Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Phone: ++49-711-685-68026, Fax: ++49-711-685-67983
Email: mail@ikr.uni-stuttgart.de, http://www.ikr.uni-stuttgart.de

Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Phone: ++49-711-685-68026, Fax: ++49-711-685-67983
Email: mail@ikr.uni-stuttgart.de, http://www.ikr.uni-stuttgart.de

Approximate Analysis of General Queuing Networks by Decomposition

PAUL J. KUEHN, MEMBER, IEEE

Abstract—In this paper an approximate method for the analysis of general queuing networks is proposed. The queuing network is of the open network type, having N single server queuing stations with arbitrary interconnections. Customers may enter the network at any queuing station. The interarrival times of the exogenous arrival processes and the service times at the queuing stations are generally distributed. The analysis is based on the method of decomposition where the total network is broken up into subsystems; e.g., queuing stations of the type $GI/G/1$ or subnets. The subsystems are analyzed individually by assuming renewal arrival and departure processes. All related processes are considered with respect to their first two moments only. An analysis procedure is reported which reduces the total problem to a number of elementary operations which can be performed efficiently with the aid of a computer. Numerical results are reported together with simulation results to demonstrate the accuracy of the new method. The paper concludes with a short discussion of possible extensions of the method.

1. INTRODUCTION

THE global traffic flow within computer and computer communications systems can be described by queuing networks. The analysis of complex queuing networks, however, results often in difficulties because of a too large number of system states or the lack of exact methods so that there is a need for accurate approximate methods.

Exact solutions are known by Jackson [1] and Gordon and Newell [2] for open and closed networks with exponential interarrival and service time distributions, respectively. These solutions have a closed product form for the stationary multidimensional state probabilities where the single product terms are the solutions of isolated exponential queuing stations. These basic solutions were extended by Baskett *et al.* [3] to open, closed, and mixed networks with different classes of customers for exponential service times under FCFS (first-come, first-served) or phase-type service times under PS (processor-sharing) and preemptive-resume LCFS (last-come, first-served) strategies. Further generalizations and properties of this class of networks are reported in [4, 5]. It has also been shown that a parametric analysis can be performed in these cases by reducing the network to a suitable subsystem, cf. Chandy, *et al.* [6]. This principle was also extended to general queuing networks approximately [7].

Another class of solution techniques is that of *decomposition*, where the network is broken up into subsystems which are analyzed in isolation. This can be done either by considering the related input and output processes of subsystems or by separation of the total system into a hierarchy of "aggregate systems" with only few interactions between the various levels, cf. Disney and Cherry [8] or Courtois [9], respectively.

Among the approximation methods for queuing networks, the *diffusion approximation* has been intensively studied in recent years [10-13]. This technique is based on the assumption that the number of events in a given time interval is approximately normally distributed. Generally, this method yields good results in case of heavy traffic.

The solution technique used in this paper belongs to the decomposition method considering basically input and output processes of the related subsystems. In the second Chapter, the general queuing network model will be defined. The third Chapter describes the analysis method in detail. In the fourth Chapter, numerical results are shown to demonstrate the accuracy of the proposed method. The fifth Chapter summarizes the results, relates them to other known results, and discusses extensions with respect to multiserver stations, multi-class customers, subnets, fixed delay elements, cyclic queues, and synchronously operated switches.

2. QUEUING NETWORK MODEL

2.1 Network Structure

The queuing network consists of various elements as servers, queues, transition paths, feedback loops, decomposition points (splitting of processes), and composition points (superposition of processes). Fig. 1 shows an elementary queuing station (a) and a network example (b).

The elementary queuing station no. i consists of a single server, a single queue with unlimited capacity, an input terminal (composition point C_i), and an output terminal (decomposition point D_i). The general queuing network is built from N elementary queuing stations according to Fig. 1(a) by arbitrary interconnections. It is assumed that exogenous arriving customers enter the queuing network at arbitrary composition points and that customers may depart from the network at an arbitrary decomposition point having a path to the outside world. There is at least one exogenous arrival process and at least one station from which customers can leave the network (open network).

The following parameters define the network structure with respect to the network topology and the routing of customers:

N	Total number of queuing stations.
$Q = (q_{ij})$	Routing matrix,
q_{ij}	Routing probability for customers leaving station i and changing to station j, $i = 1, 2, \cdots, N, j = 0, 1, \cdots, N$.

Paper approved by the Editor for Computer Communication of the IEEE Communications Society for publication after presentation at the 8th International Teletraffic Congress (ITC), Melbourne, Australia, November 1976. Manuscript received January 6, 1978; revised July 14, 1978. This work was supported by the Federal Ministry of Research and Technology (BMFT) of the Federal Republic of Germany.

The author is with the Gesamthochschule Siegen, Siegen, Germany.

0090-6778/79/0100-0113$00.75

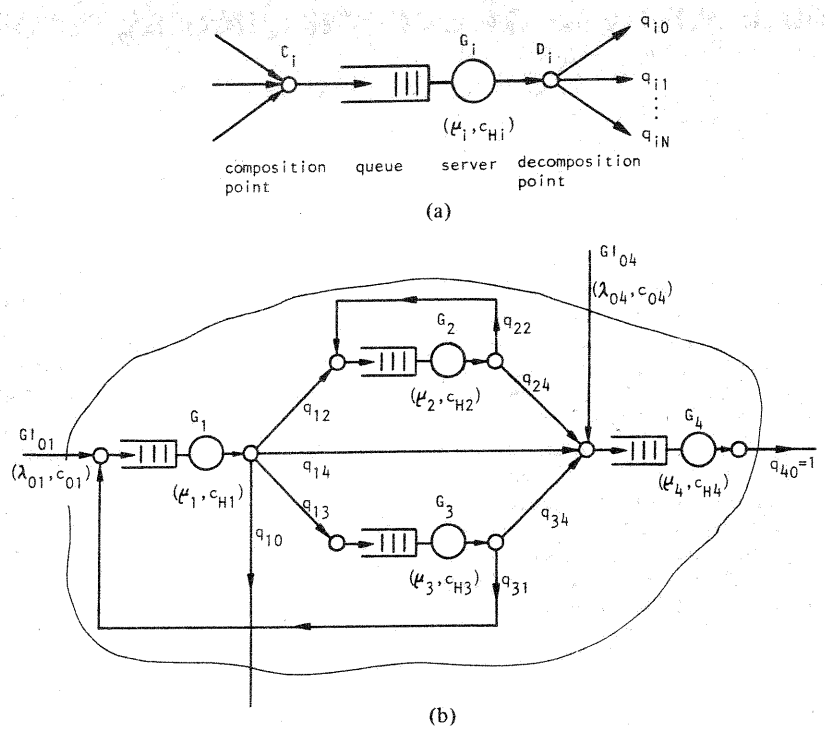

Fig. 1. Queuing network structure. (a) Elementary queuing station no. i. (b) Network model (example with 4 stations).

Herewith, "station 0" represents the outside of the queuing network.

2.2 Arrival and Service Processes

Customers arrive from the outside of the queuing network according to general exogenous arrival processes and they are served at the various stations according to general service processes, where:

$GI_0 = (GI_{0i})$	Vector of exogenous arrival processes.
$\lambda_0 = (\lambda_{0i})$	Vector of exogenous arrival rates, where $a_{0i} = 1/\lambda_{0i}$ is the mean exogenous interarrival time at station i.
$c_0 = (c_{0i})$	Vector of the coefficients of variation of the exogenous arrival processes.
$G = (G_i)$	Vector of service processes.
$\mu = (\mu_i)$	Vector of service rates, where $h_i = 1/\mu_i$ is the mean service time at station i.
$c_H = (c_{Hi})$	Vector of the coefficients of variation of the service processes.

At each station i the exogenous interarrival times T_{0i} and service times T_{Hi} are mutually independent and identically distributed with probability distribution function (df) $A_{0i}(t)$ and $H_i(t)$, respectively, $i = 1, 2, \cdots, N$. The latter assumption includes the independence assumption that successive service times of the same customer are independent of each other [14]. The following notations will be used for the df, the kth ordinary moment, and the coefficient of variation of a random variable T:

$$F(t) = P\{T \leqslant t\} \tag{1a}$$

$$E[T^k] = \int_{0-}^{\infty} t^k \, dF(t), \qquad k = 1, 2, \cdots \tag{1b}$$

$$c = \sqrt{\frac{E[T^2]}{E[T]^2} - 1}. \tag{1c}$$

These notations are used for interarrival times T_A, service times T_H, interdeparture times T_D, and interarrival times in arbitrary network paths analogously. For a short notation of the process type, the usual abbreviations are used as M and D for Markovian and deterministic processes, E_k and H_k for Erlangian and hyperexponential processes of order k, respectively. The network is considered to be in the stationary state.

2.3 Routing and Queuing Disciplines

All customers in the network are treated equally (one class of customers only). A customer leaving station i is routed to a station j independently according to the probability q_{ij}, $i = 1, 2, \cdots, N$, $j = 0, 1, \cdots, N$. Queuing customers are scheduled for service according to an arbitrary queue discipline which does not depend on the service time (e.g., FCFS, LCFS, RANDOM).

3. ANALYSIS BY DECOMPOSITION

3.1 Outline of the Basic Analysis Principles

The analysis method was developed according to the following principles.

1. Decomposition of the queuing network into subsystems; e.g., single queuing stations or subnetworks.
2. Analysis of the subsystems in isolation. The subsystems are related to their network surroundings by input (arrival) and output (departure) processes.
3. Approximation of all nonrenewal processes by stationary renewal processes.
4. Consideration of two moments (mean, coefficient of variation) of all processes consistently.

5. Reduction of the total analysis to few elementary operations to be performed efficiently by a computational algorithm.

The key points of the analysis method are principles 3 and 4. Stationary renewal processes are used because of their mathematical tractability for the necessary operations. Additionally, principle 3 is motivated by an analogy argument between Markovian queuing networks and networks with more general arrival and service processes: Markovian networks can be decomposed into subsystems exactly, where the arrival and departure processes of the subsystems can be assumed to be Markovian despite the fact that they are not (with the exception of networks without feedbacks [15]). In other words, for Markovian networks the global product solution [1-4] is not affected by the nonrecurrence of processes. This phenomenon is transferred to general networks approximately.

Principle 4 rests on a number of observations in queuing and teletraffic theory where characteristic mean values are mainly (sometimes only) influenced by the mean and variance of a random variable. Examples are the queuing station $M/G/1$ where the mean waiting time depends on the mean and variance of the service time only (Pollaczek-Khintchine), and overflow systems where the blocking probability basically depends on the first two moments of the offered overflow traffic. Although there is an indication by counterexamples [16] or particular computer applications [17] that a two moment approach is misleading or not accurate enough, the two moment principle has been chosen for reasons of tractability. For special applications, this principle has to be revised and augmented (see Chapter 5 and [18]).

We finally mention for the sake of completeness that similar methods of decomposition have been suggested [12] and [19]. Apart from different methods for elementary operations, the present approach differs in the consistent consideration of two moments in general networks with feedback loops.

Exact methods for the analysis of general queuing networks are only known in some special cases (e.g., 2 queues in a closed network with one exponential and one general server). Also, for smaller networks the exact solution can be carried out numerically by using the method of phases to represent general df's, cf. [20].

3.2 Elementary Standard Operations

In the following sections, basic operations are discussed which are elements of the analysis algorithm in Section 3.3.

3.2.1 Mean Arrival Rates: Under the assumption of stationarity, the mean arrival rate λ_i of queuing station i is obtained from the following set of linear equations representing the conservation of flow [21]:

$$\lambda_i = \lambda_{0i} + \sum_{j=1}^{N} \lambda_j q_{ji}, \qquad i = 1, 2, \cdots, N. \tag{2}$$

In the stationary case, for all stations it must hold

$$\rho_i = \lambda_i/\mu_i < 1, \qquad i = 1, 2, \cdots, N. \tag{3}$$

ρ_i is the server utilization of station i. The transition rate λ_{ij} of the path from station i to station j follows from

$$\lambda_{ij} = \lambda_i q_{ij}, \qquad \begin{matrix} i = 1, 2, \cdots, N \\ j = 0, 1, \cdots, N. \end{matrix} \tag{4}$$

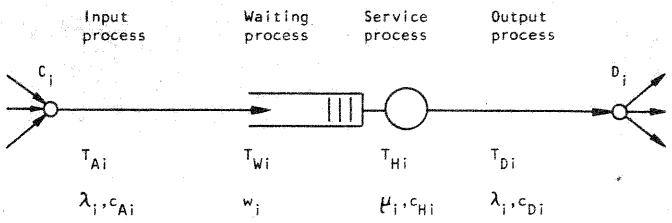

Fig. 2. The general queuing system $GI/G/1$.

3.2.2 Mean Values of the Queuing System GI/G/1: In the general network case we consider queuing stations of the type $GI/G/1$ given in Fig. 2.

The input (arrival) process is a renewal process with general df (GI), mean arrival rate λ_i, and coefficient of variation c_{Ai}. The service process is also general (G) with mean service or holding time $h_i = 1/\mu_i$ and coefficient of variation c_{Hi}. We are interested in the mean values of the waiting time T_{Wi} and flow time $T_{Fi} = T_{Wi} + T_{Hi}$ of an arbitrary customer and the number of customers X_i at the station, viz.,

$$w_i = E[T_{Wi}] \tag{5a}$$

$$f_i = E[T_{Fi}] = w_i + h_i \tag{5b}$$

$$N_i = E[X_i] = \lambda_i f_i = \Omega_i + \rho_i. \tag{5c}$$

In eq. (5c), Ω_i defines the mean queue length at station i. Exact results for this queuing system are only known in some special cases as $M/G/1$, $GI/M/1$, or can be obtained numerically by using a phase-type representation for general df's [34]. The analysis algorithm allows *any* $GI/G/1$ results as far as they rest on two moments of the input process and render the mean waiting time.

In the following, we apply a new approximation formula by Kraemer and Langenbach-Belz for $GI/G/1$ queuing stations which rests only on the first two moments of the arrival and service processes [22]. This formula is essentially an interpolation between known results for special systems and the general heavy traffic result and has been validated by extensive simulations for a wide range of arrival and service process combinations. The mean waiting time of this approximation is

$$w_i = h_i \cdot \frac{\rho i}{2(1-\rho_i)} \cdot (c_{Ai}^2 + c_{Hi}^2) \cdot g(\rho_i, c_{Ai}^2, c_{Hi}^2) \tag{6}$$

where

$$g(\rho_i, c_{Ai}^2, c_{Hi}^2) = \begin{cases} \exp\left\{-\dfrac{2(1-\rho_i)}{3\rho_i} \cdot \dfrac{(1-c_{Ai}^2)^2}{c_{Ai}^2 + c_{Hi}^2}\right\}, \\ c_{Ai} < 1 \\ \exp\left\{-(1-\rho_i) \cdot \dfrac{c_{Ai}^2 - 1}{c_{Ai}^2 + 4c_{Hi}^2}\right\}, \\ c_{Ai} \geqslant 1. \end{cases}$$

A similar expression has also been derived for the probability of delay W_i, cf. [22].

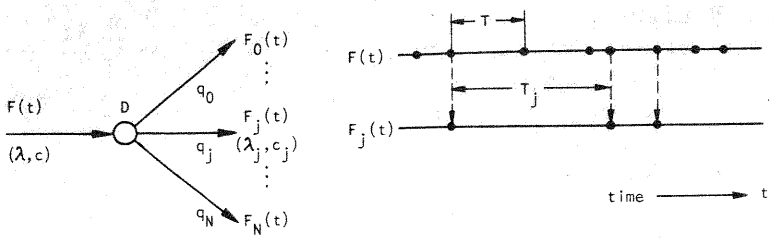

Fig. 3. Decomposition of a renewal process into $(N+1)$ component processes.

3.2.3 Output Process of the Queuing System GI/G/1: Output processes are generally difficult to describe since in almost all cases those processes are no longer recurrent. Explicit results are known for queuing stations of the type *M/M/n* by Burke [23], *M/D/1* by Pack [24], and *GI/M/n* with Interrupted Poisson Input by Heffes [25]. More general results are described in [26].

Under the approximation assumption 3 in Section 3.1, the output process is characterized by the df of the interdeparture times T_{Di}, viz.,

$$D_i(t) = P\{T_{Di} \leqslant t\}. \tag{7}$$

Under the further assumption 4 we concentrate only on the first and second moments of T_{Di}. The first moment is the reciprocal of the arrival rate λ_j. For the second moment, an exact relationship is known between the mean delay w_i and the coefficient of variation of T_{Di} for *GI/G/1* queuing stations, cf. Marshall [27]:

$$c_{Di}^2 = c_{Ai}^2 + 2\rho_i^2 c_{Hi}^2 - 2\rho_i(1-\rho_i)\cdot\frac{w_i}{h_i}. \tag{8}$$

Substituting eq. (6) into eq. (8) we obtain

$$c_{Di}^2 = c_{Ai}^2 + 2\rho_i^2 c_{Hi}^2 - \rho_i^2(c_{Ai}^2 + c_{Hi}^2) \cdot g(\rho_i, c_{Ai}^2, c_{Hi}^2) \tag{9a}$$

or with the simplification $g(\rho_i, c_{Ai}^2, c_{Hi}^2) \cong 1$

$$c_{Di}^2 \cong c_{Ai}^2 + \rho_i^2(c_{Hi}^2 - c_{Ai}^2). \tag{9b}$$

The solutions eq. (9a, b) include the known exact results for *M/G/1*, cf. Makino [28], as well as for *GI/G/1* for $\rho_i \to 0$ and $\rho_i \to 1$, respectively. It was shown by a number of simulations that eq. (9a) fits extremely well with respect to a wide range of arrival and service processes, cf. Chapter 4. Even the simpler result eq. (9b) is sufficiently accurate for a first characterization.

3.2.4 Decomposition of Renewal Processes: Given a stationary renewal point process as a sequence of events (arrivals of customers). The time between two successive events is a random variable T with df $F(t)$. At the decomposition point D, an arriving customer is routed into direction j according to a fixed and independent probability q_j, $j = 0, 1, \cdots, N$, cf. Fig. 3. We want to know the characteristics of the component processes; i.e., the df $F_j(t)$ of the interarrival times T_j, $j = 0, 1, \cdots, N$.

The random interarrival times T_j of the component process are constituted as sums of a random number X of successive realizations of the identically distributed random interarrival time T of the original process. Thus, the component process is found from the compound distribution through standard techniques. The Laplace-Stieltjes (LS) transform of $F_j(t)$ of the component process j is [21]

$$\phi_j(s) = \frac{q_j\phi(s)}{1-(1-q_j)\phi(s)} \tag{10}$$

where $\phi(s)$ is the LS transform of $F(t)$. From eq. (10) the transition rate λ_j and the coefficient of variation c_j of the component process j are derived

$$\lambda_j = E[T_j]^{-1} = \lambda q_j \tag{11a}$$

$$c_j^2 = q_j c^2 + (1-q_j), \qquad j = 0, 1, \cdots, N. \tag{11b}$$

Whereas $\lambda_0 + \lambda_1 + \cdots + \lambda_N = \lambda$ reflects the law of the conservation of flow in node D, an interesting relation is found from eq. (11b)

$$c_0^2 + c_1^2 + \cdots + c_N^2 = c^2 + N. \tag{11c}$$

The results hold exactly only in case of a recurrent process $F(t)$. For nonrecurrent (output) processes, these relations were also proved to be in good accordance with simulations, cf. Chapter 4. If $F(t)$ is a Markovian process, all component processes $F_j(t)$ are Markovian processes again.

3.2.5 Composition of Renewal Processes: The dual problem with the decomposition of a process is the composition (superposition) of a number of independent processes, given $(N+1)$ component processes which are stationary renewal point processes with df $F_j(t)$, $j = 0, 1, \cdots, N$. We are now interested in the characteristic of the resulting process when all component processes are superposed. For the sake of clearness, we will solve the more basic problem of two component processes $F_1(t)$ and $F_2(t)$ at first, cf. Fig. 4.

The resulting process is in the general case nonrecurrent. As shown in the Appendix by means of the forward recurrence times T_{V1}, T_{V2}, and T_V, the df $F(t)$ of the resulting process is

$$F(t) = 1 - \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\cdot\left\{F_1^c(t)\cdot\int_t^\infty F_2^c(u)\,du + F_2^c(t)\cdot\int_t^\infty F_1^c(u)\,du\right\} \tag{12}$$

where $F_j^c(t) = 1 - F_j(t)$ the complementary df of $F_j(t)$, $j = 1, 2$. Unfortunately, the moments of $F(t)$ cannot be given

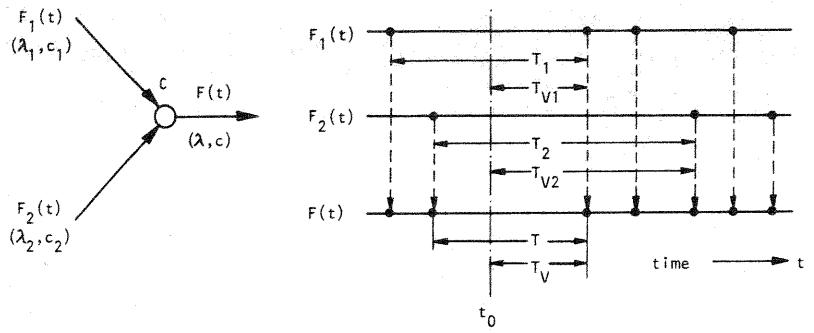

Fig. 4. Composition of two renewal component processes.

explicitly as a function of the moments of $F_1(t)$ and $F_2(t)$ except for the first moment; in this case, we obtain the plausible result

$$\lambda = E[T]^{-1} = \lambda_1 + \lambda_2 \tag{13}$$

which reflects again the conservation of flow in node C. If the component processes $F_1(t)$ and $F_2(t)$ were known explicitly, the higher moments could be calculated from eq. (12) straightforwardly. In the context of the two-moment network analysis approach, however, the component processes are only known by their first and second moments. To overcome this difficulty, we proceed as follows.

a) Construction of "substitute" component processes which agree with the real ones in their first and second moments.

b) Calculation of the second moment from eq. (12) using the concept a) of substitute component processes.

As substitute processes we choose a simple combination of phases allowing the approximation of *any* process with respect to their given first and second moments. Two cases of these phase representations are shown in Fig. 5(a), (b). In case of hypoexponential process types ($0 \leqslant c_j \leqslant 1$), we choose a series of a deterministic (D) and an exponential (M) phase, and in the case of hyperexponential process types ($c_j \geqslant 1$), an alternative of two exponentials (H_2). Mathematically, these processes are described by

$$F_j(t) = \begin{cases} \begin{cases} 0, & 0 \leqslant t \leqslant t_{j1} \\ 1 - \exp\{-\epsilon_{j2}(t - t_{j1})\}, & t \geqslant t_{j1}, \end{cases} \\ \qquad 0 \leqslant c_j \leqslant 1 \qquad (14a) \\ 1 - p_{j1}\exp(-\epsilon_{j1}t) - p_{j2}\exp(-\epsilon_{j2}t), \\ \qquad c_j \geqslant 1. \qquad (14b) \end{cases}$$

Using the abbreviation $E[T_{j\nu}] = t_{j\nu} = 1/\epsilon_{j\nu}$, $\nu = 1, 2$, the parameters of the substitute processes are determined from the given parameters as follows:

$0 \leqslant c_j \leqslant 1$:

$$\epsilon_{j1} = \lambda_j/(1 - c_j), \qquad \epsilon_{j2} = \lambda_j/c_j \tag{15a}$$

$c_j \geqslant 1$:

$$\epsilon_{j1,2} = \lambda_j \left\{ 1 \pm \sqrt{\frac{c_j^2 - 1}{c_j^2 + 1}} \right\}, \qquad p_{j1,2} = \epsilon_{j1,2}/2\lambda_j,$$
$$(p_{j1}t_{j1} = p_{j2}t_{j2}). \tag{15b}$$

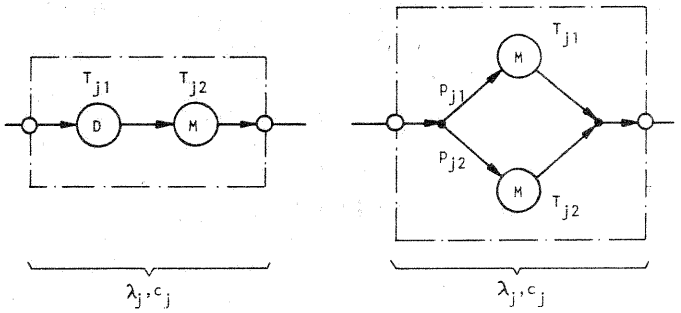

Fig. 5. Representation of substitute processes by phases. (a) Hypoexponential df ($0 \leqslant c_j \leqslant 1$). (b) Hyperexponential df ($c_j \geqslant 1$).

The superposition according to eq. (12) is carried out with these two basic substitute processes yielding the coefficient of variation c explicitly (see the Appendix). We remark that the results are slightly dependent on the choice of the substitute process types. If there is evidence of a much different process characteristic, appropriate other substitutes can be chosen, too. Finally, we note that if the component processes are Markovian, the resulting process is Markovian again.

The extension from two component processes to the general case of $(N + 1)$ processes is performed recursively in N steps according to Fig. 6.

3.2.6 Reconfiguration by Substitution of Nodal Feedbacks: During the development of this method, it became clear that strong nodal feedbacks are critical with respect to the assumption of renewal processes, since input and output processes of such a queuing system are correlated strongly. To eliminate this effect, a substitute queuing system without feedback is formed according to Fig. 7. In the original system, a customer is served according to a geometrically distributed number of service phases which may be interleaved by phases of other customers. In the substitute system, a customer gets its total service time continuously. As shown by a similar derivation as in the case of decomposed processes, the substitute service time df $H_i^*(t)$ is given by its LS transform

$$\Psi_i^*(s) = \frac{(1 - q_{ii})\Psi_i(s)}{1 - q_{ii}\Psi_i(s)} \tag{16}$$

where $\psi_i(s)$ is the LS transform of $H_i(t)$. From eq. (16) and Fig. 7 it follows:

$$\mu_i^* = \mu_i(1 - q_{ii}) \tag{17a}$$

$$c_{Hi}^{*2} = q_{ii} + (1 - q_{ii})c_{Hi}^2 \tag{17b}$$

$$q_{ij}^* = q_{ij}/(1 - q_{ii}), \qquad j \neq i. \tag{17c}$$

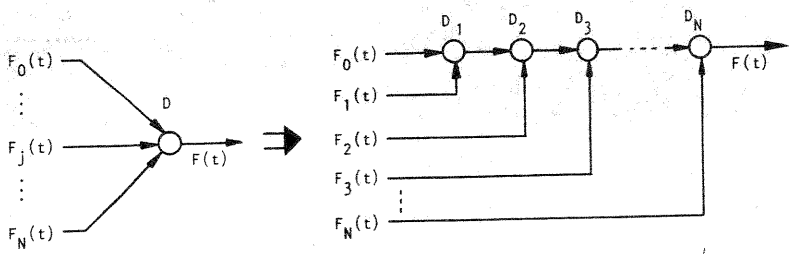

Fig. 6. Composition of $(N+1)$ component processes by N recursive compositions of two processes.

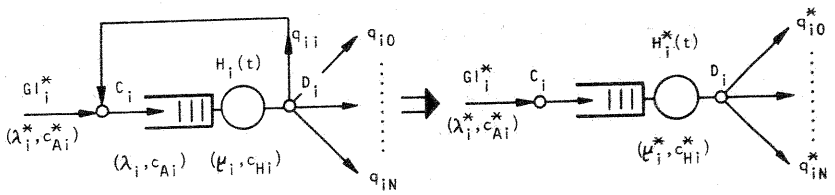

Fig. 7. Substitution of nodal feedbacks.

By this procedure, the nodal feedback loop is eliminated and the queuing system i is considered only with respect to customers arriving from or departing to other stations via the input and output ports C_i and D_i, respectively. Applying this procedure to each station with feedback, the queuing network becomes reconfigured. The reconfigured network differs from the original one with respect to arrival rates, service time distributions, and the routing matrix. After that reconfiguration procedure, the usual quantities without asterisk will be used to describe the reconfigured network.

A proof of an exact analogy between stations with and without nodal feedback with respect to the distribution of queue lengths and mean flow times was given by Takács [30] in the case of $M/G/1$ stations. The extension to general arrival processes is an approximation. It has been shown by simulations that the reconfiguration step yields good accuracy, whereas the analysis without that step results in considerable inaccuracies.

3.2.7 Mean Number of Visits at a Station and Mean Flow Times: To calculate the mean flow times, the expected number of visits at a certain station must be known. These are:

e_i Expected number of visits at station i with respect to an arbitrary customer.

$e_i(a)$ Expected number of visits at station i with respect to those customers entering the network at station a

where $i, a = 1, 2, \cdots, N$. Defining λ_i and $\lambda_i(a)$ as total or partial arrival rates at station i with respect to all customers or all customers entering the network at station a, respectively, we have

$$e_i = \lambda_i/\lambda, \quad \text{where} \quad \lambda = \sum_{a=1}^{N} \lambda_{0a} \tag{18a}$$

$$e_i(a) = \lambda_i(a)/\lambda_{0a}. \tag{18b}$$

It remains to calculate the various arrival rates in eq. (18a-b). The arrival rates λ_i are the solutions of eq. (2). The same procedure according to eq. (2) can be used to calculate the rates $\lambda_i(a)$ by setting $\lambda_{0j} = 0$ for $j \neq a$, cf. Fig. 8(b).

The flow times T_F and $T_F(a)$ define the random lifetime o an arbitrary customer or a customer of input a, respectivel The mean values are found directly by considering a "test cu tomer" moving through the network

$$f(a) = E[T_F(a)] = \sum_{i=1}^{N} e_i(a) f_i \tag{19a}$$

$$f = E[T_F] = \sum_{a=1}^{N} \frac{\lambda_{0a}}{\lambda} \cdot f(a) = \sum_{i=1}^{N} e_i f_i = \frac{1}{\lambda} \cdot \sum_{i=1}^{N} E[X_i]. \tag{19b}$$

The last expression is identical to Little's theorem applied the total network.

3.3 Queuing Network Analysis Algorithm

3.3.1 Standard Procedures: The algorithm is based on number of procedures for standard operations as discussed Section 3.2. These are:

MEANRATE Calculation of the mean arrival rates λ_i of queuing stations, $i = 1, 2, \cdots, N$.

RECONF Reconfiguration of the queuing network substitution of stations with feedba through stations without feedback and tra formation of the network parameters.

COMPOS Composition of $(N+1)$ component proc ses (λ_{ji}, c_{ji}) at C_i of station $i, j = 0, 1, \cdots, N$

DECOMP Decomposition of the departure process (c_{Di}) of queuing station i into $(N+1)$ co ponent processes (λ_{ij}, c_{ij}) at D_i of statior $j = 0, 1, \cdots, N$.

CDEPART Calculation of the coefficient of variati c_{Di} of the departure process of station i.

QVALUE Calculation of the characteristic values N_i, w_i, and f_i of queuing station i.

FLOWTIME Calculation of expectations of network fl times with respect to all customers or all c tomers of a certain input, respectively.

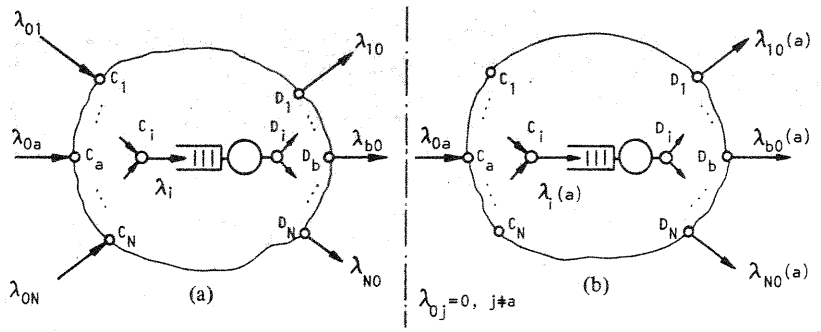

Fig. 8. Total and partial arrival rates at station i. (a) Network with complete exogenous arrivals. (b) Network with exogenous arrivals only at input a.

3.3.2 Analysis Algorithm: In networks without feedback, the analysis can be carried out in a straightforward way. There exists always a sequence for analyzing the network station by station starting from one of those stations having only exogenous arrivals.

In the general case of networks with feedback, the composition operation at a station i cannot always be carried out since there are not all component processes known with respect to their coefficients of variation c_{ji}, $i, j = 1, 2, \cdots, N$. This problem is solved by *iteration,* with the additional advantage of being applicable without regard to the sequence of stations to be analyzed.

The principal flow chart of the algorithm is shown in Fig. 9 (details are omitted). The algorithm is very fast and needs about $5N^2$ words of storage capacity for data. It has been implemented by an Algol computer program [31], and its results were checked by simulation programs for general queuing networks [32]. Finally, we state that the algorithm yields the exact results in the special case of pure Markovian networks with state-independent arrival and service rates.

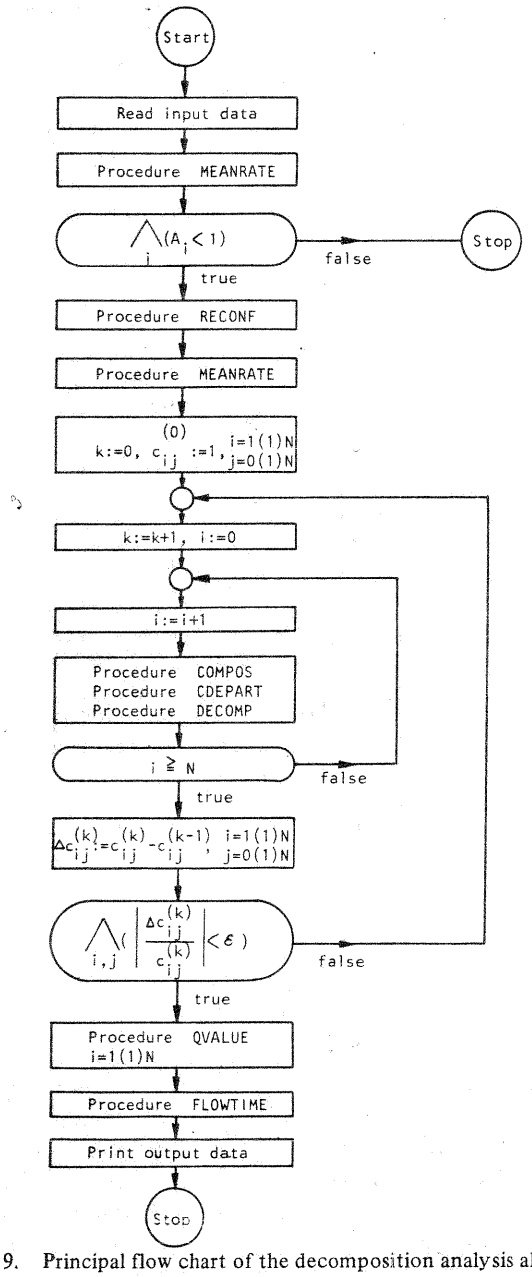

Fig. 9. Principal flow chart of the decomposition analysis algorithm.

4. NUMERICAL RESULTS AND VALIDATIONS

In this chapter, several results are reported to show the accuracy of the algorithm for basic operations and whole networks. All simulations were performed with at least 100 000 events. The results are given together with their 95% confidence levels.

4.1 Standard Operations

4.1.1 Mean Values and Output Process of the Queuing System GI/G/1: The mean values w_i and W_i cited in Section 3.2.2 have been checked by intensive simulations yielding an acceptable accuracy, cf. [22], and will not be repeated here.

Concerning the output process, in Fig. 10(a), (b) the functions $c_D{}^2(\rho)$ are given for queuing systems of the type $E_2/G/1$ and $H_2/G/1$ according to eq. (9a). The simulation results show a reasonable accuracy of the approximate solution.

4.1.2 Decomposition of Point Processes: In Fig. 11 the functions $c_j{}^2(c^2)$ are given for the decomposition of a renewal process or a nonrenewal process with coefficient of variation c and routing probability q_j as parameter according to Section 3.2.4, respectively. The results of the decomposition operation are shown in Fig. 11(a) for renewal processes and in Fig. 11(b) for nonrenewal processes. The nonrenewal processes were generated as output processes of queuing systems of the type $M/G/1$ for $\rho = 0.6$ (note that the output process is renewal again for $\rho \to 0$ and $\rho \to 1$). The curves hold exactly only if the original process is renewal, cf. Fig. 11(a). For nonrenewal processes the calculated curves are still within the confidence levels of the simulation, cf. Fig. 11(b). The accuracy does not depend remarkably on the parameter q_j.

4.1.3 Composition of Point Processes: The results of the

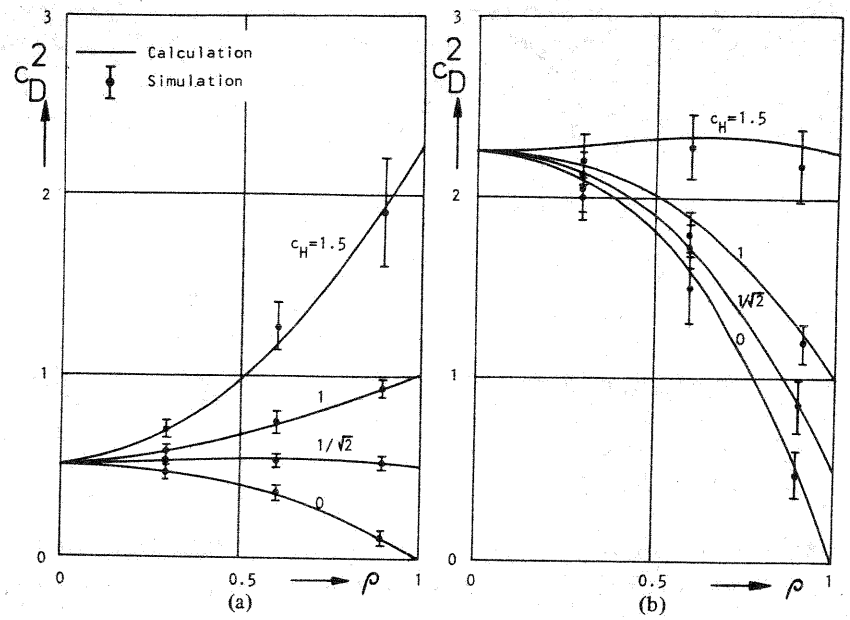

Fig. 10. Squared coefficient of variation $c_D{}^2$ versus the server utilization ρ (parameter c_H). (a) Queuing stations $E_2/G/1$. (b) Queuing stations $H_2/G/1(c_A = 1.5)$.

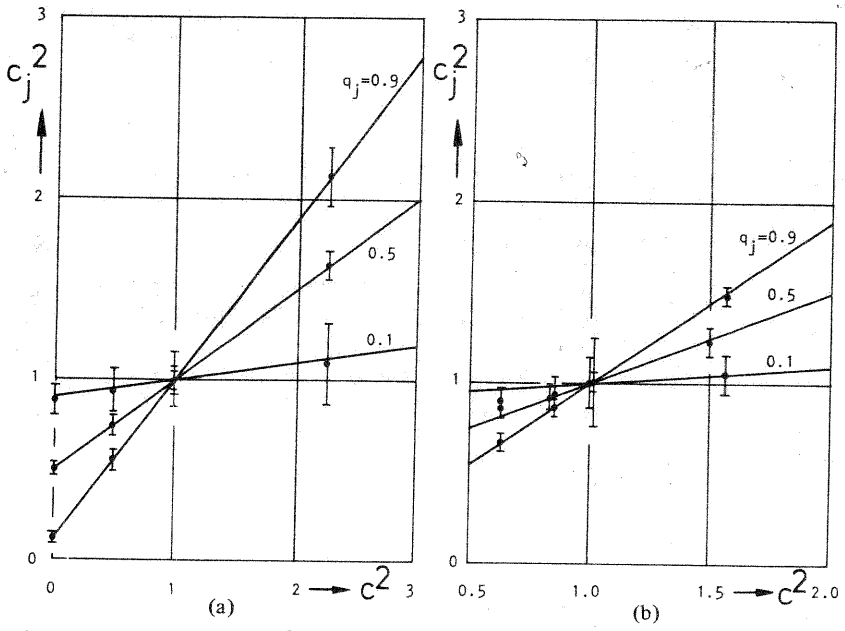

Fig. 11. Squared coefficient of variation $c_j{}^2$ of decomposed processes versus squared coefficient of variation c^2 of the original process (parameter q_j). (a) Decomposition of renewal processes. (b) Decomposition of nonrenewal processes.

composition operation on two component processes according to Section 3.2.5 are shown in Fig. 12. The figures represent the squared coefficient of variation c^2 of the superposed process dependent on the squared coefficients of variation $c_1{}^2$ (abscissa) and $c_2{}^2$ (parameter) for two renewal component processes (cf. Fig. 12(a)) or nonrenewal component processes (cf. Fig. 12(b)), respectively. The nonrenewal processes were realized as output processes of $M/G/1$ at $\rho = 0.6$. Whereas the composition of renewal processes fits extremely well with simulation, the composition of nonrenewal processes yields errors up to 15% in the worst case of the superposition of two output processes of two $M/D/1$ stations.

4.1.4 Substitution of Nodal Feedbacks: In Fig. 13 the calculated results of the reconfigured queuing station without nodal feedback according to Section 3.2.6 are compared with simulations for the original system with nodal feedback (the subscript i is omitted). The figures show the squared coefficient of variation $c_D{}^{*2}$ of the output process of customers leaving the station (cf. Fig. 13(a)) as well as the mean total flow time f of all outside arriving customers (cf. Fig. 13(b)) versus the squared coefficient of variation $c_H{}^2$ of service times of the original system with parameter $c_A{}^*$ of the arrival process of outside arriving customers. Both results for $c_D{}^{*2}$ and f are in acceptable accordance with simulations (note that the calculated results for $c_A{}^* = 1$ are exact). The comparatively large confidence levels for f reflect the high degree of variability caused by the nodal feedback of the simulated original system. Similar results were obtained for the comparison of

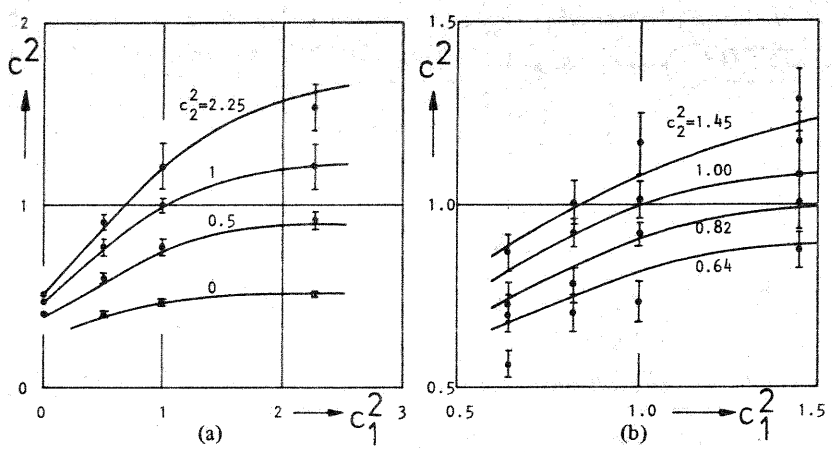

Fig. 12. Squared coefficient of variation c^2 of the superposed process dependent on squared coefficients of variation c_j^2 of the component processes ($\lambda_1 = \lambda_2$). (a) Composition of two renewal processes. (b) Composition of two nonrenewal processes.

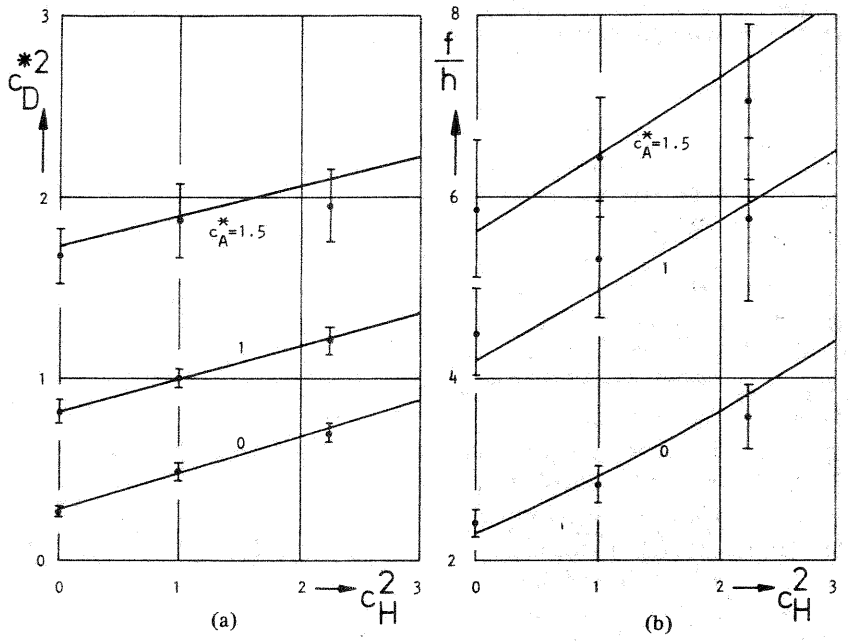

Fig. 13. Comparison of equivalent queuing stations with and without nodal feedback. Parameters: $\lambda^* = 0.3$, $h = 1.0$, $q = 0.5$ ($\rho = 0.6$), $c_A{}^* = 0$, 1, 1.5. (a) Squared coefficient of variation $c_D{}^{*2}$ of the output process versus $c_H{}^2$. (b) Mean normalized flow time f/h versus $c_H{}^2$.

both systems by simulations only, although the equivalence of both systems with respect to f (and therefore also for $c_D{}^*$) has been proved only in the case of Poisson arrivals [30].

4.2 Queuing Networks

In this section, calculated results of the suggested decomposition method for queuing networks are given and compared with simulation results. First, we summarize general validation experiences. Then, we give two specific network examples with detailed results. The method has also been compared with results of two diffusion approximation methods. Although the decomposition method yielded in those cases a slightly better overall accuracy than the diffusion approximations, the sample values are too few for a representative comparison with all existing diffusion approximation methods. This problem remains for some future studies.

4.2.1 General Validation Experiences: Subsequently, we summarize some general experiences made with the validation of the described decomposition approach. The method yields generally an increasing accuracy under conditions of

a) low or heavy traffic;
b) increasing randomness in the arrival and service processes;
c) increasing network complexity;
d) decreasing "closedness" of the global network.

Observation a) results from the fact that in both low and heavy traffic regions the renewal assumption is usually better fulfilled. Observation b) rests on the fact that the results are the more robust with respect to the renewal assumption the closer the network is to a Markovian queuing network: In Markovian networks the decomposition approach yields even the exact results, although the processes are not renewal (except in case of networks without feedbacks). Also, more general arrival processes have a worse effect on the accuracy than more general service processes. Observation c) results from a general result of renewal theory which states that the superposition of

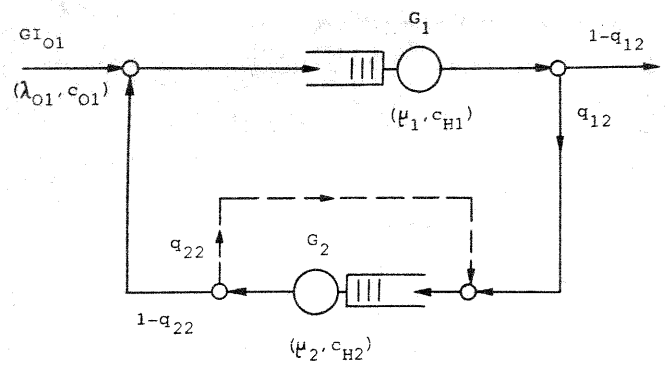

Fig. 14. Queuing network example with 2 queuing stations.

increasingly many component processes yields (in the limit) a Poisson process. This effect was observed with increasing network complexity (see also results of the decomposition and composition operations). The substitution of nodal feedback loops d) reduces the extent of the renewal assumption and results in a far better accuracy. Similarly, the closer the network to a closed queuing network, the greater the dependence effects. Examples of smaller closed queuing networks yielded errors up to 30%. With increased network extension and complexity, the "closedness" becomes again less important. For increasing numbers of input and output network terminals, the "randomness" is enforced, yielding in turn better accuracies of the decomposition approach.

4.2.2 Queuing Network Example with 2 Queuing Stations: As an example of a small queuing network, we consider a queuing network of two queuing stations according to Fig. 14 representing an interactive computer system with a CPU and a disk-*I/O* subsystem. Results of the decomposition technique and simulation results are given in Table 1 for the combinations of traffic and routing parameters shown in Table 2.

The comparison shows generally good accordance between calculated and simulated results in case of low and medium traffic. The inaccuracies in case of heavier traffic result in this case from the facts of a small and relatively "closed" network where the renewal assumption becomes critical, especially in case of more deterministic traffics. A further source of errors lies in the *GI/G/*1 approximation (usually less than 10%) and in the superposition of nonrenewal processes.

4.2.3 Queuing Network Example with 9 Queuing Stations: As a last example we consider a larger queuing network of 9 queuing stations shown in Fig. 15 together with the routing probabilities. The exogenous arrival processes are Markovian. The service processes are varied from deterministic ($c_H = 0$) to hyperexponential ($c_H = 2$).

The network example was investigated for homogeneous servers ($h_{1-9} = 1$, c_{H1-9} identical) as well as heterogeneous servers ($h_{1-9} = 1$, c_{H1-3} different to c_{H4-9}). Fig. 16 shows the mean normalized total flow time f/h and the mean flow time f_4/h of the interior station number 4 (solid curves) versus the coefficient of variation c_{H1-9} in case of homogeneous servers (cf. Fig. 16(a)) or versus c_{H4-9} with parameter c_{H1-3} in case of heterogeneous servers (cf. Fig. 16(b)), respectively.

In Fig. 16(a) two more curves are shown for comparison with different analysis methods when the arrival process at each station is assumed to be Markovian (dashed curves), or when the arrival and service processes at each station are assumed to be Markovian (dotted curves), respectively.

TABLE 1
COMPARISON OF CALCULATED RESULTS AND SIMULATION RESULTS FOR THE AVERAGE FLOW TIME OF CUSTOMERS IN A QUEUING NETWORK ACCORDING TO FIG. 14.

System Number	λ_{01}	f_{CALC}	f_{SIM} ± Confid. Levels	$\frac{f_{CALC}-f_{SIM}}{f_{SIM}} \cdot 100\%$
1	0.15	4.509	4.24 ± 0.08	6.34%
	0.30	8.220	7.51 ± 0.29	9.45%
	0.45	33.474	27.27 ± 5.63	22.75%
2	0.15	5.937	5.95 ± 0.89	- 0.22%
	0.30	10.888	10.91 ± 0.64	- 0.20%
	0.45	46.309	49.49 ± 6.80	- 6.43%
3	0.15	6.108	6.09 ± 0.19	0.29%
	0.30	11.595	11.08 ± 0.63	4.65%
	0.45	51.265	61.72 ± 17.99	- 16.94%
4	0.15	4.691	4.50 ± 0.16	4.24%
	0.30	8.712	7.96 ± 0.55	9.45%
	0.45	34.942	29.91 ± 6.14	16.82%
5	0.15	3.627	3.66 ± 0.05	- 0.90%
	0.30	4.926	5.35 ± 0.16	- 7.92%
	0.45	12.801	18.29 ± 3.27	- 30.01%
6	0.15	3.490	3.43 ± 0.04	1.75%
	0.30	4.420	4.83 ± 0.07	- 8.49%
	0.45	9.791	13.59 ± 1.79	- 27.95%
7	0.15	4.603	4.73 ± 0.12	- 2.68%
	0.30	8.590	9.04 ± 0.56	- 4.98%
	0.45	35.876	46.83 ± 10.64	- 23.39%
8	0.15	4.194	3.67 ± 0.09	14.27%
	0.30	7.494	5.78 ± 0.16	29.65%
	0.45	29.567	17.46 ± 1.67	69.34%

TABLE 2

System	$GI_{01}/G_1/G_2$	c_{01}	μ_1	μ_2	c_{H1}	c_{H2}	q_{12}	q_{22}
1	$M/H_2/E_4$	1.0	1.0	1.0	1.5	0.5	0.5	0
2	$M/H_2/E_4$	1.0	1.0	0.5	1.5	0.5	0.5	0
3	$M/H_2/E_4$	1.0	1.0	1.0	1.5	0.5	0.5	0.5
4	$M/H_2/H_2$	1.0	1.0	1.0	1.5	1.5	0.5	0
5	$M/E_4/E_4$	1.0	1.0	1.0	0.5	0.5	0.5	0
6	M/D/D	1.0	1.0	1.0	0	0	0.5	0
7	$H_2/H_2/E_4$	1.5	1.0	1.0	1.5	0.5	0.5	0
8	$E_4/H_2/E_4$	0.5	1.0	1.0	1.5	0.5	0.5	0

Compared with the simulation results, the proposed two-moment method yields acceptable results; whereas, neglect of the second moment of the arrival processes or both of the arrival and service processes yields principally worse results. The comparatively small difference between the solid and dashed curves results from the fact that in a complex network with many compositions and decompositions of processes the component processes tend to become Markovian again. Then the renewal assumption is also better justified as in special cases as closed networks, or, e.g., series of queues with constant service times.

5. DISCUSSION AND GENERALIZATIONS OF THE METHOD

The suggested decomposition method allows the analysis of open queuing networks with general exogenous arrival processes and general service processes. The method does not allow state-dependent arrival or service rates. The method has been described for the case of open networks only. Although an extension to closed networks is possible, we do not suggest its

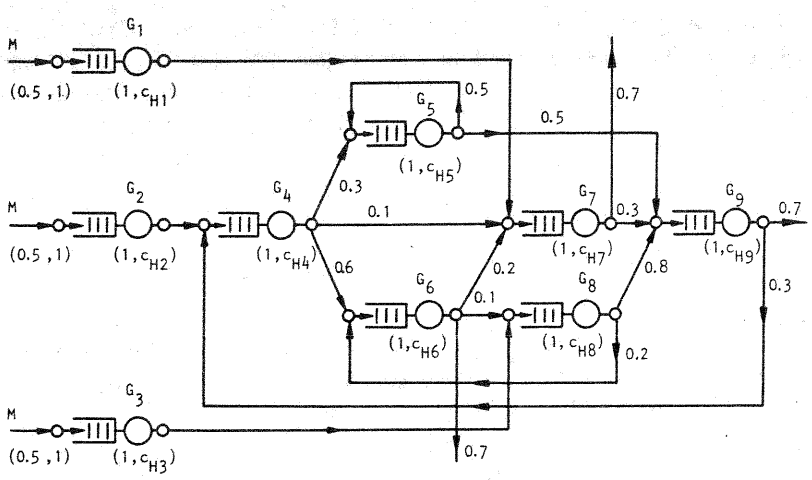

Fig. 15. Queuing network example with 9 queuing stations.

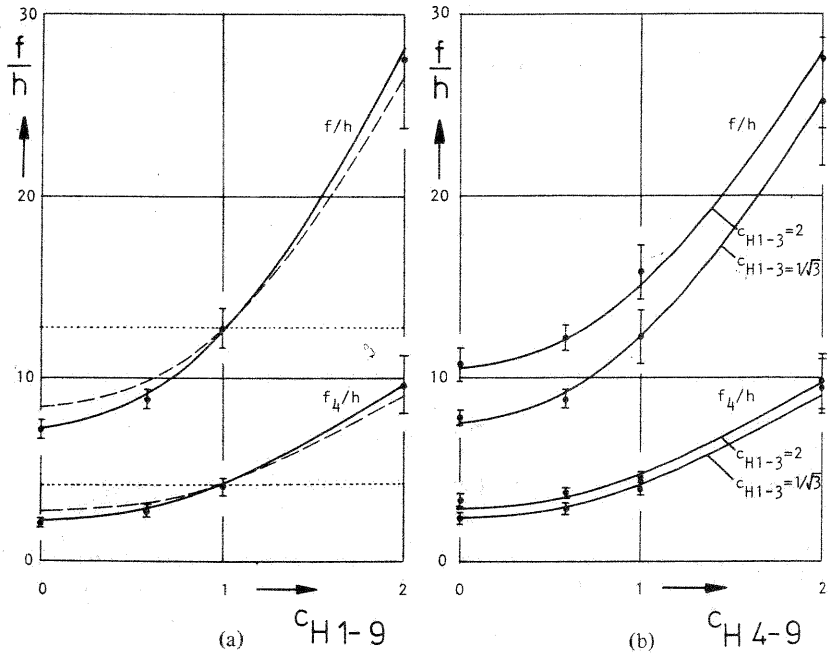

Fig. 16. Mean flow times f and f_4 versus the coefficient of variation c_H of network service stations. (a) Homogeneous servers. (b) Heterogeneous servers. —— Decomposition algorithm. – – – – Markovian arrival processes at queuing stations. - - - - Markovian queuing network result.

application except for cases of larger closed networks with a complex structure where the renewal process approximation is justified. The network analysis algorithm has been described for the case of single-queue, single-server stations with only one class of customers. The modular concept of the algorithm, however, easily allows a number of generalizations, as briefly discussed in the following sections.

5.1 Multi-Server Queuing Stations

Any single server queuing station no. i can be replaced by a multi-server station. The analysis is performed analogously by inserting the corresponding results for a $GI/G/n_i$ queue with respect to the coefficient of variation c_{Di} of the output process and mean values w_i, W_i, f_i, and N_i.

5.2 Multi-Queue Stations With Several Classes of Customers

A further extension is the introduction of R classes of customers. Customers are classified according to their origin, preceding path, urgency, or importance and may also change their class membership, cf. [3]. At each queuing station i, arriving customers are separated into R distinct queues according to their class index r, $r = 1, 2, \cdots, R$. Waiting customers are selected for service by a schedule with any nonpreemptive discipline; e.g., a nonpreemptive priority. A customer of class r leaving station i is routed to station j and changes into class s with probability $q_{ir,js}$, $r, s = 1, 2, \cdots, R$, $i = 1, 2, \cdots, N$, $j = 0, 1, \cdots, N$.

For the analysis, in a first step the arrival rate λ_{ir} of class r customers at station i must be determined from a system of $N \cdot R$ equations analogously to eq. (2). Herewith, the total arrival rate λ_i and the rates $\lambda_{ir,js}$ are also determined. Then, the general R-class queuing system $GI/G/1$ must be analyzed considering only two moments of processes and the underlying scheduling discipline. The consistency of the coefficients of variation of all processes in the network must be achieved again by iteration as described in Section 3.3.2. For the expected

number of visits at a station, a test customer of a certain class is considered moving through the network as described in Section 3.2.7 analogously.

5.3 Subnets

Up to now, the elementary subsystems of the network were single-stage service stations. In a further generalization, subsystems can also be subnets having one input port and one output port. The only difference to the described analysis algorithm is the analysis of the subnet itself, considering two moments of the input and output processes.

The concept of subnets is favorable in cases where the global subnet behavior with respect to input and output processes is sufficient to describe its influence on the residual network behavior, or in special cases when the decomposition method becomes worse if it is applied to the elements of the subnet. As an example of the first case, let the subnet be a queuing network model of a computer system and the residual queuing network the model of a large-scale computer communications network. As an example for the second case, consider subnets with strong dependencies between their stations as, e.g., in the case of closed loops or tandem queues with constant service times [33], where the decomposition method yields too bad results. For these reasons, it is useful to analyze such "aggregate systems" in isolation and to put them into the algorithm as a whole.

5.4 Further Generalizations

The described decomposition method has been further generalized in cases including fixed delay elements, cyclically served queues, and synchronously operated switches in connection with the investigation of network models for switching system control structures [18]. Fixed delay elements can be included since the process characteristic (2 moments) is completely maintained. Cyclically served queues with nonzero switch-over times are again described by their input/output behavior in terms of 2 moments of the interarrival times. For networks with synchronously operated switches the characterization of the interface traffic between subsystems has to be augmented; in this case the traffic is adequately characterized through the distribution (or its moments) of the batch size of customers transferred at the switching instants. The approximation consists in this case essentially of the assumption of independence between successively transferred batch sizes. For such networks with such mixed types of elements as cyclically served queues, fixed delay elements, synchronously operated switches with "batch service" and "batch arrival" effects, and priorities between classes of customers, a decomposition approach is the only successful analysis method at the current state of the art.

APPENDIX
COMPOSITION OF TWO RENEWAL PROCESSES

Given two stationary renewal point processes with random interarrival times T_1 and T_2 and df $F_1(t)$ and $F_2(t)$, respectively; both processes are superposed at the composition point C. The df of the resulting process is $F(t) = P\{T \leqslant t\}$, where T is the interarrival time of the superposed process, cf. Fig. 4.

Following the theory of renewal processes according to Cox and Miller [29], the forward recurrence times T_{V1}, T_{V2}, and T_V are introduced as the intervals between an arbitrary instant t_0 and the following event of the component and superposed processes, respectively. In the stationary case, the forward recurrence times T_{Vj} are independent of t_0 and their density function is given by

$$V_j'(t) = \lambda_j \cdot F_j^c(t), \qquad j = 1, 2 \tag{A.1}$$

where $F_j^c(t) = 1 - F_j(t)$ the complementary df of $F_j(t)$. From eq. (A. 1) the complementary df of T_{Vj} follows by integration

$$V_j^c(t) = \int_{u=t}^{\infty} \lambda_j F_j^c(u)\, du, \qquad j = 1, 2. \tag{A.2}$$

Since

$$P\{T_V > t\} = P\{T_{V1} > t\} \cdot P\{T_{V2} > t\} \tag{A.3}$$

the df of T_V is given by

$$V(t) = 1 - \left\{ \int_{u=t}^{\infty} \lambda_1 F_1{}^c(u)\, du \right\} \cdot \left\{ \int_{u=t}^{\infty} \lambda_2 F_2{}^c(u)\, du \right\}. \tag{A.4}$$

Assuming the renewal property and inserting the general relation

$$V'(t) = \lambda \cdot F^c(t) \tag{A.5}$$

between $V(t)$ and $F(t)$ of the superposed point process into eq. (A. 4), we find the final result eq. (12):

$$F(t) = 1 - \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \cdot \left\{ F_1{}^c(t) \cdot \int_{u=t}^{\infty} F_2{}^c(u)\, du + F_2{}^c(t) \cdot \int_{u=t}^{\infty} F_1{}^c(u)\, du \right\}. \tag{A.6}$$

Unfortunately, we are not able to give the moments of $F(t)$, viz.,

$$E[T^k] = \int_{0-}^{\infty} t^k F'(t)\, dt = -\frac{1}{\lambda} \cdot \int_{0-}^{\infty} t^k V''(t)\, dt \tag{A.7}$$

in terms of the moments of the component processes, except in the case of $k = 1$. Then, we obtain the plausible result $\lambda = \lambda_1 + \lambda_2$, which reflects the law of the conservation of flow in node C.

The second moment $E[T^2]$ of the superposed process, and herewith its coefficient of variation c, can be calculated from eq. (A. 6), (A. 7) in a straightforward way, using the concept of hypo- and hyperexponential substitute processes according to eq. (14a, b). For the algebraic manipulations, a subdivision

into three cases of hypo- and hyperexponential process-type combinations has to be considered. The explicit results are summarized below:

$$c^2 = 2 \cdot \frac{t_1 + t_2}{t_1 \cdot t_2} \cdot E[T_V] - 1 \tag{A.8}$$

where $t_j = 1/\lambda_j$, $j = 1, 2$. $E[T_V]$ can be expressed as a sum of four components

$$E[T_V] = \frac{1}{t_1 t_2} \cdot (I_1 + I_2 + I_3 + I_4). \tag{A.9}$$

The results of the components are as seen below.

Case 1: Superposition of 2 hypoexponential processes (where $t_{11} \leqslant t_{21}$)

$$I_1 = t_{11}^2 \left(\frac{t_2}{2} - \frac{t_{11}}{3} - t_{12}\right) + t_{12} t_1 (t_2 - 2t_{12}) + t_{12}[t_{12}(2t_{21} + 2t_{12} - t_2) - t_{21} t_{22}] \cdot \exp\left(-\frac{t_{21} - t_{11}}{t_{12}}\right)$$

$$I_2 = \frac{t_{12} t_{22}^2}{(t_{12} + t_{22})^2} \cdot (t_{12} t_{21} + t_{12} t_{22} + t_{21} t_{22}) \cdot \exp\left(-\frac{t_{21} - t_{11}}{t_{12}}\right)$$

$$I_3 = \tfrac{1}{2} t_1 t_{11}^2 - \tfrac{1}{3} t_{11}^3$$

$$I_4 = t_{12}^2 \left[t_{11} - t_{21} + \frac{t_{22}}{(t_{12} + t_{22})^2} \cdot (t_{12} t_{21} + t_{12} t_{22} + t_{21} t_{22})\right] \cdot \exp\left(-\frac{t_{21} - t_{11}}{t_{12}}\right).$$

Case 2: Superposition of a hyperexponential process (1) and a hypoexponential process (2)

$$I_1 = p_{11} t_{11}^2 t_2 \left[1 - \left(1 + \frac{t_{21}}{t_{11}}\right) \exp\left(-\frac{t_{21}}{t_{11}}\right)\right] + p_{12} t_{12}^2 t_2 \left[1 - \left(1 + \frac{t_{21}}{t_{12}}\right) \exp\left(-\frac{t_{21}}{t_{12}}\right)\right] - p_{11} t_{11} \left[2t_{11}^2 - (2t_{11}^2 + 2t_{11} t_{21} + t_{21}^2) \cdot \exp\left(-\frac{t_{21}}{t_{11}}\right)\right] - p_{12} t_{12} \left[2t_{12}^2 - (2t_{12}^2 + 2t_{12} t_{21} + t_{21}^2) \cdot \exp\left(-\frac{t_{21}}{t_{12}}\right)\right]$$

$$I_2 = p_{11} \cdot \frac{t_{11} t_{22}^2}{(t_{11} + t_{22})^2} (t_{11} t_{21} + t_{11} t_{22} + t_{21} t_{22}) \cdot \exp\left(-\frac{t_{21}}{t_{11}}\right) + p_{12} \cdot \frac{t_{12} t_{22}^2}{(t_{12} + t_{22})^2} (t_{12} t_{21} + t_{12} t_{22} + t_{21} t_{22}) \exp\left(-\frac{t_{21}}{t_{12}}\right)$$

$$I_3 = p_{11} t_{11}^3 \left[1 - \left(1 + \frac{t_{21}}{t_{11}}\right) \exp\left(-\frac{t_{21}}{t_{11}}\right)\right] + p_{12} t_{12}^3 \left[1 - \left(1 + \frac{t_{21}}{t_{12}}\right) \exp\left(-\frac{t_{21}}{t_{12}}\right)\right]$$

$$I_4 = p_{11} \cdot \frac{t_{11}^2 t_{22}}{(t_{11} + t_{22})^2} (t_{11} t_{21} + t_{11} t_{22} + t_{21} t_{22}) \cdot \exp\left(-\frac{t_{21}}{t_{11}}\right) + p_{12} \cdot \frac{t_{12}^2 t_{22}}{(t_{12} + t_{22})^2} (t_{12} t_{21} + t_{12} t_{22} + t_{21} t_{22}) \exp\left(-\frac{t_{21}}{t_{12}}\right).$$

Case 3: Superposition of two hyperexponential processes

$$I_1 = p_{11} p_{21} \cdot \frac{t_{11}^2 t_{21}^2}{(t_{11} + t_{21})} \quad I_2 = p_{11} p_{22} \cdot \frac{t_{11}^2 t_{22}^2}{(t_{11} + t_{22})}$$

$$I_3 = p_{12} p_{21} \cdot \frac{t_{12}^2 t_{21}^2}{(t_{12} + t_{21})} \quad I_4 = p_{12} p_{22} \cdot \frac{t_{12}^2 t_{22}^2}{(t_{12} + t_{22})}.$$

ACKNOWLEDGMENT

The author thanks Prof. Dr.-Ing. A. Lotze, Director of the Institute of Switching and Data Technics of the University of Stuttgart, for the steady support of this work. He also thanks Dipl.-Ing. R. Ertelt for his contributions during the development and the implementation of the algorithm, and his colleagues Dipl.-Ing. W. Bux and Dipl.-Ing. L. Truong for their help while the author was on leave of absence. The constructive comments of three anonymous reviewers are greatly appreciated.

REFERENCES

[1] Jackson, J. R.: Networks of waiting lines. *Opns. Res.* 5 (1957), pp. 518-521.

[2] Gordon, W. J., Newell, G. F.: Closed queuing systems with exponential servers. *Opns. Res.* 15 (1967), pp. 254-265.

[3] Baskett, F., Chandy, K. M., Muntz, R. R., Palacios, F.: Open, closed and mixed networks of queues with different classes of customers. *J.ACM* 22 (1975), pp. 248-260.

[4] Chandy, K. M., Howard, J. H., Towsley, D. F.: Product form and local balance in queuing networks. *J.ACM* 24 (1977), pp. 250-263.

[5] Lam, S. S.: Queuing networks with population size constraints. *IBM J. Res. Develop.* 21 (1977), pp. 370-378.

[6] Chandy, K. M., Herzog, U., Woo, L.: Parametric analysis of queuing networks. *IBM J. Res. Develop.* 19 (1975), pp. 36-42.

[7] —: Approximate analysis of general queuing networks. *IBM J. Res. Develop.* 19 (1975), pp. 43-49.
[8] Disney, R. L., Cherry, W. P.: Some topics in queuing network theory. In: *Lecture Notes in Economics and Math. Systems, Operations Res. No. 98,* Springer-Verlag, Berlin/Heidelberg/New York (1974), pp. 23-44.
[9] Courtois, P. J.: Decomposability, instabilities, and saturation in multiprogramming systems. *C. ACM* 18 (1975), pp. 371-377.
[10] Kobayashi, H.: Application of the diffusion approximation to queuing networks. (Parts I and II), *J. ACM* 21 (1974), pp. 316-328 and pp. 459-469.
[11] Reiser, M., Kobayashi, H.: Accuracy of the diffusion approximation for some queuing systems. *IBM J. Res. and Develop.* 18 (1974), pp. 110-124.
[12] Gelenbe, E., Pujolle, G.: The behaviour of a single queue in a general queueing network. *Acta Informatica* 7 (1976), pp. 123-136.
[13] Halachimi, B., Franta, W. R.: A diffusion approximate solution to the G/G/k queuing system. *Comput. and Ops. Res.* 4 (1977), pp. 37-46.
[14] Kleinrock, L.: *Communication nets, stochastic message flow and delay.* McGraw-Hill Book Comp., New York/San Francisco/Toronto/London (1964).
[15] Burke, P. J.: Output processes and tandem queues. *Proc. Symp. on Computer Communications Networks and Teletraffic,* New York (1972). Polytechn. Press of the PIB, Vol. 22 (1972), pp. 419-428.
[16] Wolff, R. W.: The effect of service time regularity on system performance. In: *Computer Performance* (Eds. K. M. Chandy and M. Reiser). North Holland Publ. Company (1977), pp. 297-304.
[17] Lazowska, E. D.: The use of percentiles in modeling CPU service time distribution. In: *Computer Performance* (Eds. K. M. Chandy and M. Reiser). North Holland Publ. Company (1977), pp. 53-66.
[18] Kuehn, P.: Analysis of switching system control structures by decomposition. (Forthcoming paper).
[19] Sevcik, K. C., Levy, A. I., Tripathi, S. K., Zahorjan, J. L.: Improving approximations of aggregated queuing network subsystems. In: *Computer Performance* (Eds. K. M. Chandy and M. Reiser). North Holland Publ. Company (1977), pp. 1-22.
[20] Kuehn, P.: Zur optimalen Steuerung des Multiprogramminggrades in Rechnersystemen mit Virtuellem Speicher und Paging. *Lecture Notes in Computer Science,* No. 34. Springer-Verlag, Berlin/Heidelberg/New York (1975), pp. 567-580.
[21] Kleinrock, L.: *Queuing systems.* Vol. 1: *Theory.* J. Wiley and Sons, New York/London/Sydney/Toronto (1975).
[22] Kraemer, W., Langenbach-Belz, M.: Approximate formulae for the delay in the queuing system GI/G/1. *Congressbook, 8th Internat. Teletraffic Congress,* Melbourne (1976), pp. 235-1/8. Also (under slightly changed title), approximate formulae for general single-server systems with single and batch arrivals. *Angewandte Informatik* (1978) pp. 396-402.
[23] Burke, P. J.: The output of a queuing system. *Opns. Res.* 4 (1956), pp. 699-704.
[24] Pack, C. D.: The output of an M/D/1 queue. *Opns. Res.* 23 (1975), pp. 750-760.
[25] Heffes, H.: On the output of a GI/M/n queuing system with interrupted Poisson input. *Opns. Res.* 24 (1976), pp. 530-542.
[26] Daley, D. J.: Notes on queuing output processes. In: *Math. Methods in Queuing Theory.* Springer-Verlag, Berlin/Heidelberg/New York (1974), pp. 351-354.
[27] Marshall, K. T.: Some inequalities in queuing. *Opns. Res.* 16 (1968), pp. 651-665.
[28] Makino, T.: On a study of output distributions. *J. of the Operations Res. Soc. of Japan* 8 (1966), pp. 109-133.
[29] Cox, D. R., Miller, H. D.: *The theory of stochastic processes.* Chapman and Hall Ltd., London (1965).
[30] Takács, L.: A single-server queue with feedback. *BSTJ* 42 (1963), pp. 505-519.
[31] Ertelt, R., Kuehn, P.: Analysis of complex queuing networks for computer systems. *Monograph,* Institute of Switching and Data Technics, University of Stuttgart (1975).
[32] Bux, W., Kraemer, W., Kuehn, P., Wucher, P., Ivancevic, P., Truong, L.: Simulation programs for general queuing networks. *Monographs,* Institute of Switching and Data Technics, University of Stuttgart (1975-1976).
[33] Kraemer, W.: Investigations of systems with queues in series. 22nd Report on Studies in Congestion Theory, Institute of Switching and Data Technics, University of Stuttgart (1975).
[34] Bux, W., Herzog, U.: The phase concept: approximation of measured data and performance analysis. In: *Computer Performance* (Eds. K. M. Chandy and M. Reiser). North Holland Publ. Company (1977), p. 23-38.

★

Paul J. Kuehn (M'77) was born in Gruessau, Germany, in 1940. He received the Dipl.-Ing. and Dr.-Ing. degrees in communications engineering from the University of Stuttgart, Germany, in 1967 and 1972, respectively.

From 1967 to 1973 he was Assistant Professor and from 1973 to 1977 Head of a research group for traffic research in computer and computer communications systems at the Institute of Switching and Data Technics (Prof. Dr.-Ing. A. Lotze) at the University of Stuttgart. From 1975 to 1977 he also was Lecturer for communications switching systems at the University of Erlangen-Nuernberg, Germany. In 1977 he joined Bell Telephone Laboratories in Holmdel, NJ, where he was working in the field of computer communications. Since August 1978 he has been Professor for Communications Switching and Transmission at the Gesamthochschule Siegen, Germany.

Dr. Kuehn is a member of the German Communications Society (NTG), the German Informatics Society (GI), and the German Chapter of ACM.