi}^{(M)}(x_1, \ldots, x_J) = \overline{G}(M,\ J)^{-1} \prod_{j=1}^{J} \left[\prod_{k=1}^{n_j} \frac{\alpha_j(t_j(k),\ s_j(k))}{\mu(j)\ a_j(k)} \right],$$

where $\overline{G}(M,\ J)$ is the normalizing constant,

$$\alpha_j(t,\ s)=\nu(t)\ 1_{(r(t,\ s)=j)},\quad j \in \tilde{J},\ t \in \bigcup_{m=0}^{M} T(m),\ 1 \leqslant s \leqslant S(t)\,,$$

$$a_j(k)=\begin{cases} k & \text{if } k \leqslant m(j)-1 \\ m(j) & \text{if } k \geqslant m(j) \end{cases},\ j \in \tilde{J},\ k \in \mathbb{N}\,.$$

The structure of the paths along which sojourn time distributions can be given in closed product-form expressions is basically the following. Such paths can be divided into three subpaths. The first and last subpath consist of infinite server nodes, these subpaths being 'uncritical'. The central subpath is 'critical'; it may begin and end with a *multiserver* node which is not an infinite server node, but in between it fulfills the topological overtake-free condition and the non-overtaking node structure of FCFS single server tandems.

DEFINITION 2.2.

(i) For $i,\ j \in \tilde{J}$ write $i-(t) \rightarrow j$ if $r(t,\ s)=i,\ r(t,\ s+1)=j$ for $t \in \bigcup_{m=0}^{M} T(m),\quad 1 \leqslant s < S(t)$.

(ii) A *t-relevant path* $[j_1, j_2, \ldots, j_K]$ from node j_1 to node j_K is a sequence of nodes such that $j_k-(t_k) \rightarrow j_{k+1},\ 1 \leqslant k < K$, holds

$$\text{with} \begin{cases} t_k \in \bigcup_{m=0}^{M} T(m) \text{ if } t \in T(0)\,, \\ t_k \in \bigcup_{\substack{m=0 \\ m \neq \overline{m}}}^{M} T(m) \text{ if } t \in T(\overline{m})\,,\quad 1 \leqslant \overline{m} \leqslant M\,. \end{cases}$$

(Note that in general a path need not be a route or a part of a route.)

(iii) For $t \in \bigcup_{m=0}^{M} T(m)$, let $[r(t,u), r(t,u+1), \ldots, r(t,v)]$ be a section of different nodes of route t, $1 \leqslant u \leqslant v \leqslant S(t)$. This section is overtake-free for type t if every t-relevant path from $r(t,\ u')$ to $r(t,\ v')$ includes node $r(t,u'+1)$, $u \leqslant u' < v' \leqslant v$.
(Note that this is a purely topological property of the network.)

(iv) For type $t \in \bigcup_{m=0}^{M} T(m)$ the route t is *quasi overtake-free* if the following holds:
There exist u,v, $1 \leqslant u \leqslant v \leqslant S(t)$ such that

$$m(r(t,1)) = \cdots = m(r(t,u-1)) = \infty\ ,$$
$$m(r(t,u+1)) = \cdots = m(r(t,v-1)) = 1\ ,$$
$$m(r(t,v+1)) = \cdots = m(r(t,S(t))) = \infty\ ,$$

and the section $[r(t,u), \ldots, r(t,v)]$ of t is overtake-free for type t.

(Note that the section $[r(t,u), \ldots, r(t,v)]$ of t is the critical part of the path, which may begin and end with a *multiserver* node which is *not* an infinite server node.)

THEOREM 2.4.
Let $T_s^{(n)}$, $1 \leqslant s \leqslant S(t)$, denote the successive sojourn times at the nodes of route t of the n-th type-t customer entering route t after the start of the system. Assume that t is quasi overtake-free, applying the notation of Definition 2.2(iv). For $\Theta_s \geqslant 0$, $1 \leqslant s \leqslant S(t)$,

$$\lim_{n\to\infty} E\ [\exp(-\Theta_1 T_1^{(n)} - \ldots - \Theta_{S(t)} T_{S(t)}^{(n)})] =$$

$$\left[\prod_{s=1}^{S(t)} \frac{\mu(s)}{\mu(s)+\Theta_s}\right] \sum_{(n_1,\ldots,n_{S(t)})} p(n_1,\ldots,n_{S(t)}) \prod_{s=u}^{v} \left[\frac{\mu(s)m(s)}{\mu(s)m(s)+\Theta_s}\right]^{[n_s+1-m(s)]_+}\ , \tag{2.7}$$

where we have used the abbreviations

$$\mu(r(s,t)) =: \mu(s),\ m(r(s,t)) =: m(s),\ 1 \leqslant s \leqslant S(t)\ ,$$
$$[a]_+ = \max(0,a)\ ,$$

and $p(n_1, \ldots, n_{S(t)})$ denotes the limiting (and stationary) probability that at arrival times of a type-t customer n_s other customers are present at node $r(t,s)$, $1 \leqslant s \leqslant S(t)$, and where the summation is performed over all feasible joint queue lengths of other customers seen by a type-t customer.
The joint steady state distribution of the successive sojourn times of a type-t customer at the nodes of route t is also given by (2.7).

The proof of Theorem 2.4 can be found in SCHASSBERGER and DADUNA [1987]. The main work to be done is to compute for node $r(t,u)$, having $2 \leqslant m(r(t,u)) < \infty$ service channels, the joint distribution of a type-t customer's sojourn time there and the state of the network at the departure from node $r(t,u)$ of that customer. Having obtained this joint distribution, a *splitting formula* follows from the strong Markov property of X which splits the sojourn times in route t into a part concerning nodes $r(t,1), \ldots, r(t,u)$ (given by an explicit expression!) and into a part concerning the rest of route t in an implicitly given term.
Remarkably enough, this *splitting formula* holds without the assumptions posed by the quasi overtake-free property on nodes $r(t,u+1), \ldots, r(t,v)$. These assumptions are then needed to convert the implicitly given term into an explicit formula — and this is done utilizing an inductive argument similar to that in the proof of BOXMA, KELLY and KONHEIM [1984] (see Theorem 2.3).

We close this section with a discussion of the main result.

REMARK 2.2.

(1) Computing $p(n_1, \ldots, n_{S(t)})$ is easy due to the simple product form of the system's equilibrium queue length distribution and the arrival theorem (LAVENBERG and REISER [1980], SEVCIK and MITRANI [1981]), which gives the limiting and stationary distribution of the system seen by type-t customers entering route t.

(2) The nodes outside route t can have a very general structure as long as the steady state distribution remains of product form.

(3) Let $(T_1, \ldots, T_{S(t)})$ denote a vector having the (limiting) distribution obtained in the theorem. Then the set of random variables $T_1, \ldots, T_{u-1}, T_{v+1}, \ldots, T_{S(t)}$ and the vector $(T_u, \ldots, T_v)$ form a stochastically independent collection. If additionally route t is external, i.e., $t \in T(0)$, and for some $u \leqslant u_1 < u_2 < ... < u_k \leqslant v$ nodes $r(t,u_i)$, $1 \leqslant i \leqslant k$, are visited by external customers only, then $T_1, \ldots, T_{u-1}, T_{u_1}, T_{u_2}, \ldots, T_{u_k}, T_{v+1}, \ldots, T_{S(t)}$ and the random vector

$$(T_u, \ldots, T_{u_1-1}, T_{u_1+1}, \ldots, T_{u_k-1}, T_{u_k+1}, \ldots, T_v)$$

again form an independent family.

(4) Setting $\Theta_1 = ... = \Theta_{S(t)} =: \Theta$ in (2.7) yields the LST of the passage time distribution through route t for type-t customers, i.e., the distribution of the sum of the sojourn times at the nodes. This expression was computed in DADUNA [1984].

(5) A remarkable refinement of Theorem 2.4 was obtained for the totally closed network case by HEMKER [1987,1989], who obtained the *splitting formula* generalizing the technique of time reversal in the proof of Theorem 2.3. In terms of the theorem he allowed the multiserver node $r(t,u)$ to have different service intensities at the various channels (servers). This requires that the node works under shift protocol, i.e., service channels are numbered $1, \ldots, m(r(t,u))$, a customer entering service always goes to the free channel with the lowest number, and if a customer departs from the node, other customers being served in channels with a higher number are shifted such that the gap is closed — after that a possibly waiting customer enters channel $m(r(t,u))$. Clearly HEMKER's result holds for open and mixed networks as well.

(6) The theorem is formulated with respect to sojourn times of a customer traversing a complete route $t = [r(t, 1), \ldots, r(t,S(t))]$. It holds as well if a customer is only observed on a specified section of a path which fulfills the quasi overtake-free requirements. On the other hand, if internal customer $m \in M$ has possible routes t, $t' \in T(m)$ such that the itinerary $[t, t']$ pasted together is quasi overtake-free, the theorem may be applied again (formally this may be shown by redefinition of types).

(7) KAWASHIMA and TORIGOE [1983] have dealt with the case of a closed star-shaped network of multiserver nodes. Using the reversibility of the joint queue length process they computed the joint sojourn time distribution for a customer during one cycle, i.e., the period between entering the central node and returning there after visiting exactly one of the external nodes. WONG [1979] and SEKINO [1972] have investigated a closed two-stage tandem of a multiserver and an infinite server.

(8) Because two nodes which a customer may subsequently visit form a quasi overtake-free path, the joint sojourn time distribution for "two-stations walks" can be computed explicitly from the theorem. A "two-stations walk" may even consist of two successive sojourns at the same node (feedback). This follows from the "splitting formula" mentioned in the proof of the theorem (see KELLY and POLLETT [1983], and DADUNA [1983]).

II.4. NUMERICAL EVALUATION OF THE SOJOURN TIME FORMULAS

In principle Theorem 2.4 gives a complete answer to the questions about joint sojourn time distributions and passage time distributions along quasi overtake-free paths, the latter distributions possibly being more important from a practical point of view. Formula (2.7), written down in terms of product-form queue length probabilities and LST's, can be inverted by inspection, leading to a mixture of convolutions of ERLANG distributions. But it turns out that numerical problems arise due to iterated integration and to small queue length probabilities, which in the closed or mixed network case are burdened with complicated normalizing constants.

HARRISON [1984] proved for closed cyclic queues that the product-form passage time result of SCHASSBERGER and DADUNA [1983] may be inverted to provide a formula which expresses the cycle time distribution as a mixture of ERLANG distributions, in the same way as CHOW [1980] stated his two-queue result. Applying the notation of Theorem 2.3 we have

COROLLARY 2.1. (HARRISON [1984])
Consider the cyclic model of Theorem 2.3. Assume that the service rates μ_j, $j=1,\ldots,J$, of the servers are all distinct. Then the probability density of the cycle time $T_1+T_2+\ldots+T_J$ in equilibrium is

$$f(N,J;s) = G(N-1,J)^{-1}\left(\prod_{j=1}^{J}\mu_j\right)\frac{s^{N-1}}{(N-1)!}\sum_{j=1}^{J}\prod_{\substack{i=1\\ i\neq j}}^{J}\frac{1}{\mu_i-\mu_j}e^{-\mu_j s}, \quad s\geq 0. \tag{2.8}$$

HARRISON [1989] shows that for tree-like closed networks similar expressions can be found, but again the formulas become very complicated. Therefore he developed methods for evaluating these formulas - which in the closed network case results in an algorithm like BUZEN's algorithm (see e.g. LAVENBERG [1983]). Another approach was developed by MCKENNA [1988]: He observed that a new technique to compute normalizing constants for the evaluation of queue length distributions in closed networks, called RECAL, can be modified to compute passage time quantiles for single server overtake-free paths in GORDON-NEWELL networks. A discretization approach to evaluate the sojourn time formulas is presented in HARRISON [1982].

Moments of passage times
Differentiating the passage time LST leads to passage time moments. In the case of closed networks it was observed by REISER [1981] that differentiation of the cycle time LST leads to an iterative scheme to compute cycle time moments of any order in a closed single server cycle. For the closed cycle defined in Section II.3 he obtained the following scheme to calculate the cycle time moments

$$m^{(i)}_{N,J} = E\left[\left(\sum_{j=1}^{J}T_j\right)^i\right], \quad i=1,2,\ldots,I.$$

ALGORITHM:
INITIALIZE:

$$G(k,0)=0, \quad G(0,j)=1, \quad m^{(i)}_{k,0}=0, \quad m^{(i)}_{0,j}=0,$$

for $k=0,1,\ldots,N$, $j=1,\ldots,J$, $i=1,\ldots,I$.

LOOP:

For $j=1,\ldots,J$, and $k=1,\ldots,N$ do

$$G(k,j) := \mu_j^{-1}\, G(k-1,j)+G(k,j-1), \quad p^0_{k-1,j} := G(k-1,j-1)\, G(k-1,j)^{-1},$$

$$m^{(1)}_{k,j} := \mu_j^{-1}+(1-p^0_{k-1,j})m^{(1)}_{k-1,j}+p^0_{k-1,j}m^{(1)}_{k,j-1},$$

$$m^{(2)}_{k,j} := 2\mu_j^{-1}\, m^{(1)}_{k,j}+(1-p^0_{k-1,j})m^{(2)}_{k-1,j}+p^0_{k-1,j}m^{(2)}_{k,j-1},$$

$$\ldots$$

$$m^{(I)}_{k,j} := I\mu_j^{-1}\, m^{(I-1)}_{k,j}+(1-p^0_{k-1,j})m^{(I)}_{k-1,j}+p^0_{k-1,j}m^{(I)}_{k,j-1}.$$

A similar algorithm for the case of the general Theorem 2.4 (applied to closed networks) was obtained by HEMKER [1987]. His algorithm applies in particular to the multiserver case of BURKE's sojourn time theorem, and to his own generalization of that theorem (cf. Remark 2.2 (5)).

II.5 The non-overtaking condition revisited

The main result on sojourn time distributions in Theorem 2.4 utilizes the quasi overtake-free property of the path under consideration to perform the step from the splitting formula (which holds without non-overtaking requirements) to an explicit expression. In our opinion, the quasi overtake-free property is the heart of the matter, and to understand overtaking and non-overtaking might be the most important step towards further possible generalizations of the theorem (see Whitt [1984a] for a discussion on overtaking in networks of queues, which includes some quantitative definitions of the *amount* of overtaking). Our discussion has to take into consideration two types of overtaking.

(i) overtaking due to the topological structure of the network, and
(ii) overtaking due to the internal node structure.

(i) *Overtaking due to the topological structure of the network*
The discovery of the central position of the overtake-free condition for solving sojourn time and passage time problems by Walrand and Varaiya [1980] and Melamed [1982] led to a discussion of whether such results could be obtained for more general paths — paths with overtaking. The main theme of this discussion was the simplest open network with overtaking, given as Example (B) in Section II.2 (the Simon-Foley network). Simon and Foley [1979] and Walrand and Varaiya [1980] proved that the sojourn times in nodes 1 and 3 of a customer passing through (1, 2, 3) are dependent. Recently Foley and Kiessler [1989] have shown that a customer's sojourn times in nodes 1 and 3 are positively correlated; actually they proved the stronger result that these sojourn times are 'associated' random variables. But up to now an *explicit* expression for the passage time is not known.

From extensive simulation experiments it was known that the passage time distribution for the path (1,2,3) is *almost* the same as the convolution of the sojourn time distributions at the individual nodes (see Kiessler et al. [1988]). That this convolution is *not* the passage time distribution was proved by Melamed [1983] using brute force numerical techniques (cf. Section III.4). Kiessler and Disney [1982] attacked the problem by pointing out how to compute the conditional sojourn times given the system state, and Fayolle, Iasnogorodski and Mitrani [1983] showed that the sojourn time problem for the Simon-Foley network can be converted into a Riemann-Hilbert-Carleman boundary value problem which can be solved up to numerical integration of a Fredholm integral equation.

A second fundamental example of overtaking due to the topological structure of the network is provided by investigating successive cycles in closed tandem systems. In the same way as the Simon-Foley network has found so much interest as a simple representative of a whole class of network problems, the two-stage tandem of Chow [1980] was investigated by several researchers later on. The problem of interest which up to now is only partly solved is: Determine the distribution of the time required for n cycles of a customer, $n \geqslant 2$. Boxma [1984] computed the time for two cycles and obtained a very involved expression. Daduna [1986a] obtained a scheme, recursive in the number of cycles and the population size, to compute n-fold cycle times. The same technique was applied in Daduna [1986b] to solve the case where one of the nodes allows immediate feedback.

A third example where overtaking appears due to the topological structure of the network is the single server node with feedback. Consider the $M/M/1$ queue with feedback probability p, as depicted in Fig. 2.5.

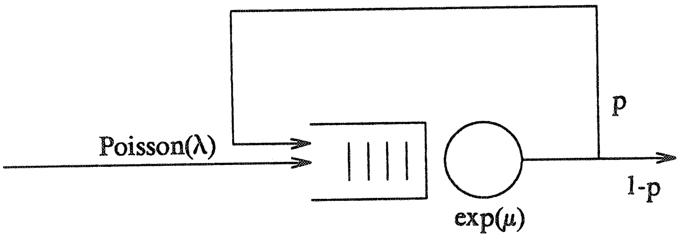

Fig. 2.5.

With respect to sojourn times, even this simple system with immediate so-called Bernoulli feedback shows a complicated structure. The main idea to compute a customer's sojourn time distribution goes back to TAKÁCS [1963]. Under the ergodicity assumption $(1-p)\mu>\lambda$ in equilibrium the feedback structure leads to the evaluation of an iterated transformation. Introduce, with $q=1-p$,

$$U_0(s,z)=\frac{1-\dfrac{\lambda}{\mu q}}{1-\dfrac{\lambda z}{\mu q}},$$

$$U_{k+1}(s,z)=\frac{\mu}{\mu+\lambda(1-z)+s}\; U_k\left[s,\frac{\mu(q+pz)}{\mu+\lambda(1-z)+s}\right],\quad k=0,1,\ldots .$$

Then the LST of the distribution of a customer's total sojourn time is

$$\phi(s)=q\sum_{k=1}^{\infty}p^{k-1}U_k(s,\,1)=\left[1-\frac{\lambda}{\mu q}\right]q\sum_{k=1}^{\infty}p^{k-1}\frac{1}{a_k(s)-b_k(s)},\quad s\geqslant 0\,, \tag{2.9}$$

where $a_k(s)$, $b_k(s)$, $(k=1,2,\ldots)$ are given by

$$\begin{bmatrix}a_k(s)\\ b_k(s)\end{bmatrix}=\begin{bmatrix}\dfrac{\mu+\lambda+s}{\mu} & -q\\ \dfrac{\lambda}{\mu} & p\end{bmatrix}^k\begin{bmatrix}1\\ \dfrac{\lambda}{\mu q}\end{bmatrix}.$$

For the case of a general distribution of the number of feedback cycles that a customer has to perform, LAM and SHANKAR [1981] have derived the transform of the total customer response time, conditioned on the number of cycles. They have investigated this system as a model of a computer system in time-sharing mode where the time quanta are exponentially distributed. TAKÁCS's result on feedback systems is also generalized in two papers of VAN DEN BERG et al. [1989], where the feedback probabilities are allowed to depend on the number of services already obtained; the system can be modeled as a product-form network by using different customer types. They obtain the LST of the joint distribution for the first $k(\geqslant 1)$ successive sojourn times of a customer, who is fed back at least $k-1$ times; the formula again reflects the feedback structure by some iterated transformation. As a by-product they show that the successive sojourn times all have an identical exponential marginal distribution. Furthermore, by letting the feedback probabilities approach one and the mean service time at each loop approach zero, such that the total required mean service time remains constant, they provide a novel method to analyze the sojourn time distribution in the $M/G/1$ processor sharing queue.

A somewhat more involved system which fits into this section is the queue with delayed feedback depicted in Fig. 2.6. It combines the feedback and cyclic structures (see FOLEY and DISNEY [1981] and KÖNIG and MIYAZAWA [1988]). Theorem 2.4 applies to determine the joint distribution of the sojourn times of a customer during one cycle. No further explicit results seem to exist.

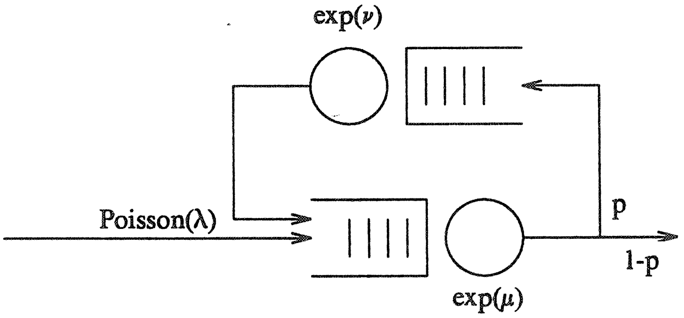

Fig. 2.6.

(ii) *Overtaking due to the internal node structure*
The first noteworthy result on sojourn times in a network with overtaking due to node structure was reported by BURKE [1969] (see Section II.1). Until now there are only some isolated examples where passage time distributions can be explicitly computed when such nodes appear.

COFFMANN, FAYOLLE and MITRANI [1986] dealt with an open exponential tandem, where the first node is a processor sharing (PS) node and the second one is either a single server FCFS queue or PS node. In both cases the passage time problem is reduced to the solution of boundary value problems (see also MITRANI [1985]). In contrast to this complicated situation, in a tandem of two exponential nodes under LCFS preemptive-resume regime (where overtaking appears!) successive sojourn times of a customer are independent (see KELLY [1979], Ex. 2.2.3).

For closed two-node cycles of one FCFS node and one m-limited PS discipline (i.e., at most m customers can share the processor) KAWASHIMA [1987] computed the joint sojourn time distribution for a cycle; for the case of two PS nodes the cycle time distribution is given in DADUNA [1985].

We close this section with a possibly somewhat disturbing observation which points out the problems arising in even very simple systems. We consider the following system of exponential multiserver nodes:

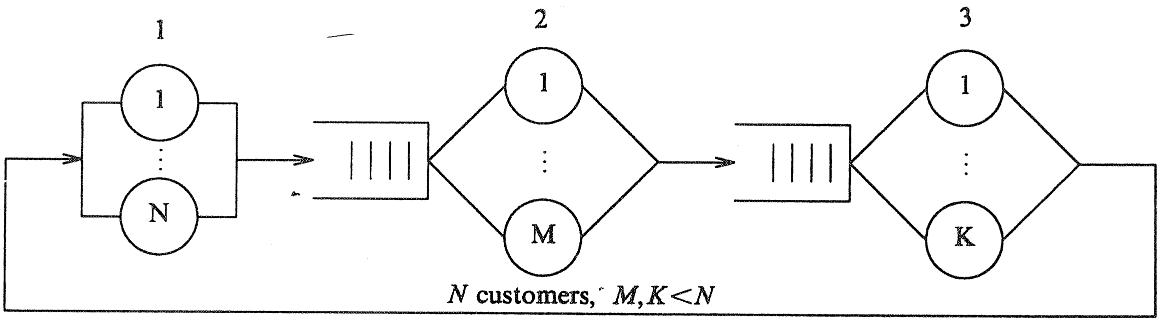

Fig. 2.7.

A tagged customer C performing one cycle can be physically overtaken several times, while also, due to potential waiting, influences generated by C can overtake him and re-influence his further behaviour. The implication of Theorem 2.4 for the passage times is: The passage time distributions for the paths (1, 2, 3) and (2, 3, 1) are identical and have product-form LST — but they differ from the passage time distribution for path (3, 1, 2).

II.6. MEAN SOJOURN TIMES

As was pointed out in the introduction, passage time quantiles are a typical performance measure for communication systems. The few known results have been largely discussed in the previous sections. Fortunately, it is relatively easy to obtain *mean* passage times. For the whole class of product-form networks we have access to mean passage times through *any* prescribed sequence of nodes. Even more: A set of efficient algorithms has been developed to compute the individual mean sojourn time at any node; *linearity* of the expectation operator then yields mean passage time through any path. The sensitivity of the mean cycle time in an exponential central server system with respect to perturbation of the CPU service intensity is investigated by WOODSIDE [1984]; for general open and closed networks CAO and HO [1987] deal with similar problems, using the recently developed perturbation analysis of queues.

There is a very extensive literature on exact and approximate algorithms for the determination of *mean queue lengths* in product-form networks. Details and reviews can be found in LAVENBERG [1983], Bruell and Balbo [1980], Agrawal [1985]. A prominent exact algorithm is the MVA (Mean Value Analysis) algorithm, which calculates mean queue lengths in closed networks without having to compute the normalizing constant. Some recent developments can be found in Conway and Georganas [1986], McKenna [1988] and De Souza e Silva and Lavenberg [1989].

In principle, equilibrium *mean sojourn times* are obtainable from equilibrium mean queue lengths via Little's formula; this is in fact utilized in the formulation of the MVA algorithm, together with the already mentioned arrival theorem which in a popular formulation states that in equilibrium an arriving customer sees the other customers distributed as if the arriving customer is cancelled (Lavenberg and Reiser [1980], Sevcik and Mitrani [1981]). The mean value analysis algorithm therefore can be generalized to hold in the fairly general class of product-form networks, as we described it in Section II.3, including now the so-called symmetric server nodes (Kelly [1979]).

Such nodes also play a central role in the following discussion of conditional mean passage times. Therefore we discuss them in some detail. The structure of a symmetric node, say node j, is as follows: There is an unlimited sequence of positions $(p_1,p_2,...)$. If n_j customers are present, they occupy positions $(p_1,\ldots,p_{n_j})$. Given there are n_j customers present, a newly arriving customer moves into position $k\in\{1,\ldots,n_j+1\}$ with probability $\delta_j(k,n_j+1)$, shifting customers previously on $(p_k,\ldots,p_{n_j})$ to $(p_{k+1},\ldots,p_{n_j+1})$. When n_j customers are present, a total service effort is supplied at rate $\phi_j(n_j)$ and a proportion $\gamma_j(k,n_j)$ of this effort is given to the customer in position $k\in\{1,\ldots,n_j\}$. The symmetry condition is now:

$$\gamma_j(k,n_j) = \delta_j(k,n_j) \ \forall n_j\geqslant 1,\ 1\leqslant k\leqslant n_j\ .$$

If this condition is fulfilled for node $j\in\tilde{J}$, the service time distribution of customers at this node may be drawn from a general distribution, depending on the type of the customer and the stage of the route he has actually reached. (We may define further "general exponential" nodes utilizing functions γ_j,δ_j as above without requiring symmetry. Then to obtain product-form steady state distributions, the service time distributions have to be exponential with node-specific mean, not depending on the customer's type.)

In the context of Generalized Semi-Markov Processes, Barbour and Schassberger [1981] proved the following remarkable result on conditional holding times for these processes, which can be translated into the terminology of queueing networks:

Theorem 2.5. (Schassberger [1985])
Consider a closed network of queues in equilibrium as described in Section II.3 allowing general exponential and symmetric servers. For a customer of type t let his route t include symmetric server nodes at stages $s_1,s_2,\ldots,s_m$, $1\leqslant s_1<\cdots<s_m\leqslant S(t)$. Let $T_t(x_1,\ldots,x_m)$, $x_i\geqslant 0$, $i=1,\ldots,m$, denote the customer's conditional expected passage time through route t (in equilibrium) given he requests exactly x_i units of processing time at stage s_i (visiting node $r(t,s_i)$) of his route, $i=1,\ldots,m$.
Then

$$T_t(x_1,\ldots,x_m) = \sum_{i=1}^{m} x_i\ \mu_{s_i}\ E[T_{s_i}] + \sum_{s\in\{1,\ldots,S(t)\}-\{s_1,\ldots,s_m\}} E[T_s]\ ,$$

where $E[T_s]$ is the steady state mean sojourn time of a type-t customer at node $r(t,s)$ of his route, $1\leqslant s\leqslant S(t)$, and μ_s^{-1} is the mean requested processing time of the type-t customer at node $r(t,s)$ of his route t, $1\leqslant s\leqslant S(t)$.

For details see Barbour and Schassberger [1981], and for the generalization to open and mixed networks see Schassberger [1986]. Similar results are obtained by Cohen [1979], Jansen and König [1980] and Jansen [1984]; Tsoucas and Walrand [1983] provide an intuitive explanation why the mean sojourn time in a symmetric queue of an open network of so-called quasi-reversible nodes depends on the service time distribution only through its mean. Schassberger [1985] points out that the general theory of Generalized Semi-Markov processes allows the following refinement of the network theory and the theorem:

Given we have symmetric server nodes at stations $s_1, \ldots, s_m$ of route t, the service time requests of type-t customers at those nodes may be drawn according to a general joint distribution. This will not change the steady state distribution obtained under the usual independence assumptions as long as marginal expected service time requests are not changed. And this implies that the theorem on the mean conditional passage times holds again as it stands.

A similar theory for discrete systems featuring network structures of even more generality has been developed in DADUNA and SCHASSBERGER [1983], SCHASSBERGER and DADUNA [1983a]. They distinguish "geometrical" servers and "doubly-stochastic" servers, the latter replacing the symmetric servers. For such networks a simple theorem on conditional expected passage times given the service time requests at the doubly-stochastic nodes holds as well.

II.7. APPROXIMATION METHODS FOR PRODUCT-FORM NETWORKS

We distinguish two kinds of approximation procedures, according to the following two reasons for the need for approximations:

(i) Although explicit sojourn time formulas as we presented before are at hand, the size and complexity of the network do not allow an exact evaluation of these formulas, and

(ii) no explicit formulas are known for the given problem.

(i.a) Let us consider again the cyclic network of single server queues of Section II.3, where the joint distribution of the successive sojourn times is derived. Let us assume that all the service times have the same mean $\mu_j^{-1}=1$ and let $\underset{\sim}{T}^n=(T_1^n, T_2^n, \ldots, T_J^n)$, $n \geqslant 1$, be the vector of sojourn times of a tagged customer C during his nth cycle, where it is assumed that the system starts with C commencing his first cycle at time 0, seeing the other customers in equilibrium. KELLY [1984] investigated the effect of the number, N, of customers in the system becoming large. He proved:

THEOREM 2.6.

Let $\underset{\sim}{U}=(U_1, \ldots, U_J)$ *be a random vector uniformly distributed on the simplex* $\{(u_1, \ldots, u_J) : u_j \geqslant 0, j \in \tilde{J}, \sum_{j=1}^{J} u_j = 1\}$.
For the first K cycles of the tagged customer, as $N \to \infty$, *we have convergence in distribution*

$$N^{-1}(\underset{\sim}{T}^1, \underset{\sim}{T}^2, \ldots, \underset{\sim}{T}^K) \overset{d}{\to} (\underset{\sim}{U}, \underset{\sim}{U}, \ldots, \underset{\sim}{U}).$$

As KELLY [1984] states:
"Hence, if traffic is heavy enough, a customer's normalized sojourn times can be approximated by a constrained uniform distribution. Further, a customer's sojourn times at successive visits to a given queue are approximately equal."
Another heavy traffic approximation for the cyclic queue is given by BOXMA [1988]. Assuming all the service rates to be distinct, and (without loss of generality, cf. Remark 2.1 (3)) $\mu_1 < \mu_2 < \cdots < \mu_J$, i.e., node 1 is the slowest server, he obtained the following result for the stationary cycle time of a customer:

$$E[\exp(-\Theta(T_1 + \cdots + T_J))] = \left[\frac{\mu_1}{\mu_1+\Theta}\right]^N \left\{1+O\left[\left[\frac{\Theta+\mu_1}{\Theta+\mu_2}\right]^N\right]\right\}, \quad N \to \infty, \ \Theta \geqslant 0\,.$$

This result suggests that in the cyclic queueing system in heavy traffic the slowest server strongly determines the cycle time distribution. Especially for large population size N, the cycle time should be well approximated by just the sum of N service times at the slowest server. Comparison of the exact and approximating densities in some test cases shows an astonishing correspondence.

(i.b) As mentioned below Corollary 2.1 in Section II.4, for an overtake-free path of single server exponential queues in closed networks McKenna [1988] has pointed out that the passage time quantiles which are determined by Theorem 2.4 can be numerically obtained using the RECAL procedure. But this algorithm becomes inadequate if the network size grows too much. He recommends the TREE RECAL algorithm for larger networks exhibiting sparseness and locality with respect to the customers' movement.

If a large network does not exhibit such properties but if every route includes at least one infinite server node, while moreover the network satisfies some normal usage condition (which guarantees that the service centers are not overloaded), McKenna [1988] shows that approximate expressions can be obtained combining RECAL and some asymptotic techniques which are developed from integral representations of the network partition function. Similar asymptotic expansion approximations are derived by McKenna [1987] for the distribution function of the sojourn time of jobs in exponential multiserver nodes in closed networks.

(ii) Let us consider the situation where sojourn time distributions and passage time distributions are not known for a customer traversing a prescribed path in a product-form network. We present some variants of a general technique to deal with such a situation, without trying to review the whole bulk of papers which are concerned with special models and ad-hoc methods.

Independent sojourn time approximation
This is the simplest way to overcome almost all difficulties. Due to the arrival theorem for product-form networks for exponential multiserver nodes, the sojourn time distribution at such nodes is known. The "independent sojourn time approximation" for open networks then simply assumes the independence of the successive sojourn times at the nodes of the path, approximating the passage time distribution by the convolution of the individual sojourn time distributions. For general paths Harrison [1981] and Hohl and Kühn [1988] state the following

Approximation Postulate
"Arrival state distributions seen by a test customer on the arrival at successive network stations are independent of each other and equal the stationary arrival state distributions at these stations seen by an arbitrary customer." (IFTA, "Independent Flow Time Approximation")

With this approximation the total flow time (passage time) distribution is the convolution of the individual flow time (sojourn time) distributions at the nodes.

Hohl and Kühn [1988] even apply their method to paths with PS nodes and LCFS-PR nodes and show by simulation that the method fits well, especially when there is no feedback on the investigated path. Feedback generally seems to reduce the accuracy of IFTA. Further they observe that for larger networks the approximation becomes more accurate, an observation also made by Harrison [1981].

The use of IFTA does not account for the covariances between successive sojourn times. Therefore Harrison [1985], [1986] suggests a pairwise analysis of servers on a prescribed path, giving a recursive scheme to compute passage time distribution functions. This scheme is derived under the assumption that the arrival state distribution, found by the test customer at the ith node in traversing the path, depends on the queue lengths at his arrival in nodes $1,2,\ldots,i-1$ of the path only through that found at node $i-1$.

The IFTA or independent sojourn time approximation seems to be a good approximation for exponential multiserver paths without cycles (feedforward networks), even if overtaking may occur as in the Simon-Foley network (example (B) in Section II.2).

A detailed discussion of the latter model is given by Kiessler et al. [1988] and by Melamed [1983], using simulation results as well as general Markov chain techniques (see Section III.4). It is pointed out that including cycles into the path leads to large deviations from the results obtained using IFTA. On the other hand, there seem to be cases where the IFTA assumption works well, even if cycles appear. One such example is provided by Shanthikumar and Buzacott [1984], who report about an open network of queues with Markovian symmetric routing, i.e.: If there are J nodes, on departure from node i a customer jumps to node $j \neq i$, $j \in \tilde{J}$, with probability J^{-1}, and leaves the network with probability J^{-1}; immediate feedback to node i is not allowed. Their approach is further developed by SHANTHIKUMAR and SUMITA [1988].

II.8. MISCELLANEA IN PRODUCT-FORM NETWORKS

Besides the - rather limited - results for sojourn time distributions in networks that have been discussed in Section II.3, there exist several isolated results. Some of those not already mentioned in Section II.5 will be sketched here.

(i) Infinite server networks (LEMOINE [1986])
At a totally open network of infinite server queues with different customer classes and Markovian routing and type selection, customers of class $m \in \{1, \ldots, M\}$ arrive at node $j \in \{1, \ldots, J\}$ in a Poisson stream with intensity function $\{\lambda_{j,m}(t),\ t \geqslant 0\}$. The service time distribution may depend on the actual type of the customer, the type he will adopt in his next station, and the actual node and the node he will visit next.
Due to the infinite server nodes customers travel independently through the network. This enables one to develop formulas for the customer's travel time distribution through the network. Clearly the effective evaluation of that quantity depends strongly on the complexity of routing and type selection mechanism.

(ii) Network flow equations (LEMOINE [1987])
For single server JACKSON networks with Markovian routing (in steady state) LEMOINE derives a set of *network flow equations* for the residual sojourn time LST of a customer given the node he actually enters. By differentiation of the LST formula he obtains moment formulas as a system of linear equations but, unfortunately, this system still contains too many unknowns. For the case of one node with Bernoulli feedback, martingale methods enable him to derive additional equations and thus to find the sojourn time variance (cf. the discussion in Section II.5 (i)). From Theorem 2.4, the network flow equations can even be derived for networks with multiserver nodes and different customer classes, cf. DADUNA [1989].

(iii) Exchangeable items in networks
BERG and POSNER [1985] consider the following problem. Customers bring failed items to a repair center. The items are assumed to be exchangeable in the sense that a customer does not necessarily want the particular item back that he brought into the repair station. So customers give their items to the station and join a queue outside of the station, waiting to get back any item. The customer queue outside the station is FCFS while the repair station may be any network of queues where items circulate until they are repaired, leave the station and are given to the customer at the head of the customer queue.
BERG and POSNER [1985] model the repair station by an $M/M/c/\infty$ system and give an explicit expression for a customer's delay distribution as well as computational formulas for the numerical evaluation of moments. Although these formulas are quite complicated, it follows from LITTLE's formula that the mean delay is the same as if a customer obtains his own item back. But in BERG and POSNER [1986] it is shown that the delay variance is minimized by using the FCFS regime in the customer queue, independent of the internal structure of the repair station. For the case of a repair station composed of two independent $M/M/1/\infty$-FCFS systems, the customer delay time LST is computed by DADUNA [1987].

(iv) Cycle times in a starshaped network with state-dependent routing
Consider a central server connected with some peripheral service stations, all exponential under FCFS with a single server. Customers served at a peripheral device go to the central processor, and being served there they are again sent to one of the peripheral devices (with fixed routing probabilities this system fits into the scope of Theorem 2.4). Blocking of some nodes is possible due to overload, but the way this routing depends on the system state is chosen such that a product-form steady state distribution is obtained. In DADUNA [1985a,1987a] the joint distribution of the successive sojourn times in a cycle is derived, revealing a product form similar to that obtained in Theorem 2.4; further, a computational algorithm for higher moments is presented.

(v) Exit time distributions
KÜHN [1983] has posed the following problems, generalizing the well known busy period problem for single server nodes: Determine, in an exponential closed two-stage cycle in equilibrium, the distribution of the time (a) during which a specified node is continuously busy ('busy period'), and (b) during which both nodes are continuously busy ('simultaneous busy periods'). KÜHN [1983] points out that, from the viewpoint of Markov process theory, these problems are sojourn time problems for staying in a prescribed subset of the state space - thus being of the same type as our passage time problems. He applies first-entrance-time techniques of the theory of Markov processes, deriving a set of linear transform equations.

Recently MASSEY [1987], BACCELLI and MASSEY [1988] and BACCELLI, MASSEY and WRIGHT [1988] have investigated the busy period problems for some more general open and closed tandem systems. They show that the distribution functions of (simultaneous) busy periods conditioned on the starting state can be expressed in terms of so-called lattice Bessel functions.

(vi) Central server systems
In MITRA and MORRISON [1985], MORRISON [1987] and several other papers (to be found in their references), the sojourn time in a PS node of a central server system consisting of this PS node and an infinite server is investigated. The key point is an asymptotic analysis of the sojourn time distribution, based on an integral representation of the network partition function.

III. Non product-form networks

III.1. Product-form approximations

In Chapter II we have dealt with an obviously very limited class of networks. Most realistic queueing models have an internal structure that destroys the product form: Priority classes of customers; state-dependent routing; blocking due to overload of nodes; non-exponential interarrival time distributions; non-exponential service time distributions at non-symmetric servers; type-dependent exponential service time distributions at non-symmetric servers; concurrency control problems which lead to splitting of jobs and to synchronization of different jobs; service disciplines which are not work conserving.

Our justification for the emphasis on sojourn time results in product-form networks is two-fold:

- the only known explicit general results are in the range of product-form networks;
- non product-form networks are frequently being approximated by product-form networks.

Indeed, even if a given system has no product-form steady state distribution, it is often possible to find a suitable system *with* product form, which yields reasonably accurate performance predictions for the original system. Such an approximate approach is in particular recommended when no or very few data are at hand from a real or projected network. If, e.g., only mean values of service times and interarrival times are given, a suitable 'one-parameter approximation' is selected by assuming that these service and interarrival times are exponentially distributed with the prescribed mean.

A typical application of the principles behind product-form approximations is the use of NORTON'S theorem (stemming from electrical circuit theory) in hierarchical investigations of general networks. NORTON'S theorem provides rules how to replace a part of a network by an 'equivalent' single node. In CHANDY, HERZOG and WOO [1975] NORTON'S theorem for queueing networks with product form is proved. Although this theorem only holds for queue lengths in product-form networks, it is mainly applied to non product-form networks assuming the equivalent substitute to be an exponential node with state-dependent service rate - and this approximation is usually claimed to be sufficiently accurate. One example is the study of CHANDY, HERZOG and WOO [1975a]: With respect to any node of a (generally non product-form) network these authors replace the rest of the network by one 'equivalent' node, using some auxiliary product-form networks. This yields a set of two-node cyclic queues which consists of one 'original' node and one 'equivalent' exponential node substituted for the rest of the network. These two-stage cycles are solved for the equilibrium queue length and waiting time *distributions* of the original node. Using linearity of the expectation operator, from this one obtains at least approximate mean passage times through any path of the original network. Other well known approximations for non product-form networks have been developed by SHUM and BUZEN [1977] and MARIE [1979] (see AGRAWAL [1985], pp. 13-16).

Clearly, for obtaining *mean* passage times it suffices to obtain mean queue lengths and throughputs, which indeed are provided by the above-mentioned procedures. Another approach is to apply the MVA algorithm (see Section II.6) directly to non product-form networks, or to apply it after some adjustments of the system parameters which are then inserted into a product-form approximation. Such techniques have led to a generalized MVA algorithm for (i) priority systems (CHANDY and LAKSMI [1983], BRYANT and KRZESINSKI [1983], VAN DOREMALEN, WESSELS and WIJBRANDS [1986]), (ii) systems with type-dependent mean service time at exponential FCFS nodes (BARD [1979]), (iii) systems with a general service time distribution at FCFS nodes (SHUM and BUZEN [1977], VAN DOREMALEN and WESSELS [1988]). These techniques are reviewed in AGRAWAL [1985], pp. 146-153, which monograph is a detailed reference for these and similar approximation procedures.

An interesting method in connection with our themes 'sojourn times', 'passage times', 'response times', is the principle of Response Time Preservation (RTP) transformation of AGRAWAL, BUZEN and SHUM [1984] (in AGRAWAL [1985], pp. 222-259). This transformation leads to a so-called RTP based approximation, which can be characterized as an alternative to NORTON'S theorem. The idea behind this approximation is the following.

In a queueing network which includes some subnetwork that destroys the product form, replace this non product-form subnetwork by an 'equivalent' product-form node. 'Equivalent' here means: The mean passage time through the subnetwork in isolation, with Poisson arrivals, is the same as a customer's mean sojourn time in the equivalent node in isolation with identical input stream. The problem here is the adjustment of the arrival intensity for the Poisson stream arriving at the isolated systems. Because the network does not have a product-form steady state distribution, the throughput for the subnetwork generally cannot be computed. Therefore an iterative procedure is proposed: Starting with an initial guess for the subnetwork's throughput X, the equivalent server is constructed; after having replaced the subnetwork by the equivalent node, the throughput X' can be computed for this node; if X' is 'sufficiently near' X then stop, else repeat the procedure with X' as new guess.

A somewhat related approximation procedure for closed queueing networks is studied by KELLY [1989]. Skipping the concept of the node in isolation, he represents the mean sojourn time at a queue as a function of the throughput at that queue, and derives a set of fixed point equations for the throughputs of the various job classes in the network.

We close this section by mentioning an approximation technique for response time distributions which applies whenever part of a customer's response time consists of a geometrically distributed number of successive cycles within a subnetwork. SALZA and LAVENBERG [1981] have applied this method to a multiprogrammed computer system in which memory contention is represented. The idea behind the approximation is that the sum of a geometrically distributed number of identically distributed random variables (the 'cycle times') is nearly exponentially distributed if the mean number of summands is large. It is somewhat surprising that the approximation seems to be quite good even if the latter condition is not fulfilled. SHANTHIKUMAR and SUMITA [1988] have shown that this exponential approximation fits well for a large class of networks.

III.2. APPROXIMATIONS WITHOUT PRODUCT-FORM ASSUMPTIONS

In Section III.1 we argued that the use of product-form models as approximation often leads to satisfactory results. But this method sometimes has disadvantages. E.g., due to the well known insensitivity properties of symmetric servers (KELLY [1979], p. 77), the mean passage time through a network of symmetric servers does *not* depend on the service time distributions, apart from their first moment. This is a nice property if it really holds, but its application in non product-form networks makes it impossible to study the influence of service time variance on mean passage time in such networks - although this influence may exist.

SHANTHIKUMAR and BUZACOTT [1984] and SHANTHIKUMAR and SUMITA [1988] have suggested the following procedure for computing passage time distributions in general networks with J nodes, Markovian routing, one customer type, independent Poisson arrival streams at the nodes, node-specific service time distributions drawn from independent renewal streams, and FCFS and 'shortest processing time first' (SPT) service disciplines. The total time in system for an arbitrary customer is given by

$$T = \sum_{i=1}^{J} \sum_{r=1}^{N_i} T_{ir}, \tag{3.1}$$

where N_i is the number of visits at node i of the customer, and T_{ir} is this customer's sojourn time at node i at his rth visit there.

The general approximation assumption that has been introduced to analyze T in (3.1) is: 'All sojourn times of the tagged customer at the nodes are independent, and distributed like the sojourn times in the isolated nodes in equilibrium'. (Note that, due to the Markovian routing, the $\{N_i, i = 1,...,J\}$ are independent of the customers' sojourn times at the nodes.) T_{ir} is obtained by some suitable computation in a steady-state $M/G/1$ system with the same service discipline, where the intensity of the arrival process is determined by the standard traffic balance equations for the network. This approximation assumption is stronger than the IFTA assumption of Hohl and Kühn [1988] and Harrison [1981], because the distributions are obtained from isolated nodes. In the case of open product-form networks, the approximation reduces to the IFTA approximation.

Under the above strong assumption, Shanthikumar and Buzacott [1984] obtain an approximate mean $E\hat{T}$ and variance $Var(\hat{T})$ of the passage time, while Shanthikumar and Sumita [1988] present an approximation for passage time distributions. The problem in the latter case remained: How does one compute the sojourn time distribution at an isolated node? For FCFS nodes an explicit expression for (the LST of) this sojourn time distribution is available, whereas for SPT nodes an approximation or simulation must be used. Applying the usual two-parameter approximation technique for non-negative random variables, Shanthikumar and Sumita have developed parametric approximations for the distribution of the total sojourn time T in the network, using only the approximations $E\hat{T}$ and $Var(\hat{T})$. They recommend the following procedure.

Denote by $\beta_s = Var(\hat{T})/[E\hat{T}]^2$ the squared coefficient of variation of the approximate total sojourn time $\hat{T}$.

(1) If $\beta_s \approx 1$ then

$$Pr\{T \leqslant t\} \approx 1 - exp[-t/E\hat{T}], \quad t \geqslant 0,$$

(exponential approximation).

(2) If $\beta_s \ll 1$ then

$$Pr\{T \leqslant t\} \approx 1 - \sum_{i=0}^{k-2} e^{-rt} \frac{(rt)^i}{i!} - ae^{-rt} \frac{(rt)^{k-1}}{(k-1)!}, \quad t \geqslant 0,$$

where

$$k = \lceil \beta_s^{-1} \rceil; \quad a = \frac{b + \sqrt{kb}}{1 + \beta_s}; \quad b = 1 - (k-1)\beta_s; \quad r = \frac{k-1+a}{E\hat{T}};$$

(generalized Erlang approximation).

(3) If $\beta_s \gg 1$ then

$$Pr\{T \leqslant t\} \approx 1 - ae^{-r_1 t} - (1-a)e^{-r_2 t}, \quad t \geqslant 0,$$

where

$$a = \frac{1}{2}(1 - \frac{\sqrt{\beta_s - 1}}{\sqrt{\beta_s + 1}}); \quad r_1 = \frac{2a}{E\hat{T}}; \quad r_2 = \frac{2(1-a)}{E\hat{T}};$$

(hyperexponential approximation with balanced means).

Similar two-parameter approximations were used by several authors to simplify the computations of customer flows and service time distributions in complicated networks: Interarrival processes and service time processes are assumed to be renewal processes with distributions of the renewal times being selected suitably (see Whitt [1983], p. 2783, where further procedures are mentioned). Then the nodes of the network are investigated individually to obtain the required node-specific performance measures; and for obtaining mean, variance and even quantiles of a customer's sojourn time distribution, usually a procedure similar to that of Shanthikumar and Sumita [1988] is applied (see, e.g., Whitt [1983], p. 2810). Whitt [1982] suggests a different approximation for distributions with small variability, which may also be used in this context.

III.3. Light and heavy traffic approximations

Single node systems are known to admit quite accurate closed form simple approximations if the traffic intensity is near zero ('light traffic') or near one ('heavy traffic'). If both light and heavy traffic approximations can be derived, one can use a linear or polynomial interpolation procedure to obtain better medium traffic results. For a short review of such procedures see Reiman and Simon [1988a].

For open *networks* of queues, a heavy traffic approximation based on diffusion processes is well developed. We shall review the results here in some detail. We close the section with a brief discussion of the sparse results on light traffic approximations in queueing networks.

(1) Diffusion approximation

Reiman [1984] considers 'renewal Jackson networks' of FCFS single server nodes $\underset{\sim}{J}=\{1,...,J\}$. The network is open, and customers (of a single type) arrive at the nodes from outside according to independent renewal processes, the interarrival times having mean $\gamma_j^{-1}>0$ and variance $a_j \geqslant 0$, $j \in \underset{\sim}{J}$ ($\gamma_j=0$ is allowed and means: No external arrival at node j). The service times at node j are i.i.d. with mean $\mu_j^{-1}>0$ and variance $s_j \geqslant 0$, and are independent of anything else. The routing is Markovian, governed by an irreducible substochastic matrix $R = (r(i,j),\ i,j \in \underset{\sim}{J})$. The traffic intensity at node $j \in \underset{\sim}{J}$ is $\rho_j := \lambda_j / \mu_j$, with $\{\lambda_j,\ j \in \underset{\sim}{J}\}$ the unique solution of the set of traffic equations $\lambda_j = \gamma_j + \sum \lambda_i r(i,j)$, $j \in \underset{\sim}{J}$.

Reiman considers a sequence of queueing networks of the type described above where the means and variances of the defining quantities converge to finite limits in such a way that $\rho_j(n) \rightarrow 1$ for $n \rightarrow \infty$ ('heavy traffic condition'), where $\rho_j(n)$ is the traffic intensity at node j of the nth network of the sequence, $j \in \underset{\sim}{J}$. Further imposing a technical (Lindeberg) condition on the defining distributions of the network sequence he proves that the sequence of joint queue length processes $\{Q^{(n)}(t),\ t \geqslant 0\}$, $n \in \mathbb{N}$, suitably normalized, converges in distribution.

If $Z^{(n)}(t) := n^{-1/2} Q^{(n)}(nt)$, $0 \leqslant t \leqslant 1$, $n \in \mathbb{N}$, then we have weak convergence $Z^{(n)} = \{Z^{(n)}(t),\ t \in [0,1]\} \rightarrow Z$, $n \rightarrow \infty$, where $Z = \{(Z_1(t), \cdots, Z_J(t)),\ t \in [0,1]\}$ is a reflected (or regulated) Brownian motion on $\mathbb{R}_+^J$, i.e., on the interior of $\mathbb{R}_+^J$ Z behaves like ordinary Brownian motion while on the boundaries (where some node becomes 'empty') it reflects instantaneously with a fixed direction of reflection for each boundary hyperplane. This result is used to approximate sojourn time processes for customers who enter the network at node $k_0 \in \underset{\sim}{J}$ traversing the network with fixed itinerary $(k_0, k_1, \ldots, k_K) =: \mathbb{K}$. For this path set $(h_1,...,h_J) =: \mathbf{h} \in \mathbb{N}^J$, where h_i is the number of visits at node $i \in \underset{\sim}{J}$ during a customer's run through path $\mathbb{K}$. If $W_{k_0,\mathbf{h}}(t)$, $t \geqslant 0$, is the total sojourn time in the network of the next customer arriving after t at node k_0 from outside having 'visit vector' $\mathbf{h}$, and $\alpha_{k_0,\mathbf{h}}^{(n)}(t) = n^{-1/2} W_{k_0,\mathbf{h}}^{(n)}(nt)$, $0 \leqslant t \leqslant 1$, the normalized total sojourn time of this customer in the nth network, then

$$\alpha_{k_0,\mathbf{h}}^{(n)} = \{\alpha_{k_0,\mathbf{h}}^{(n)}(t),\ t \in [0,1]\} \rightarrow \alpha_{k_0,\mathbf{h}}, \quad n \rightarrow \infty, \tag{3.2}$$

in distribution, where

$$\alpha_{k_0,\mathbf{h}} = \sum_{j=1}^{J} \frac{h_j}{\mu_j} Z_j, \tag{3.3}$$

and $Z = \{Z_j,\ j \in \underset{\sim}{J}\}$. (Convergence in distribution ($\rightarrow$) here means convergence in D, the space of $\mathbb{R}_+^J$-valued right-continuous functions having lefthand limits.)

In discussing his results, Reiman points out several easily obtained generalizations, allowing, e.g., dependencies, batch arrivals and different routing types.

The most remarkable observation concerning the limiting result (3.2) is the explicit form of the limiting distribution given by (3.3): It depends on the prescribed path $\mathbb{K}$ only through the visiting vector $\mathbf{h} = (h_1, \ldots, h_J)$, i.e., different paths with possibly different entrance nodes lead to the same passage time distribution for a customer entering the network at a prescribed time, as long as paths have identical visiting vectors. This is a remarkable insensitivity property of the network, as pointed out by REIMAN (it should be noted that up to now we have not imposed any equilibrium assumption on the network). And astonishingly enough, in the diffusion limit all problems with overtaking (cf. Sections II.2, II.3 and II.5) disappear.

REIMAN [1982] elaborates further on the sojourn time diffusion approximation under the assumption that the limit is obtained through a sequence of stationary exponential JACKSON networks. For the case of positive recurrent reflected Brownian motion $Z = \{Z(t),\ t \geqslant 0\}$ (the 'joint queue length' process), HARRISON and REIMAN [1981] obtain a partial differential equation which can be solved explicitly in the present case. Via (3.3) it yields the stationary sojourn time distribution for a customer entering the network at node k_0 with visiting vector $\mathbb{h} = (h_1, \ldots, h_J)$:

$$W_{k_0,\mathbb{h}}(0) \sim \sum_{j=1}^{J} h_j \tau_j; \tag{3.4}$$

here $\tau_1, \ldots, \tau_J$ are independent, and $\tau_j \sim \exp(\mu_j - \lambda_j)$, $j = 1, \ldots, J$, i.e., τ_j is distributed as the equilibrium sojourn time in node j if this node is in isolation with a Poisson(λ_j) arrival stream. The form of (3.4) suggests the following interpretation. Each time the tagged customer enters node j on his itinerary $\mathbb{K}$ (with visiting vector $\mathbb{h}$), he has exactly the same sojourn time at that node. We cite REIMAN [1982], p. 413: *"It is as if the customer takes a snapshot of the network when he enters and all queues remain at the same value during the customer's sojourn throughout the network. This remarkable fact was first pointed out by G.J. FOSCHINI and is a result of his observation that on the diffusion time scale, customers spend zero time in the network."* For closed cyclic queues in heavy traffic, a similar observation was made by KELLY [1984] (see Theorem 2.6).

Let us consider the case of a customer's itinerary $\mathbb{K}$ where all nodes are visited at most once, i.e., the visiting vector $\mathbb{h} = (h_1, \ldots, h_J)$ fulfills $h_i \in \{0,1\}$, $i = 1, \ldots, J$. Then the equilibrium sojourn time distribution is given by

$$W_{k_0,\mathbb{h}}(0) \sim \sum_{j=1}^{J} h_j \tau_j = \sum_{k \in \mathbb{K}} \tau_k,$$

and it follows:

(i) if $\mathbb{K}$ is overtake-free then the diffusion limit is exact (see Section II.2);

(ii) if $\mathbb{K}$ is cycle-free then the diffusion limit gives the same result as the independent flow time approximation (IFTA) of HOHL and KÜHN [1988] and P.G. HARRISON [1981], and the approximation of SHANTHIKUMAR and SUMITA [1988] would also produce the same result in this special case.

Some preparations for, and special comments on, these very general results may be found in the early papers of M.J. HARRISON [1973,1978], where open tandem systems are investigated, and in REIMAN [1988], where a multiclass feedback queue is studied.

(2) Further heavy traffic approximations

BOXMA [1988] has investigated the influence of the slowest server on the cycle time distribution in a closed J-stage tandem of queues. For exponential servers, analytic limiting expressions for the sojourn time distribution are obtained for growing population size (see already Section II.7). For general service time distributions he applies the structural result from the exponential case, supporting the result by simulations and by some simple non-exponential examples for which explicit formulas are known.

WHITT [1984] proposes to approximate large closed networks by open ones. He focuses on queue lengths and throughputs, but his approach should also lead to very comfortable approximations with respect to sojourn time distributions. E.g., an overtake-free path in an exponential FCFS network will lead to a simple convolution approximation of the complicated expression for the sojourn time distribution in the closed network (compare the remarks in Section II.7).

(3) Light traffic approximation

General light traffic approximations for sojourn times in queueing networks have only recently been developed, by REIMAN and SIMON [1988]. The class of networks introduced in their paper consists of generalized open JACKSON networks with Markovian routing and type selection, type-dependent priorities, type-dependent phase-type service time distributions, and Poisson(λ) arrival stream. Let $W(\lambda)$ denote a random variable distributed like the stationary sojourn time of a tagged customer, and

$$W(x,\lambda) = Pr\{W(\lambda)>x\}, \quad x \in \mathbb{R}_+.$$

In REIMAN and SIMON [1989] a method is derived to compute the nth order light traffic limit of $W(\lambda)$, $n=0,1,\ldots$, which is defined by

$$W^{(0)}(x,0) = W(x,0) = \lim_{\lambda \to 0+} W(x,\lambda),$$

and

$$W^{(n)}(x,0) = \lim_{\lambda \to 0+} \frac{1}{\lambda}[W^{(n-1)}(x,\lambda) - W^{(n-1)}(x,0)].$$

It is proved that $W^{(n)}(x,0)$ can be computed by solving the 'k customer problem' for the network, $k=1,\ldots,n+1$, which is:

'Suppose a tagged customer enters a system with at most $k-1$ other customers present. Determine the tagged customer's conditional sojourn time given the positions of the customers at the entrance moment of the tagged customer, and perform the de-conditioning.'

The actual computation is done by using the matrix exponential representation for the tagged customer's sojourn time distribution after deriving the correct initial state distribution which is obtained knowing that at most $k-1$ other customers arrived at the system over $(-\infty,\infty)$ and that the tagged customer arrived at $t=0$. REIMAN and SIMON [1988] investigate, among other things, the light traffic sojourn time distribution for a tandem with a PS and a FCFS node in series, and for feedback queues.

With respect to interpolation, these light traffic results suggest to approximate the sojourn time distribution in medium traffic by using a TAYLOR expansion of $W(x,\lambda)$ in $\lambda=0$, which is suitably adjusted taking into account the earlier discussed heavy traffic results (and possibly using some heuristics). In this way, REIMAN and SIMON [1988a] and FLEMING and SIMON [1989] obtain quite accurate expressions for feedback nodes and for a two-stage tandem queue. SIMON and WILLIE [1986] aggregate results obtained from heavy traffic theory with data from simulation experiments to obtain suitable interpolations for sojourn time characteristics.

III.4. COMPUTATIONAL METHODS FOR PASSAGE TIME EVALUATION IN MARKOVIAN NETWORKS

For ease of computation, in performance analysis one almost always formulates the system dynamics in terms of Markov processes with discrete state space (this can be done using phase-type distributions to approximate general service time distributions). Then the problem of passage (response) time distribution can be transformed into a first entrance time problem for such Markov processes. This is done by lumping together, into one absorbing state Δ, all those states of the system for which the customer under observation is not present on his path. Starting that process in a suitably normalized 'conditional steady state' π, and governed by an intensity matrix Q which is defined such that it lets the process jump into Δ when the customer departs from the last node of his path, the new process is a transient Markov process with one absorbing state. The passage time of the customer in the original network is just the time until absorption of the new process. So one can apply traditional methods from the theory of first entrance times in Markov processes, as well as existing techniques for numerical computations for these problems. These methods are elaborated in the book of KEILSON [1979] from an application-oriented point of view. Although that book does not deal with queueing networks, the general techniques presented there can be carried over directly (the same remark applies to the book of ALDOUS [1989], where general heuristics for Markov chain hitting times are developed).

In this section we are mainly interested in those numerical procedures. We may assume that the process has a finite state space, what has to be guaranteed by some appropriate state space truncation. Then one obtains a set of convolution equations for the set of residual passage time distributions given the actual state of the system. Applying Laplace-Stieltjes transformation yields a finite system of linear equations that can be solved. If the path of the customer is prescribed, then this linear system is recursively solvable (see SCHASSBERGER [1985]). An example in connection with cycle times in the 'equivalent network' of NORTON'S theorem is provided by SCHASSBERGER [1985], p. 124. It should be pointed out that inverting the obtained passage time LST often poses difficulties; but at least one can obtain moments of any order from this LST (the better way usually is: Differentiate the linear system of residual passage time LST's; this yields a recursive scheme for conditional moments of any order).

To obtain the expected passage time, KOHLAS [1986] provides a numerically stable algorithm from the set of mean residual passage times given the actual state of the system. Suitably applying Gaussian elimination, he points out that the elimination of unknowns can be interpreted as transforming the passage time problem into a reduced passage time problem for a semi-Markov instead of a Markov model. Fortunately, this is not a drawback of the reduction process because the procedure can be applied to semi-Markovian absorption problems as well. This is of special interest to our problem because the introduction of phase-type approximations in non-exponential systems usually blows up the state space considerably.

An explicit expression for the *distribution* of the customer's passage time T is given by

$$Pr\{T>s\}=\pi\ exp(Qs)\ e,\quad s\geqslant 0, \tag{3.5}$$

where Q is the intensity matrix of the transient process, π its initial distribution row vector and e the column vector having all entries equal to one. Here one has to evaluate a series of matrix powers: The problem again is to apply suitable numerical procedures.

A well-known technique for computing sojourn time distributions for Markov processes with bounded intensity matrix is the uniformization procedure. This approach has been worked out by MELAMED [1983] and MELAMED and YADIN [1984, 1984a]. Starting-point is the following observation. The fact that the sojourn time distribution of the process in some state is state-dependent exponential, poses considerable numerical difficulties as exponential distributions of different means have to be convoluted to compute the sojourn time distribution. The uniformization procedure introduces dummy jumps for the process evolution, such that the point process counting the jumps of the new system is a Poisson process. At dummy jump instants the system immediately jumps back into the state it just left (such behaviour is not allowed in classical process theory!). Although we have a new (and usually different) counting process for the number of jumps in the system, the Markovian state process of the original and the uniformized system are stochastically identical. In particular, the absorption times are identically distributed. Now the absorption time T of the uniformized process is obtained as follows: Determine the random number N of jumps that occur until absorption. Then

$$Pr\{T<s\} = \sum_{n=1}^{\infty} Pr\{N=n\}[1-e^{-\lambda s}]^{n*},\quad s\geqslant 0, \tag{3.6}$$

where $\lambda>0$ is the intensity of the Poissonian jump counting process obtained by uniformization and where $n*$ denotes n-fold convolution. The essential point of this method: Determining the distribution of N only requires the evaluation of the absorption time distribution of a discrete time Markov chain. So computing the sojourn time distribution is again reduced to matrix computations.

The power of this method was pointed out by MELAMED [1983] who proved, using this approach, that the passage time of a customer through the SIMON-FOLEY network is *not* distributed as the convolution of the successive sojourn times at the nodes (see Section II.2, Example (B), and Section II.5). The uniformization procedure enabled him to numerically prove bounds of the passage time distribution function, thus bounding it away from the convoluted distribution function of the successive sojourn times.

III.5. Queues in series with dependent service times

In Example A of Chapter I it was observed that, in a queueing model of a message switching communication network, the service times of a customer at the successive service stations are generally strongly related: These service times have a fixed ratio, corresponding to the ratio of channel capacities of the channels that those service stations represent. For the case of *identical* service times of a customer at the successive queues of a tandem connection (identical channel capacities), a quite detailed exact analysis of queue lengths, sojourn times etc. is possible. This analysis will be discussed in (i) below. Subsequently some (ii) light traffic and (iii) heavy traffic results will be surveyed.

(i) Queues in series with identical service times

Consider a system of two single server FCFS queues Q_1, Q_2 in series, Q_1 being an $M/G/1$ queue with Poisson(λ) arrival process and service time distribution $B(.)$ with mean μ^{-1}. After completion of his service at Q_1, a customer immediately enters Q_2, *requiring exactly the same service time* as he did in Q_1. Boxma [1979] studies the two-dimensional embedded Markov chain $\{(Z_1^{(n)},T_2^{(n)}),\ n=1,2,...\}$, with $Z_1^{(n)}$ the queue length in Q_1 immediately after the nth departure from Q_1 after $t=0$, and with $T_2^{(n)}$ the sojourn time of this customer in Q_2. The joint distribution of $Z_1^{(n)}$ and $T_2^{(n)}$ is obtained, leading to the following expression for $S^{(2)}(w)$, the steady-state distribution of $T_2^{(n)}$ - which exists iff $\rho := \lambda/\mu < 1$:

$$S^{(2)}(w) = (1-\rho)\frac{B(w)}{1-B(w)}(1-m(w))Y(w), \quad w \geqslant 0; \tag{3.7}$$

where $m(w)$ is the steady-state distribution of the supremum of the service times of the customers served in a busy period of Q_1, determined for all $w \geqslant 0$ as the unique zero in $[0,1]$ of

$$m(w) = \int_{t=0}^{w-} \exp[-(1-m(w))\lambda t]\, dB(t), \tag{3.8}$$

and $Y(w)$ is the steady-state distribution of the amount of work in Q_2 at an epoch at which a busy period of Q_1 starts:

$$Y(w) = \exp[-\lambda \int_{q=w}^{\infty} (1-m(q))\, dq]. \tag{3.9}$$

In fact,

$$G(w) := (1-\rho)\frac{B(w)}{1-B(w)}(1-m(w)), \quad w \geqslant 0,$$

is a proper probability distribution if $\rho<1$; it is the limiting distribution for $n \to \infty$ of the supremum of the service times in Q_1 of the nth customer, C_n, and of the customers who have arrived before C_n and belong to the same busy period of Q_1 as C_n.

Formula (3.7) ($S^{(2)}(w)=G(w)Y(w)$) can readily be interpreted by realizing that, with τ_{n+1} the service time of C_{n+1} at both servers and δ_{n+1} the time interval that Q_1 is empty between the departures of C_n and C_{n+1} from Q_1,

$$\begin{aligned} T_2^{(n+1)} &= \max\{T_2^{(n)},\tau_{n+1}\} && \text{if } Z_1^{(n)}>0, \\ &= \max\{T_2^{(n)}-\delta_{n+1},\tau_{n+1}\} && \text{if } Z_1^{(n)}=0. \end{aligned} \tag{3.10}$$

The joint steady-state distribution of the waiting times of a customer at Q_1 and Q_2 is also calculated

in BOXMA [1979], Part I. The results are used in Part II of that paper to compare the correlation of the waiting times for two $M/M/1$ queues in series with identical respectively i.i.d. service times at the two queues. Identical service times lead to a high positive correlation of waiting times, much higher than that calculated by KRÄMER [1973] for the case of independent service times. In Part II the results of Part I are also used to make a comparison - mainly using numerical and asymptotic techniques - of the queueing behaviour at the first and second queue. In the case of exponential service times, this also yields a comparison of the queueing behaviour at the second of two queues in series with identical and with independent service times, respectively. These comparisons make it possible to assess the influence of the first queue on the second queue, and to give an indication of the accuracy of the *Independence Assumption* (see Example A of Chapter I) in tandem models. The main quantities under consideration are $R(E,S)$ respectively $R(V,S)$, the ratio of expectation respectively variance of sojourn times at Q_2 and Q_1. Apart from extensive numerical results, the following heavy traffic results are presented in BOXMA [1979, Part II].
Without any assumption on $B(.)$, one can already show that, for $\rho \to 1$, $R(E,S) \to 0$ and $R(V,S) \to 0$; if $B(.)$ has a finite support, stronger statements can be obtained:

$$R(E,S) \to 0 \quad \text{as} \quad 1-\rho \text{ for } \rho \to 1;$$

$$R(V,S) \to 0 \quad \text{as} \quad (1-\rho)^3 \text{ for } \rho \to 1.$$

Indeed, Formula (3.10) shows that $S^{(2)}(w)$ has the same finite support as $B(.)$. Simulation results of KLEINROCK [1964] and of MITCHELL et al. [1977], for $J \geqslant 2$ queues in series, also underline the strong regularizing influence of a positive correlation of service times on sojourn times.

CALO [1979] proves for a J-stage tandem, with identical service times at all queues, that the sojourn times experienced by any customer at the successive queues $Q_2, Q_3, \cdots, Q_J$ are nondecreasing, this relationship holding without any assumptions regarding the stochastic nature of the interarrival process and service time distribution. In CALO [1981] the arrival process at Q_1 is taken to be a Poisson process, and the service time distribution is a step distribution with two steps. CALO then determines the LST of the steady-state distribution of the total waiting time of a customer in the system, and the mean waiting times at all stages. MACFADYEN and EVERITT [1984] combine his results with light traffic results to approximate mean waiting times in a tandem queue with successive service times of a customer being scaled versions of each other ("unequal channel capacities").
An important contribution is made by VINOGRADOV [1986] for the case of a J-stage tandem queue with Poisson arrival process and identical service times. He obtains the joint distribution of the sojourn time in Q_1 and of the total sojourn time in the rest of the system.

(ii) Light traffic
As remarked above, in the model of two queues in series with identical service times and heavy traffic, there is a strong reduction of mean and variance of sojourn times in Q_2 as compared to Q_1. On the other hand, it had been observed in BOXMA [1979, Part II] that in light traffic the mean sojourn time at Q_2 may *exceed* that at Q_1. A more detailed study of this phenomenon has been undertaken in PINEDO and WOLFF [1982], WOLFF [1982a]. The first of these studies concerns the two-stage tandem model with identical exp(μ) service times. It is shown that, for $\rho \to 0$, the mean waiting time at Q_2 is 1.75 times as large as the mean waiting time at Q_1 (or at Q_2 in the model with *independent* exp(μ) service times). More general light traffic results are derived in WOLFF [1982a]. Here the model under consideration is a J-stage tandem queue with Poisson arrivals, where the J successive service times of a customer have an arbitrary joint distribution. For the case of identical service times, an explicit light traffic formula for the total mean delay is derived.

(iii) Heavy traffic
VINOGRADOV [1984,1986] considers a J-stage tandem queue with Poisson(λ) arrival process at the first queue and with identical service times at all stages. He derives the heavy traffic limiting behaviour of

the properly normalized joint distribution of the sojourn time at Q_1 and of the total sojourn time in the rest of the system. In the case of exp(1) service times, he obtains

$$E[T_2+ \cdots +T_J] \approx (J-1)\ \ln[(1-\lambda)^{-2}] \quad \text{for } \lambda \rightarrow 1.$$

MAKARICHEV [1984] restricts himself to the two-queue model, but he allows various service disciplines. Again, the emphasis is on heavy traffic behaviour.

The results discussed in this section expose the fact that, in tandem configurations, application of the Independence Assumption generally leads to large errors. Probably the tandem model and the feedback model present the worst cases for this approximative assumption; it would be interesting to undertake a theoretical investigation of other networks with dependent service times, perhaps applying a light and heavy traffic analysis.

REMARK 3.1.
Tandem queues with identical service times and *finite intermediate waiting rooms* have also been studied. We mention the simulation results of MITCHELL et al. [1977] for intermediate waiting rooms of sizes zero and one and blocking; the light traffic results of PINEDO and WOLFF [1982] for zero intermediate waiting rooms and blocking; and the exact two-queue analysis of BOXMA [1984a] for a finite intermediate buffer and overflow of the excess part of a message.

III.6. NETWORKS OF QUEUES WITH DETERMINISTIC SERVICE TIMES

A packet switching communication network (PSCN) is a communication network in which the messages, upon arrival at the network, are decomposed into packets of fixed length. Via channels with a bounded capacity, the packets are transmitted in a store-and-forward manner to their destination, where they are re-assembled. The difference of an MSCN is that in the latter a message travels in its entirety from centre to centre. A PSCN provides a prime example of a queueing network with deterministic service times. In this section we mainly consider such queueing networks, concentrating on (i) the influence of the order of queues in a tandem connection on total sojourn time, and (ii) waiting and sojourn time distributions in queueing networks with deterministic service times.

(i) The influence of the order of queues in a tandem connection on total sojourn time

As observed in Remark 2.1 (3), the joint sojourn time distribution at the queues of a closed cycle of exponential single server FCFS stations does not depend on the order of the queues. Hence the total sojourn time distribution also does not depend on this order. The same holds for the *open* tandem connection of Theorem 2.2. In fact, the interchangeability of $./M/1$ queues in series with respect to the departure process from the last queue has been proved by WEBER [1979] for a *general* arrival process.

In the sixties, a similar interchangeability property had been shown to hold for open tandem queues with *deterministic* service times. FRIEDMAN [1965] considers a model consisting of J FCFS service stations $Q_1, \ldots, Q_J$ in series, station Q_i having infinite waiting room and containing m_i parallel servers each of which provides the same constant service time s_i. The interarrival times at Q_1 have a general distribution. FRIEDMAN shows that *the time spent in the system by each customer is independent of the order of the stages.* This observation leads to a considerable simplification of the analysis: At a stage with values of m_i and s_i such that s_i is smaller than any of the values $s_j \lfloor m_i / m_j \rfloor$, $j=1,...,i-1$, no waiting occurs; hence this stage can be discarded in the waiting time analysis, and an equivalent reduced system is obtained. Consequently a suitable ordering of stages with respect to the values of m_i and s_i may lead to a strongly reduced system. In particular, when each stage contains exactly one server, a reduction to a single stage model with service time the longest of the service times can always be established. A similar reduction is possible if one of the stations Q_i is a single server queue with variable service times that are at least as large as s_j / m_j for all $j \neq i$.

Independently of FRIEDMAN, AVI-ITZHAK [1965] studied a very similar model with constant service times but with finite waiting room at each stage except at the first one; each stage has the same number of servers. He also proved the independence of order of stages, and the resulting reduction properties. Suzuki and Kawashima [1974] showed that the results concerning independence of order and reduction of stages also hold if in Avi-Itzhak's model the number of servers in stage i depends on i.

A fundamental extension of part of Friedman's work is due to Tembe and Wolff [1974]. They show that, in general, the order of the queues in series *does* matter. Subsequently they discuss the *optimal* order - with regard to total waiting time etc. - of a class of single server queues in series with infinite intermediate waiting rooms, proving stochastic ordering results. For example, for two queues in series with *non-overlapping* service times ($Pr\{\tau_n^{(2)} \geqslant \tau_n^{(1)}\}=1$ for service times $\tau_n^{(1)}$ and $\tau_n^{(2)}$ at Q_1, Q_2), they prove that the total sojourn time is stochastically smaller when the longer service time is performed at Q_1. They extend this result for $J \geqslant 2$ queues in series with non-overlapping service times, showing that it is optimal to order the queues in decreasing order of service times.

These ordering properties are essentially sample path properties, that do not depend on the stochastic nature of the arrival process. To make quantitative statements about the influence of the order of the queues, one has to specify the arrival process. Boxma [1979a] studies a two-stage tandem queue with a Poisson arrival process at the first stage and with non-overlapping service times. For the case that the shorter services are performed at the first stage, he determines the LST of the joint distribution of the waiting times at both queues, and of the distribution of the total waiting time in the system. For the reversed case, waiting times at the second stage are zero. Hence a comparison between these two cases is now easily made, yielding quantitative information about the influence of the order of the queues.

Remark 3.2.
Whitt [1985] and Greenberg and Wolff [1988] use approximation methods to devise heuristic design principles for the optimal order of queues in tandem with general (non-deterministic) service times. Whitt's method is based on approximating the arrival process to each server by a renewal process characterized by the first two moments, and studying each queue as a $GI/G/1$ queue. Greenberg and Wolff apply a light traffic approximation for 2-stage tandem queues to select the order of the two queues that minimizes the mean total sojourn time. They allow dependence between the successive sojourn times of a customer. The design principles developed in these two papers do not agree; this seems to imply that one should be very cautious in applying approximations to develop order design procedures.

(ii) Waiting and sojourn time distributions in queueing networks with deterministic service times
The concepts of interchangeability and reduction of stages in tandem queues with deterministic service times are extremely useful in the analysis of a PSCN, as shown by Rubin [1974,1975,1976]. In his 1974 paper, Rubin essentially rediscovers Friedman's results for the case of a tandem queue of FCFS single servers, with a Poisson arrival process at the first queue and with deterministic service times. In particular, the total waiting time distribution along the path is shown to be the waiting time distribution in an $M/D/1$ queue with service time the longest service time in the tandem model. An analysis of a similar tandem model, but with general arrival process and infinite or finite intermediate waiting rooms (cf. the study of Avi-Itzhak [1965]) is performed by Labetoulle and Pujolle [1976]. They apply diffusion approximation methods to estimate mean sojourn times at the stations.

Rubin [1975] uses his 1974 results to study an isolated path in a PSCN. *Messages* arrive according to a Poisson process at the first stage. There they are subdivided into fixed-length packets, which are sent independently along the path; after the last stage, messages are re-assembled. An expression for the message delay time along the path is obtained. This quantity is observed to depend only on the longest service time (minimal channel capacity).

In the above studies, a path was considered in isolation. In Rubin [1976] the interference of other packets at the path, arriving from outside and departing, is taken into account. An approximation is presented for mean delay at a service station, as well as approximation procedures to estimate the mean sojourn time of a message in a PSCN.
Shalmon and Kaplan [1984] are able to give an exact analysis of a tandem queue with deterministic service times *and* interfering traffic, under the following two restrictive assumptions: (i) service times are non-decreasing downstream along the path, and (ii) there can only be intermediate arrivals (according to compound Poisson processes), but no intermediate departures. As a result, busy periods at the stations 'do not break' in the direction of flow. This key property is exploited to derive the LST of the joint steady-state distribution of the waiting times at the queues. As a by-product, the LST of the end-to-end delay for each source feeding into the path is obtained. In this study, service at all nodes is FCFS; Shalmon [1987], for the case of identical constant service times at all nodes, also allows (i) priority for the endogenous stream over outside arrivals, and (ii) alternating priority, where at each node the exogenous and endogenous arrivals are served exhaustively in alternating order. Konheim and Reiser [1977] had previously also studied a tandem model with deterministic service times and with intermediate arrivals (according to independent renewal processes); in their case all service times are identical, and intermediate departures are allowed. A special feature of their model is that at each stage the packets that have to travel the longest distance along the path have priority. Konheim and Reiser obtain the delay distribution for a packet originating at Q_i and departing at Q_j, $1 \leq i \leq j \leq J$.

In this and the previous section we have discussed service time structures that generally lead to a quite regular customer behaviour, with a reduced mean, and in particular a reduced variance, of sojourn times. Fendick et al. [1988] observe that in packet communication networks, several dependencies occur which may have quite an opposite effect on sojourn times. E.g., traffic is often very bursty, and there are multiple classes of traffic (packetized voice, short data messages, bulk files, etc.). The burstiness, which is sometimes characterized by dependence between interarrival times, implies that packets from the same source tend to be served consecutively. The associated dependence between interarrival and service times, combined with the dependence between interarrival times, can lead to a considerable deterioration of performance, with relatively long delays. Fendick et al. capture this phenomenon analytically by considering a multi-class single server queue with batch Poisson arrival process. Although the simple batch Poisson model does not describe packet delays well under light-to-moderate loads, it is shown using heavy traffic limit theorems that the model accurately reflects the limiting degradation of performance as the load increases.

III.7. Miscellanea in non product-form networks

There is an enormous number of papers on ad-hoc methods for determining and approximating sojourn time distributions for customers in queueing networks (or their moments). Most of these papers are restricted to the problem of mean steady-state sojourn times which can be obtained via Little's theorem. Instead of listing these papers we describe in this section a few models and approaches that - in our opinion - either cover important classes of - mostly unsolved - problems, or by dealing with specific systems may give insight into methods applicable to more general problems.

We have omitted a discussion of the extensive literature on waiting times in single server multi-queue systems (polling systems), referring to Takagi [1990] instead. We have also refrained from discussing the rather scarce literature on exact results for sojourn times in such networks as slotted ALOHA and Carrier Sense Multiple Access networks (cf. Tobagi [1982]) and Interconnection networks (cf. Kruskal, Snir and Weiss [1988]).

(i) Single server feedback queues

In Section II.5 we have presented a queue with feedback as an example of overtaking due to the topological structure of the network. The references we mentioned there concerned the product-form case, with the exception of the early paper of Takács [1963] on the $M/G/1$ queue with Bernoulli feedback.

For the same $M/G/1$ queue with Bernoulli feedback, DISNEY, KÖNIG and SCHMIDT [1984] derive sojourn time distributions for the next visit of a newly arriving customer, of a fed back customer, and of an arbitrary customer (either new or fed back). Another recent extension of the work of TAKÁCS is given by DOSHI and KAUFMAN [1988], who obtain the joint sojourn time distribution of a customer at successive loops for this model. Their results enable them to investigate the (in)accuracy of the often used assumption of independence of the successive sojourn times of a customer. A related study is presented by FONTANA and DIAZ BERZOSA [1985], who consider an $M/G/1$ queue with N nonpreemptive priorities; after a service completion, a customer may generate one or more new customers having the same or different priorities. They obtain the LST of the sojourn time of a particular sequence of customers. SIMON [1984] also studies an $M/G/1$ queue with multiple priorities - both nonpreemptive and preemptive. A customer of type i pays exactly $N(i)$ consecutive visits, at the kth visit having priority level $f(i,k)$ and service request with distribution $G_{i,k}$. SIMON derives a set of linear equations for the mean sojourn time at each visit.

(ii) Two-stage closed networks

Closed two-stage tandem queues with a fixed number of customers cycling are models for multiprogrammed computer systems with a fixed level of multiprogramming (see Section II.3 and Remark 1.2). It is known for this application that the coefficients of variation of the service times at the two stages (CPU and I/O) usually differ from one. This leads to the analysis of closed non-exponential tandem queues. BOXMA [1983] has computed the joint distribution of the sojourn times in a system where a cycle consists of a visit to a node with general service time distribution followed by a visit to an exponential node. Interestingly, this joint distribution differs from the one for the reversed situation, where the exponential node is visited first. For the case of only two customers and at least one exponential node, it is shown that the joint distributions only agree when *both* service time distributions are exponential.

DADUNA [1984a] has studied the two-stage model where the cycle starts with a visit to the exponential node. He derives a recursive scheme for obtaining the LST, and moments, of the cycle time. CARBINI et al. [1986] give an algorithm to compute the cycle time distribution when the second stage has a Coxian service time distribution. For the case of both stages having general service time distributions, DADUNA [1986] provides a recursive scheme for conditional cycle time LST's given the system state at the beginning of a cycle (see also BOXMA [1986]).

(iii) Open tandem systems

CHEN [1989] shows that one cannot expect such nice results as the independence of sojourn times in open tandems of exponential queues (Theorem 2.1) to remain valid in the case of non-exponential service times. For a two-stage of a deterministic server followed by an exponential server (fed by a Poisson stream) he proves dependence of a customer's successive sojourn times.

(iv) Jackson networks with breakdowns

Recently some reliability issues in queueing networks have been tackled by analytical methods. It has always been apparent that the possibility of server breakdowns should be included in the network modelling process. However, analytical difficulties have up to now prevented successful research to go beyond (i) the introduction of artificial customers in product-form networks whose service times represent the breakdowns of servers, or (ii) decreasing a server's service rate to take into consideration the throughput degradation caused by breakdowns. In particular, the sojourn time problem in networks with breakdowns is largely unsolved. An interesting recent study, which exposes some of the problems one has to overcome in analytical investigations of breakdown phenomena in queueing networks, is MIKOU [1988]. He considers a two-node open network with FCFS single servers, and Bernoulli feedback from node 2 to node 1. The network is subject to breakdowns. If a breakdown occurs, both servers come to a complete stop until a repair is performed. The customer arrival process is *not* interrupted. All distributions are assumed to be exponential. MIKOU reduces the problem of determining the steady-state joint queue length distribution to a RIEMANN-HILBERT boundary value problem. This enables him to derive a computational algorithm for the determination of the mean sojourn time in the system.

(v) Fork-join queueing networks

Performing computations for a program on parallel computers brings new features into the modelling process for such 'networks'. A program may be split into several subprograms which are executed independently (splitting a 'customer' at a FORK queue into several 'subcustomers' who may to some extent proceed independently through the network). On the other hand, for some further computation the previous execution of several subprograms may be required (assembling some 'subcustomers' at a JOIN queue into one 'customer', who then travels further through the network). Very little is known about sojourn times for such fork-join networks, apart from some approximations. BACCELLI, MASSEY and TOWSLEY [1987] prove some qualitative sojourn time results, using stochastic ordering methods. BRUN and FAYOLLE [1988] derive an exact expression for the LST of the sojourn time distribution in the simplest fork-join queue: Customers, arriving according to a Poisson process, are split into two customers feeding into two $./M/1$ queues; when both services are completed, the two customers are rejoined and the sojourn ends. The final form of the LST renders little hope for getting explicit results for more complicated network configurations.

Recently, networks including fork-join structures have received considerable interest in the area of flexible manufacturing systems. For example, join queues are the classical 'assembly-like' queues modelling production lines that require several incoming items for producing a further item.

(vi) Resequencing

In many computer-communication systems, a situation arises where a stream of tasks, arriving at a node, must be processed in a specific order that may differ from the order of arrival. In such a case, the tasks must first be resequenced at the node, at the expense of a resequencing delay. One context in which the resequencing problem has been addressed is the retransmission of erroneous messages, as in a selective-repeat ARQ (automatic repeat request) protocol. Under this protocol, the transmitter is continuously sending new messages and upon receipt of a negative acknowledgment only the corresponding message is retransmitted. The messages are stored in the receiving buffers until they can be further transmitted in the original order. ROSBERG and SHACHAM [1987] determine the resequencing delay distribution for this situation.

Resequencing problems also occur in error-free networks, where the possibility of messages traversing multiple parallel channels gives rise to disorder. In several studies, the disorder is modeled by an $M/G/\infty$ system. BACCELLI et al. [1984] analyze the total delay experienced by the customers, including the disorder delay, the resequencing delay and the delay at the final receiving server. They use a WIENER-HOPF type factorization method. STAFYLOPATIS and GELENBE [1988] restrict themselves to the resequencing delay distribution at the $M/G/\infty$ queue; relaxing the strict ordering rule, they consider three different partial orderings. We also refer to the latter study for further references on this kind of resequencing problem.

In distributed data bases, several control mechanisms have been developed to preserve consistency, using some timestamp procedure to determine the order in which updates must be performed. BACCELLI [1988] presents an interesting investigation of the impact of such a timestamp ordering algorithm on the performance of a fully replicated distributed data base. He pays special attention to the three synchronization primitives that arise, viz., *fork, join* and *resequencing*. He describes their interaction in terms of a queueing network model. Using analytical methods and stochastic ordering techniques that were developed, for example, in BACCELLI et al. [1984] and BACCELLI et al. [1987], he determines (i) the maximum throughput of updates that the system can process without instability, and (ii) computable upper bounds on the system response time (defined as the time required to update all copies).

ILIADIS and LIEN [1988] study the resequencing problem in a queueing system with two heterogeneous

exponential servers, which model two links of different speeds in a communication network. Two threshold-type policies are considered, that determine which server serves a particular customer. After a fairly straightforward Markovian analysis, those authors derive closed-form solutions for the resequencing delay distributions under both policies.
BACCELLI and MAKOWSKI [1989] present an extensive survey of the literature on resequencing.

(vii) Networks with blocking
Hardly any research has been published concerning the important topic of sojourn time distributions in networks with finite-capacity nodes and blocking. This research area is opened by the study of BALSAMO and DONATIELLO [1987]; they provide a recursive technique to compute the cycle time distribution in a cyclic two-stage network with exponential servers and blocking.

IV. CONCLUSIONS

This paper has presented a survey of the state-of-the-art in computing sojourn time distributions in queueing networks, a problem that is interesting and challenging both from a mathematical and an engineering viewpoint. Our aim has been

- to give a brief reference manual for the methods available to apply analytical and computational procedures, and
- to motivate and suggest further research in this - as we still believe - complicated but promising field, by elucidating the structure of the essential difficulties.

For a somewhat different presentation of structural properties connected with the sojourn time problem, we refer to Chapter 4 of WALRAND [1988] where a lucid account is given of some of the central issues concerning sojourn times and customer flows in networks. But although the theory is beginning to find its way into textbooks, the reader will have noticed that explicit general formulas for sojourn time distributions are rare, and that, although progress has been made in developing approximation techniques, much work still remains to be done in that area.

References

S.C. Agrawal (1985). *Metamodeling - a Study of Approximations in Queueing Models,* The MIT Press, Cambridge.

S.C. Agrawal, J.P. Buzen and A.W. Shum (1984). *Response time preservation: A general technique for developing approximate algorithms for queueing networks,* Proc. of the 1984 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems; reprint in Agrawal [1985].

D. Aldous (1989). *Probability Approximations via the Poisson Clumping Heuristic* (Springer, Berlin).

B. Avi-Itzhak (1965). *A sequence of service stations with arbitrary input and regular service time,* Management Sci. 11, 565-571.

F. Baccelli, E. Gelenbe and B. Plateau (1984). *An end-to-end approach to the resequencing problem,* J. Assoc. Comput. Mach. 31, 474-485.

F. Baccelli, W.A. Massey and D. Towsley (1987). *Acyclic fork-join queueing networks,* Report INRIA No. 688, to appear in J. Assoc. Comput. Mach.

F. Baccelli (1988). *A queueing model of timestamp ordering in a distributed system,* In: Performance '87, eds. G. Latouche and P.J. Courtois (North-Holland Publ. Cy., Amsterdam) pp. 413-431.

F. Baccelli, A.M. Makowski (1989). *Queueing models for systems with synchronisation constraints,* to appear in Proc. of the IEEE, Special Issue on discrete-event systems.

F. Baccelli, W.A. Massey (1988). *A transient analysis of the two-node series Jackson network,* Report INRIA.

F. Baccelli, W.A. Massey and P.E. Wright (1988). *Determining the exit time distribution for a closed cyclic network,* Report INRIA.

S. Balsamo, L. Donatiello (1987). *On the cycle time distribution in a two-stage network with blocking,* Report Department of Computer Science, University of Pisa.

A.D. Barbour, R. Schassberger (1981). *Insensitive average residence times in generalized semi-Markov processes,* Adv. in Appl. Probab. 13, 720-735.

Y. Bard (1979). *Some extensions to multiclass queueing network analysis,* In: Performance of Computer Systems, eds. M. Arato, A. Butrimenko and E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 51-62.

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.

M. Berg, M.J.M. Posner (1985). *Customer delays in M/M/c repair systems with spares,* Naval Res. Logist. Quart. 32, 287-299.

M. Berg, M.J.M. Posner (1986). *On the regulation of queues,* Oper. Res. Letters 4, 221-224.

J.L. van den Berg, O.J. Boxma and W.P. Groenendijk (1989). *Sojourn times in the M/G/1 queue with deterministic feedback,* Stochastic Models 5, 115-129.

J.L. van den Berg, O.J. Boxma (1989). *Sojourn times in feedback queues,* In: Operations Research Proceedings 1988, eds. D. Pressmar et al. (Springer, Berlin) pp. 247-257.

O.J. Boxma (1979). On a tandem queueing model with identical service times at both counters, I, II, Adv. in Appl. Probab. 11, 616-643, 644-659.

O.J. Boxma (1979a). *Two queues in series with non-overlapping service times,* In: Proc. ITC-9.

O.J. Boxma, P. Donk (1982). *On response time and cycle time distributions in a two-stage cyclic queue,* Performance Evaluation 2, 181-194.

O.J. Boxma (1983). *The cyclic queue with one general and one exponential server,* Adv. in Appl. Probab. 15, 857-873.

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.

O.J. Boxma (1984a). *Two identical communication channels in series with a finite intermediate buffer and overflow,* In: Modelling and Performance Evaluation Methodology, eds. F. Baccelli and G. Fayolle, Lect. Notes in Control and Inf. Sci. 60 (Springer, Berlin) pp. 613-638.

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.

O.J. Boxma (1986). *Models of two queues: A few new views,* In: Teletraffic Analysis and Computer Performance Evaluation, eds. O.J. Boxma, J.W. Cohen and H.C. Tijms (North-Holland Publ. Cy., Amsterdam) pp. 75-98.

O.J. Boxma (1988). *Sojourn times in cyclic queues - the influence of the slowest server,* In: Computer Performance and Reliability, eds. G. Iazeolla, P.J. Courtois and O.J. Boxma (North-Holland Publ. Cy., Amsterdam) pp. 13-24.

S.C. Bruell, G. Balbo (1980). *Computational Algorithms for Closed Queueing Networks,* North-Holland Publ. Cy., New York.

M.A. Brun, G. Fayolle (1988). *The distribution of the transaction processing time in a simple fork-join system,* In: Computer Performance and Reliability, eds. G. Iazeolla, P.J. Courtois and O.J. Boxma (North-Holland Publ. Cy., Amsterdam) pp. 203-212.

R.M. Bryant, A.E. Krzesinski (1983). *The MVA pre-empt resume priority approximation,* pre-RC, IBM Thomas J. Watson Research Center, Yorktown Heights (NY).

P.J. Burke (1956). *The output of a queueing system,* Oper. Res. 4, 699-704.

P.J. Burke (1964). *The dependence of delays in tandem queues,* Ann. Math. Stat. 35, 874-875.

P.J. Burke (1968). *The output process of a stationary M/M/s queueing system,* Ann. Math. Statist. 39, 1144-1152.

P.J. Burke (1969). *The dependence of sojourn times in tandem M/M/s queues,* Oper. Res. 17, 754-755.

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.

X.R. Cao, Y.C. Ho (1987). *Estimating the sojourn time sensitivity in queueing networks using perturbation analysis,* J. Optimization-Theory and Appl. 53, 353-375.

S.B. Calo (1979). *Delay properties of message channels,* In: Proc. 1979 Int. Conf. Commun., Boston (MA), pp. 43.5.1-43.5.4.

S.B. Calo (1981). *Message delays in repeated-service tandem connections,* IEEE Trans. Commun. 29, 670-678.

S. Carbini, L. Donatiello and G. Iazeolla (1986). *An efficient algorithm for the cycle time distribution in two-stage cyclic queues with a non-exponential server,* In: Teletraffic Analysis and Computer Performance Evaluation, eds. O.J. Boxma, J.W. Cohen and H.C. Tijms (North-Holland Publ. Cy., Amsterdam) pp. 99-115.

K.M. Chandy, U. Herzog and L. Woo (1975). *Parametric analysis of queueing networks,* IBM J. Res. Develop. 19, 36-42.

K.M. Chandy, U. Herzog and L. Woo (1975a). *Approximate analysis of general queueing networks,* IBM J. Res. Develop. 19, 43-49.

K.M. Chandy, M.S. Laksmi (1983). *An approximation technique for queueing networks with preemptive priority queues,* Technical Report, Dept. of Computer Science, University of Texas at Austin, Austin (TX).

T.M. Chen (1989). *On the independence of sojourn times in tandem queues,* Adv. in Appl. Probab. 21, 488-489.

W.-M. Chow (1980). *The cycle time distribution of exponential cyclic queues,* J. Assoc. Comput. Mach. 27, 281-286.

E.G. Coffman, Jr., G. Fayolle and I. Mitrani (1986). *Sojourn times in a tandem queue with overtaking: Reduction to a boundary value problem,* Stochastic Models 2, 43-65.

J.W. Cohen (1979). *The multiple phase service network with generalized processor sharing,* Acta Informatica 12, 245-284.

A.E. Conway, N.D. Georganas (1986). *RECAL - a new efficient algorithm for the exact analysis of multiple-chain closed queueing networks,* J. Assoc. Comput. Mach. 33, 768-791.

H. Daduna (1982). *Passage times for overtake-free paths in Gordon-Newell networks,* Adv. in Appl.

Probab. 14, 672-686.

H. DADUNA (1983). *On passage times in Jackson networks: Two-stations walk and overtake-free paths,* Zeitschr. für Oper. Res. 27, Ser. A, 239-256.

DADUNA, R. SCHASSBERGER (1983). *Networks of queues in discrete time,* Zeitschr. für Oper. Res. 27, Ser. A, 159-175.

H. DADUNA (1984). *Burke's theorem on passage times in Gordon-Newell networks,* Adv. in Appl. Probab. 16, 867-886.

H. DADUNA (1984a). *The cycle time distribution of cyclic two-stage queues with a non-exponential server,* In: Modelling and Performance Evaluation Methodology, eds. F. Baccelli and G. Fayolle, Lect. Notes in Control and Inf. Sci. 60 (Springer, Berlin) pp. 641-653.

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, Berlin), pp. 127-140.

H. DADUNA (1985a). *The cycle-time distribution in a central server network with state-dependent branching,* Optimization 16, 617-626.

H. DADUNA (1986). *Two-stage cyclic queues with nonexponential servers: Steady-state and cycle time,* Oper. Res. 34, 455-459.

H. DADUNA (1986a). *Cycle times in two-stage closed queueing networks: Applications to multiprogrammed computer systems with virtual memory,* Oper. Res. 34, 281-288.

H. DADUNA (1986b). *Repair-times in a two-echelon repair system with control,* In: M.J. Beckmann, K.-W. Gaede, K. Ritter and H. Schneeweiss (eds.), Methods of Operations Research 53, 375-386.

H. DADUNA (1987). *Exchangeable items in repair systems: Delay times,* Report Institute of Mathematical Stochastics, University of Hamburg, to appear in Oper. Res.

H. DADUNA (1987a). *Cycle times in a starlike network with state dependent routing,* J. Appl. Math. and Simulation 1, 1-12.

H. DADUNA (1989). *On network flow equations and splitting formulas for sojourn times in queueing networks,* Report Institute of Mathematical Stochastics, University of Hamburg.

H. DADUNA (1989a). *The method of adjusted transfer rates for computing delay time distributions in data communication systems with window flow control,* Report Institute of Mathematical Stochastics, University of Hamburg.

R.L. DISNEY, D. KÖNIG, V. SCHMIDT (1984). *Stationary queue-length and waiting-time distributions in single-server feedback queues,* Adv. Appl. Probab. 16, 437-446.

J. VAN DOREMALEN, J. WESSELS AND R. WIJBRANDS (1986). *Approximate analysis of priority queueing networks,* In: Teletraffic Analysis and Computer Performance Evaluation, eds. O.J. Boxma, J.W. Cohen and H.C. Tijms (North-Holland Publ. Cy., Amsterdam) pp. 117-131.

J. VAN DOREMALEN, J. WESSELS (1988). *A recursive aggregation-disaggregation method to approximate large-scale closed queueing networks with multiple job types,* In: Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen, eds. O.J. Boxma and R. Syski (North-Holland Publ. Cy., Amsterdam) pp. 325-342.

B.T. DOSHI, J.S. KAUFMAN (1988). *Sojourn time in an M/G/1 queue with Bernoulli feedback,* In: Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen, eds. O.J. Boxma and R. Syski (North-Holland Publ. Cy., Amsterdam) pp. 207-233.

G. FAYOLLE, R. IASNOGORODSKI AND I. MITRANI (1983). *The distribution of sojourn times in a queueing network with overtaking: Reduction to a boundary problem,* In: Performance '83, eds. A.K. Agrawala and S.K. Tripathi (North-Holland Publ. Cy., Amsterdam) pp. 477-486.

K.W. FENDICK, V.R. SAKSENA AND W. WHITT (1988). *Dependence in packet queues,* Report AT&T Bell Laboratories, Holmdel (NJ), to appear in IEEE Trans. Commun.

P.J. FLEMING, B. SIMON (1989). *Interpolation approximations of response time distributions,* Preprint.

R.D. FOLEY, R.L. DISNEY (1981). *Queues with delayed feedback,* Technical Report VTR 8106, Virginia Polytechnic Institute and State University, Blacksburg (VA).

R.D. FOLEY, P.C. KIESSLER (1989). *Positive correlations in a three-node Jackson queueing network,* Adv. in Appl. Probab. 21, 241-242.

B. Fontana, C. Diaz Berzosa (1985). *M/G/1 queue with N-priorities and feedback: Joint queue-length distributions and response time distribution for any particular sequence,* In: Teletraffic Issues in an Advanced Information Society, Proc. ITC-11, ed. M. Akiyama (North-Holland Publ. Cy., Amsterdam) pp. 452-458.

H.D. Friedman (1965). *Reduction methods for tandem queueing systems,* Oper. Res. 13, 121-131.

B.S. Greenberg, R.W. Wolff (1988). *Optimal order of servers for tandem queues in light traffic,* Management Sci. 34, 500-508.

P.G. Harrison (1981). *Approximate analysis and prediction of time-delay distributions in networks of queues,* Computer Performance 2, 124-135.

P.G. Harrison (1982). *Convergent discrete form for time-delay distributions in networks of queues,* Computer Performance 3, 167-175.

P.G. Harrison (1984). *A note on cycle times in tree-like queueing networks,* Adv. in Appl. Probab. 16, 216-219.

P.G. Harrison (1984a). *The distribution of cycle times in tree-like networks of queues,* The Computer Journal 27, 27-36.

P.G. Harrison (1985). *Paired centre analysis of time delays in queueing networks,* In: Modelling Techniques and Tools for Performance Analysis, ed. D. Potier (North-Holland Publ. Cy., Amsterdam) 571-588.

P.G. Harrison (1986). *An enhanced approximation by pair-wise analysis of servers for time delay distributions in queueing networks,* IEEE Trans. Computers 35, 54-61.

P.G. Harrison (1989). *Laplace transform inversion and passage time distributions in Markov processes,* to appear in J. Appl. Probab.

J.M. Harrison (1973). *The heavy traffic approximation for single server queues in series,* J. Appl. Probab. 10, 613-629.

J.M. Harrison (1978). *The diffusion approximation for tandem queues in heavy traffic,* Adv. in Appl. Probab. 10, 886-905.

J.M. Harrison, M.I. Reiman (1981). *On the distribution of multidimensional reflected Brownian motion,* SIAM J. Appl. Math. 41, 345-361.

J. Hemker (1987). *Durchlaufzeiten in geschlossenen Netzwerken von Warteschlangen,* (in German), M.Sc. Thesis, Department of Mathematical Stochastics, University of Hamburg.

J. Hemker (1989). *A note on sojourn times in queueing networks with multiserver nodes,* Report, Institute of Mathematical Stochastics, University of Hamburg, to appear in J. Appl. Probab.

S.D. Hohl, P.J. Kühn (1988). *Approximate analysis of flow and cycle times in queuing networks,* In: Proc. 3rd Int. Conference on Data Communication Systems and their Performance, eds. L.F.M. de Moraes, E. de Souza e Silva and L.F.G. Soares (North-Holland Publ. Cy., Amsterdam) pp. 471-485.

D.L. Iglehart, G.S. Shedler (1980). *Regenerative Simulation of Response Times in Networks of Queues,* Lect. Notes in Control and Inf. Sci. 26 (Springer, Berlin).

I. Iliadis, L. Y.-C. Lien (1988). *Resequencing delay for a queueing system with two heterogeneous servers under a threshold-type scheduling,* IEEE Trans. Commun. 36, 692-702.

U. Jansen, D. König (1980). *Insensitivity and steady state probabilities in product form for queueing networks,* Elektron. Informationsverarb. Kybernet. 16, 385-397.

U. Jansen (1984). *Conditional expected sojourn times in insensitive queueing systems and networks,* Adv. in Appl. Probab. 16, 906-919.

T. Kawashima, N. Torigoe (1983). *The cycle time distribution in a central server queueing system with multi-server station,* Memoirs of the National Defense Academy 23, 155-160.

T. Kawashima (1987). *On the sojourn time distribution in cyclic queueing systems with a LiPS station,* J. Oper. Res. Soc. Japan 30, 335-346.

J. Keilson (1979). *Markov Chain Models - Rarity and Exponentiality* (Springer, Berlin).

F.P. Kelly (1979). *Reversibility and Stochastic Networks* (Wiley, New York).

F.P. Kelly, P.K. Pollett (1983). *Sojourn times in closed queueing networks,* Adv. in Appl. Probab.

15, 638-656.

F.P. Kelly (1984). *The dependence of sojourn times in closed queueing networks,* In: Mathematical Computer Performance and Reliability, eds. G. Iazeolla, P.J. Courtois and A. Hordijk (North-Holland Publ. Cy., Amsterdam) pp. 111-121.

F.P. Kelly (1989). *On a class of approximations for closed queueing networks,* Queueing Systems 4, 69-76.

P.C. Kiessler, R.L. Disney (1982). *The sojourn time in a three node, acyclic, Jackson queueing network,* Technical Report VTR 8203, Virginia Polytechnic Institute and State University, Blacksburg (VA).

P.C. Kiessler, B. Melamed, M. Yadin and R.D. Foley (1988). *Analysis of a three node queueing network,* Queueing Systems 3, 53-72.

L. Kleinrock (1964). *Communication Nets - Stochastic Message Flow and Delay* (McGraw-Hill, New York; reprinted by Dover, New York, 1972).

E. Koenigsberg (1982). *Twenty five years of cyclic queues and closed queue networks: A review,* J. Opnl. Res. Soc. 33, 605-619.

J. Kohlas (1986). *Numerical computation of mean passage times and absorption probabilities in Markov and semi-Markov models,* Zeitschr. für Oper. Res. 30, Ser. A, 197-207.

A.G. Konheim, M. Reiser (1977). *Delay analysis for tandem networks,* Proc. ICC '77, pp. 12.2-265 - 12.2-269.

D. König, M. Miyazawa (1988). *Relationships and decomposition in the delayed Bernoulli feedback queueing system,* J. Appl. Probab. 25, 169-183.

W. Krämer (1973). *Total waiting time distribution function and the fate of a customer in a system with two queues in series,* In: Proc. ITC-7, pp. 322/1-322/8.

C.P. Kruskal, M. Snir and A. Weiss (1988). *The distribution of waiting times in clocked multistage interconnection networks,* IEEE Trans. Computers 37, 1337-1352.

P.J. Kühn (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.

J. Labetoulle, G. Pujolle (1976). *A study of queueing networks with deterministic service and application to computer networks,* Acta Informatica 7, 183-195.

S.S. Lam, A.U. Shankar (1981). *A derivation of response time distributions for a multi-class feedback queueing system,* Performance Evaluation 1, 48-61.

S.S. Lavenberg, M. Reiser (1980). *Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers,* J. Appl. Probab. 17, 1048-1061.

S.S. Lavenberg (ed.) (1983). *Computer Performance Modeling Handbook,* Academic Press, New York.

A.J. Lemoine (1977). *Networks of queues - a survey of equilibrium analysis,* Management Sci. 24, 464-481.

A.J. Lemoine (1979). *Total sojourn time in networks of queues,* TR 79-020-1, Systems Control, Inc., Palo Alto (CA).

A.J. Lemoine (1986). *A stochastic network formulation for complex sequential processes,* Naval Res. Logist. Quart. 33, 431-443.

A.J. Lemoine (1987). *On sojourn time in Jackson networks of queues,* J. Appl. Probab. 24, 495-510.

N.W. MacFadyen, D.E. Everitt (1984). *Approximation techniques for tandem queues with related service-times,* In: Proc. ITC Seminar, Moscow, pp. 294-297.

A.V. Makarichev (1984). *Analysis of a tandem queueing system with identical service times at both counters for various service disciplines,* In: Proc. ITC Seminar, Moscow, pp. 298-301.

R.A. Marie (1979). *An approximate analytical method for general queueing networks,* IEEE Trans. Software Eng. 5, 530-538.

W.A. Massey (1987). *Calculating exit times for series Jackson networks,* J. Appl. Probab. 24, 226-234.

J. McKenna (1987). *Asymptotic expansions of the sojourn time distribution function of jobs in closed,* *product form queueing networks,* J. Assoc. Comput. Mach. 34, 985-1003.

J. McKenna (1988). *Extensions and applications of RECAL in the solution of closed product-form queueing networks,* Stochastic Models 4, 235-276.

B. Melamed (1979). *Characterisation of Poisson traffic streams in Jackson queueing networks,* Adv. in Appl. Probab. 11, 422-438.

B. Melamed (1982). *Sojourn times in queueing networks,* Math. Oper. Res. 7, 223-244.

B. Melamed (1983). *Randomization procedures for numerical computation of delay time distributions in queueing systems,* Bull. Int. Statist. Inst., ISI Session 44, Vol. II, 755-775.

B. Melamed, M. Yadin (1984). *Randomisation procedures in the computation of cumulative-time distributions over discrete state Markov processes,* Oper. Res. 32, 926-944.

B. Melamed, M. Yadin (1984a). *Numerical computation of sojourn-time distributions in queueing networks,* J. Assoc. Comput. Mach. 31, 839-854.

N. Mikou (1988). *A two-node Jackson's network subject to breakdowns,* Stochastic Models 4, 523-552.

C.R. Mitchell, A.S. Paulson, C.A. Beswick (1977). *The effect of correlated exponential service times on single server tandem queues,* Naval Res. Logist. Quart. 24, 95-112.

D. Mitra, J.A. Morrison (1985). *Heavy-usage asymptotic expansions for the waiting time in closed processor-sharing systems with multiple classes,* Adv. in Appl. Probab. 17, 163-185.

I. Mitrani (1979). *A critical note on a result by Lemoine,* Management Sci. 25, 1026-1027.

I. Mitrani (1985). *Response time problems in communication networks,* J. Roy. Statist. Soc. B 47, 396-406.

J.A. Morrison (1987). *Conditioned response-time distribution for a large closed processor-sharing system in very heavy usage,* SIAM J. Appl. Math. 47, 1117-1129.

M. Pinedo, R.W. Wolff (1982). *A comparison between tandem queues with dependent and independent service times,* Oper. Res. 30, 464-479.

K.M. Rege, B. Sengupta (1988). *Response time distribution in a multiprogrammed computer with terminal traffic,* Performance Evaluation 8, 41-50.

E. Reich (1957). *Waiting times when queues are in tandem,* Ann. Math. Statist. 28, 768-773.

E. Reich (1963). *Note on queues in tandem,* Ann. Math. Statist. 34, 338-341.

M.I. Reiman (1982). *The heavy traffic diffusion approximation for sojourn times in Jackson networks,* In: Applied Probability - Computer Science: The Interface, eds. R.L. Disney and T.J. Ott (Birkhäuser Verlag, Boston) Vol.II, pp. 409-421.

M.I. Reiman (1984). *Open queueing networks in heavy traffic,* Math. of Oper. Res. 9, 441-458.

M.I. Reiman (1988). *A multiclass feedback queue in heavy traffic,* Adv. in Appl. Probab. 20, 179-207.

M.I. Reiman, B. Simon (1988). *Light traffic limits of sojourn time distributions in Markovian queueing networks,* Stochastic Models 4, 191-233.

M.I. Reiman, B. Simon (1988a). *An interpolation approximation for queueing systems with Poisson input,* Oper. Res. 36, 454-469.

M.I. Reiman, B. Simon (1989). *Open queueing systems in light traffic,* Math. of Oper. Res. 14, 26-59.

M. Reiser (1979). *A queueing network analysis of computer communication networks with window flow control,* IEEE Trans. Commun. 27, 1199-1209.

M. Reiser (1981). *Calculation of response-time distributions in cyclic exponential queues,* Performance Evaluation 1, 331-333.

M. Reiser (1982). *Performance evaluation of data communication systems,* Proc. of the IEEE 70, 171-196.

Z. Rosberg, N. Shacham (1988). *Buffer occupancy and message delay due to resequencing under reliable transmission protocol,* In: Proc. 3rd Int. Conference on Data Communication Systems and their Performance, eds. L.F.M. de Moraes, E. de Souza e Silva and L.F.G. Soares (North-Holland Publ. Cy., Amsterdam) pp. 69-82.

I. Rubin (1974). *Communication networks: Message path delays,* IEEE Trans. Inform. Theory 20, 738-745.

I. Rubin (1975). *Message path delays in packet-switching communication networks,* IEEE Trans.

Commun. 23, 186-192.

I. RUBIN (1976). *An approximate time delay analysis for packet-switching communication networks,* IEEE Trans. Commun. 24, 210-222.

S. SALZA, S.S. LAVENBERG (1981). *Approximating response time distributions in closed queueing network models of computer performance,* In: Performance '81, ed. F.J. Kylstra (North-Holland Publ. Cy., Amsterdam) pp. 133-145.

R. SCHASSBERGER, H. DADUNA (1983). *The time for a roundtrip in a cycle of exponential queues,* J. Assoc. Comput. Mach. 30, 146-150.

R. SCHASSBERGER, H. DADUNA (1983a). *A discrete-time technique for solving closed queueing network models of computer systems,* In: Messung, Modellierung und Bewertung von Rechensystemen, eds. P.J. Kühn and K.M. Schulz (Springer, Berlin) pp. 122-134.

R. SCHASSBERGER (1985). *Exact results on response time distributions in networks of queues,* In: Messung, Modellierung und Bewertung von Rechensystemen, ed. H. Beilner (Springer, Berlin) pp. 115-126.

R. SCHASSBERGER (1986). *Two remarks on insensitive stochastic models,* Adv. in Appl. Probab. 18, 791-814.

R. SCHASSBERGER, H. DADUNA (1987). *Sojourn times in queuing networks with multiserver nodes,* J. Appl. Probab. 24, 511-521.

A. SEKINO (1972). *Response time distribution of multiprogrammed time-shared computer systems,* Proc. 6th Annual Princeton Conf. Information Sciences and Systems, pp. 613-619.

K.C. SEVCIK, I. MITRANI (1981). *The distribution of queueing network states at input and output instants,* J. Assoc. Comput. Mach. 28, 358-371.

M. SHALMON, M.A. KAPLAN (1984). *A tandem network of queues with deterministic service and intermediate arrivals,* Oper. Res. 32, 753-773.

M. SHALMON (1987). *Exact delay analysis of packet-switching concentrating networks,* IEEE Trans. Commun. 35, 1265-1271.

J.G. SHANTHIKUMAR, J.A. BUZACOTT (1984). *The time spent in a dynamic job shop,* European J. Oper. Res. 17, 215-226.

J.G. SHANTHIKUMAR, U. SUMITA (1988). *Approximations for the time spent in a dynamic job shop with application to due-date assignment,* Int. J. Prod. Research 26, 1329-1352.

A.W. SHUM, J.P. BUZEN (1977). *The EPF technique: A method for obtaining approximate solutions to closed queueing networks with general service times,* In: Proc. Third Symposium on Measuring, Modelling and Evaluating Computer Systems, eds. H. Beilner and E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 201-220.

B. SIMON, R.D. FOLEY (1979). *Some results on sojourn times in acyclic Jackson networks,* Management Sci. 25, 1027-1034.

B. SIMON (1984). *Priority queues with feedback,* J. Assoc. Comput. Mach. 31, 134-149.

B. SIMON, J.S. WILLIE (1986). *Estimation of response time characteristics in priority queueing networks via an interpolation methodology based on simulation and heavy traffic limits,* In: Computer Science and Statistics: Proc. of the 18th Symposium on the Interface, pp. 251-256.

E. DE SOUZA E SILVA, S.S. LAVENBERG (1989). *Calculating joint queue-length distributions in product-form queuing networks,* J. Assoc. Comput. Mach. 36, 194-207.

A. STAFYLOPATIS, E. GELENBE (1988). *Delay analysis of resequencing systems with partial ordering,* In: Performance '87, eds. G. Latouche and P.J. Courtois (North-Holland Publ. Cy., Amsterdam) pp. 433-445.

T. SUZUKI, T. KAWASHIMA (1974). *Reduction methods for tandem queueing systems,* J. Oper. Res. Soc. Japan 17, 133-144.

L. TAKÀCS (1963). *A single-server queue with feedback,* The Bell System Techn. J. 42, 505-519.

H. TAKAGI (1990). Chapter in this book.

S.V. TEMBE, R.W. WOLFF (1974). *The optimal order of service in tandem queues,* Oper. Res. 22, 824-832.

F.A. Tobagi (1982). *Distributions of packet delay and interdeparture time in slotted ALOHA and Carrier Sense Multiple Access,* J. Assoc. Comput. Mach. 29, 907-927.

P. Tsoucas, J. Walrand (1983). *A note on the processor sharing queue in a quasireversible network,* Adv. in Appl. Probab. 15, 468-469.

O.P. Vinogradov (1984). *On the distribution of sojourn time in the tandem system with identical service times,* In: Proc. ITC Seminar, Moscow, pp. 449-450.

O.P. Vinogradov (1986). *A multiphase system with identical service,* Sov. J. Comput. Syst. Sci. 24, 28-31.

J. Walrand, P. Varaiya (1980). *Sojourn times and the overtaking condition in Jacksonian networks,* Adv. in Appl. Probab. 12, 1000-1018.

J. Walrand (1988). *An Introduction to Queueing Networks (Ch. 4),* Prentice Hall, Englewood Cliffs (NJ).

R.R. Weber (1979). *The interchangeability of tandem ./M/1 queues in series,* J. Appl. Probab. 16, 690-695.

W. Whitt (1982). *Approximating a point process by a renewal process I: Two basic methods,* Oper. Res. 30, 125-147.

W. Whitt (1983). *The queueing network analyzer,* The Bell System Techn. J. 62, 2779-2815.

W. Whitt (1984). *Open and closed models for networks of queues,* AT&T Bell Labs. Techn. J. 63, 1911-1979.

W. Whitt (1984a). *The amount of overtaking in a network of queues,* Networks 14, 411-426.

W. Whitt (1985). *The best order for queues in series,* Management Sci. 31, 475-487.

R.W. Wolff (1982). *Poisson arrivals see time averages,* Oper. Res. 30, 223-231.

R.W. Wolff (1982a). *Tandem queues with dependent service times in light traffic,* Oper. Res. 30, 619-635.

J.W. Wong (1978). *Distribution of end-to-end delay in message-switched networks,* Computer Networks 2, 44-49.

J.W. Wong (1979). *Response time distribution of the M/M/m/N queueing model,* Oper. Res. 27, 1196-1202.

C.M. Woodside (1984). *Response time sensitivity measurement for computer systems and general closed queuing networks,* Performance Evaluation 4, 199-210.