ANALYSIS OF COMPLEX QUEUING NETWORKS BY DECOMPOSITION

Paul Kühn

Institute of Switching and Data Technics, University of Stuttgart, Fed. Rep. of Germany

Contribution to the 8 th International Teletraffic Congress,
Melbourne, Nov. 10 - 17, 1976

ABSTRACT

In this paper an approximate method for the analysis of complex queuing networks is proposed. The queuing network is of the open network type having N single server queuing stations with arbitrary interconnections. There is only one class of customers (calls) which arrive acc.to general exogenous arrival processes. The service times of the queueing 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 G/G/1. 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 very quickly with the aid of a computer. Numerical results are reported to demonstrate the accuracy of the method. The paper concludes with a discussion on extensions of the method.

1. INTRODUCTION

The traffic flow within computer systems and data networks can be described sufficiently accurate 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 at all so that there is a need for accurate approximate methods, too.

Exact methods are known by J.R.Jackson [1] and W.J.Gordon and G.F.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 recently by F.Baskett, K.M.Chandy, R.R.Muntz, and F.G.Palacios [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. It has also been shown that a parametric analysis can be performed in that cases by reducing the network to a suitable subsystem, cf. K.M.Chandy, U.Herzog, and L.Woo [4]. This principle was also extended to general queuing networks approximately [5].

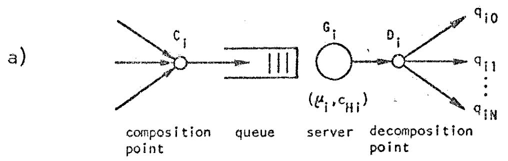
Another class of solution techniques is that of decomposition where the network is broken up into subsystems which are analyzed in isolation. This can be done either by considering the related input and output processes of subsystems or by separation of the total system into a hierarchy of "aggregate systems" with only few interactions between the various levels, cf. R.L.Disney and W.P.Cherry [6] or P.J.Courtois [7], respectively.

The solution technique used in this paper belongs to the decomposition method considering only 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 multi-server stations, multi-class customers, and subnet-configurations. Special derivations are given in the Appendix.

2. QUEUING NETWORK MODEL

2.1 NETWORK STRUCTURE

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



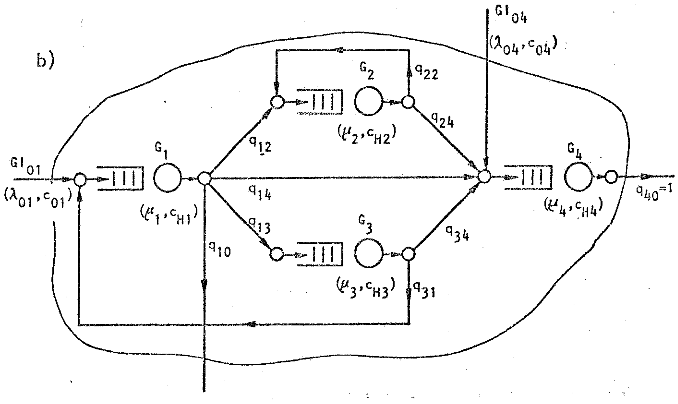


Fig.1. Queuing network structure
a) Elementary queuing station no.i
b) Network model (example with 4 stations)

The elementary queuing station no.i consists of a single server, a single queue with unlimited capacity, a composition point C_i at the input, and a decomposition point D_i at the output. The general queuing network is built from N elementary queuing stations acc.to Fig.1a by arbitrary interconnections. It is assumed that exogenous arriving customers may enter the queuing network at an arbitrary composition point in the most general case. At the other hand, 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 to the outside world (open network).

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

N	Total number of queuing stations
$\underline{Q} = (q_{ij})$	Transition matrix, where
q_{ij}	transition probability for customers branching from station i to a station j, $i = 1,2,\dots,N$, $j = 0,1,\dots,N$.

Herewith, station no. 0 represents the outside world of the queuing network.

2.2 ARRIVAL AND SERVICE PROCESSES

Customers arrive from the outside world acc.to general exogenous arrival processes and they are served at the various stations acc.to general service processes:

$\underline{GI}_o = (GI_{oi})$	Vector of exogenous arrival processes
$\underline{\lambda}_o = (\lambda_{oi})$	Vector of exogenous arrival rates, where $a_{oi}=1/\lambda_{oi}$ is the mean exogenous interarrival time at station i
$\underline{c}_o = (c_{oi})$	Vector of variation coefficients of the exogenous arrival processes
$\underline{G} = (G_i)$	Vector of service processes
$\underline{\mu} = (\mu_i)$	Vector of service rates, where $h_i=1/\mu_i$ is the mean service time at station i
$\underline{c}_H = (c_{Hi})$	Vector of variation coefficients of the service processes

At each station i the exogenous interarrival times T_{oi} and service times T_{Hi} are mutually independent and identically distributed with probability distribution function (df) $A_{oi}(t)$ and $H_i(t)$, respectively, i=1,2,...,N. The latter assumption includes the independence assumption that successive service times of the same customer are independent of each other [8]. The following notations will be used for the df, the k-th ordinary moment, and the variation coefficient of a random variable T:

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

$$E\left[T^k\right] = \int_{0-}^{\infty} t^k F'(t)dt, \quad k = 1,2,... \tag{1b}$$

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

These notations are used for interarrival times T_A, service times T_H, interdeparture times T_D, and arbitrary interarrival times in transition 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 Erlang and hyperexponential processes of the order k, respectively. Finally, we assume that the network is in the stationary state.

2.3 OPERATIONAL STRATEGIES

All customers in the network are treated equally (one class of customers only). A customer leaving station i is branched to a station j independently acc.to the transition probability q_{ij}, i=1,2,...,N, j=0,1,...,N. Queuing customers are scheduled for service acc.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 PRINCIPLES

The analysis method was developed acc.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 only two moments (mean, variation coefficient) of all processes
5. Reduction of the total analysis to few elementary operations to be performed very quickly 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 (i.e. networks with Markovian processes for exogenous arrivals and service times) 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 exception of networks without feedbacks [9]). 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. As an example, consider the queuing station M/G/1 where the mean waiting time depends on the mean and the variance of the service time only (Pollaczek-Khintchine). Additionally, the following procedures are much better to perform for two moments than for whole processes.

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 [10]:

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

In the stationary case, for all stations it must hold:

$$A_i = \lambda_i/\mu_i < 1 , \quad i = 1,2,...,N. \tag{3}$$

A_i is called the offered traffic to station i. The transition rate λ_{ij} of the path from station i to station j follows acc.to

$$\lambda_{ij} = \lambda_i q_{ij} , \quad \begin{array}{l} i = 1,2,...,N, \\ j = 0,1,...,N. \end{array} \tag{4}$$

3.2.2 Mean Values of the Queuing System G/G/1

In the general network case we consider queuing stations acc.to Fig. 2:

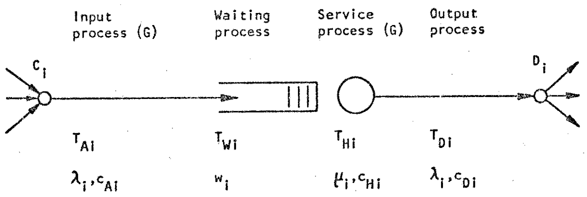


Fig.2. The general queuing system G/G/1

The input (arrival) process is a renewal process with general df G, mean arrival rate λ_i, and variation coefficient c_{Ai}. The service process is also general (G) with mean service or holding time h_i and variation coefficient 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 that station, viz.

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

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

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

In eq.(5c), Ω_i defines the mean queue length at station i.

Exact values are known only for special cases as, e.g.,for the queuing systems M/G/1 and G/M/1. For the general case, a new approximate formula has been developed heuristically which includes the case M/G/1 exactly, cf. W.Krämer and M. Langenbach-Belz [11]:

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

where

$$g(A_i, c_{Ai}^2, c_{Hi}^2) = \begin{cases} \exp\{ -\frac{2(1-A_i)}{3A_i} \cdot \frac{(1-c_{Ai}^2)^2}{c_{Ai}^2+c_{Hi}^2} \} , & c_{Ai} < 1 \\ \exp\{ -(1-A_i) \cdot \frac{c_{Ai}^2 - 1}{c_{Ai}^2+4c_{Hi}^2} \} , & c_{Ai} \geq 1. \end{cases}$$

Analogously, an expression was found for the probability of waiting $W_i = P\{T_{Wi} > 0\}$, cf. [11] . Both formulas were checked by intensive simulations.

3.2.3 Output Process of the Queuing System G/G/1

The output process of a queuing station acc.to Fig.2 is basically characterized by the df of the interdeparture times T_{Di}, viz.

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

Output processes are known explicitly for systems M/M/n and M/D/1, cf. P.J.Burke [12] and C.D.Pack [13], respectively. Apart from some special cases, little is known for more general systems, cf. [14]. Whereas the output of M/M/n is again Markovian, in almost all other cases the output processes are no longer recurrent.

Because of these difficulties we concentrate on the variation coefficient c_{Di} of the output process given that the input process is recurrent. K.T.Marshall [15] gives a general expression for the variation coefficient c_{Di} of the queuing system G/G/1:

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

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

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

or with some simplifications

$$c_{Di}^2 = c_{Ai}^2 + A_i^2(c_{Hi}^2 - c_{Ai}^2) \; . \tag{9b}$$

The solutions eq.(9a,b) include the known exact results for M/G/1, cf. T.Makino [16], as well as for G/G/1 for $A_i \to 0$ and $A_i \to 1$, 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 branched into direction j acc.to a fixed probability q_j, j=0,1,...,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,...,N.

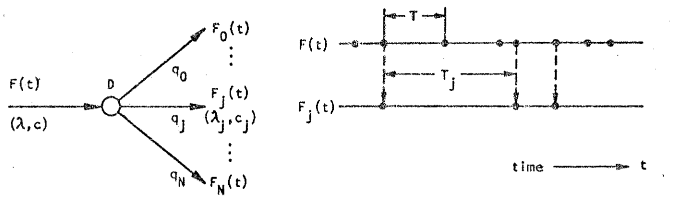


Fig.3. Decomposition of a renewal process into (N+1) component processes

As shown in App.1, the Laplace-Stieltjes (LS)-transform of $F_j(t)$ of the component process j is given by

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

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

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

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

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

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

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 to 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,...,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.

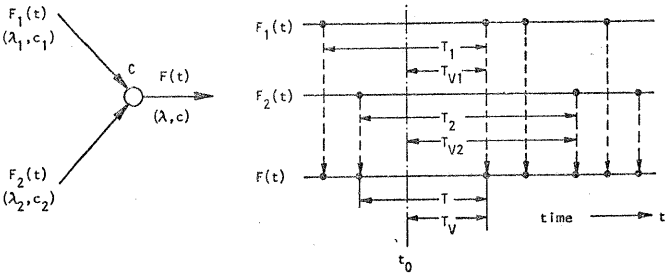


Fig.4. Composition of two renewal component processes

The resulting process is in the general case nonrecurrent. As shown in App.2 by means of the forward recurrence times T_{V1}, T_{V2}, and T_V, the df F(t) of the resulting process is

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

where $F_j^c(t) = 1 - F_j(t)$ the complementary df of $F_j(t)$,j=1,2.

From eq.(12) it follows for t=0 the plausible result for the resulting rate λ which reflects again the law of the conservation of flow in node C:

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

The calculation of the variation coefficient, however, turns out to be rather laborious. For this reason, a concept of simple substitute component processes is introduced by which the operations for the calculation of c are tractable. These substitute processes are as follows:

$$F_j(t) = \begin{cases} \begin{cases} 0 & , \; 0 \leqq t \leqq t_{j1} \\ 1-\exp\{-\varepsilon_{j2}(t-t_{j1})\}, & t \geqq t_{j1} \end{cases} , \; 0 \leqq c_j \leqq 1 & (14a) \\ \\ 1 - p_{j1}\exp(-\varepsilon_{j1}t) - p_{j2}\exp(-\varepsilon_{j2}t), \quad c_j \geqq 1. & (14b) \end{cases}$$

Both substitute processes are simple combinations of phases. Eq.(14a) represents a series of a constant phase T_{j1} and an exponential phase T_{j2}, eq.(14b) the alternative of two exponential phases T_{j1} and T_{j2}, respectively, cf. Fig.5. Both df allow the approximation of arbitrary hypo- and hyperexponential component processes exactly with respect to their first two moments.

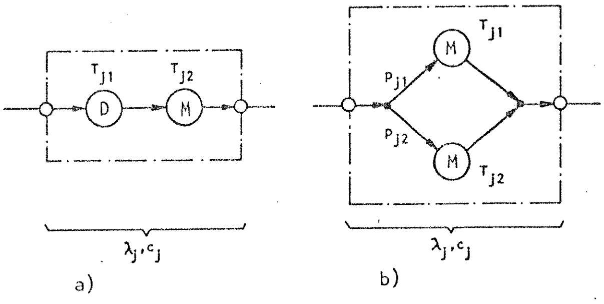


Fig.5. Representation of substitute processes by phases

a) Hypoexponential df $(0 \leqq c_j \leqq 1)$
b) Hyperexponential df $(c_j \geqq 1)$

Using the abbreviation $E\left[T_{j\nu}\right] = t_{j\nu} = 1/\varepsilon_{j\nu}, \nu = 1,2$, the parameters of the substitute processes are as follows:

$$0 \leqq c_j \leqq 1: \quad \varepsilon_{j1} = \lambda_j/(1-c_j) \quad , \varepsilon_{j2} = \lambda_j/c_j \tag{15a}$$

$$c_j \geqq 1: \quad \varepsilon_{j1,2} = \lambda_j\{1 \pm \sqrt{\frac{c_j^2-1}{c_j^2+1}}\}, \; p_{j1,2} = \varepsilon_{j1,2}/2\lambda_j . \tag{15b}$$

$$(p_{j1}t_{j1} = p_{j2}t_{j2})$$

The superposition acc.to eq.(12) is carried out with these two basic substitute processes yielding the variation coefficient c straightforwardly. 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 acc.to Fig.6:

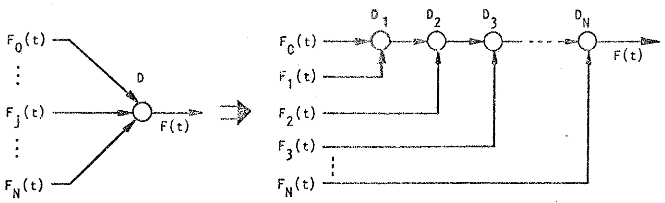


Fig.6. Composition of (N+1) component processes by N recursive compositions of two processes in each case

3.2.6 Reconfiguration by Substitution of Stage-Internal Feedbacks

It turned out that stage-internal feedbacks may affect the assumption of renewal processes in a negative way 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 acc.to Fig.7.

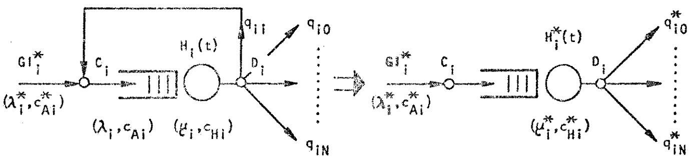


Fig.7. Substitution of stage-internal feedbacks

In the original system a customer is served acc.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 in App.3, 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 $\Psi_i(s)$ the LS-transform of $H_i(t)$. From eq.(16) and Fig.7 follows:

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

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

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

By this procedure, the stage-internal 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 stage with internal feedback, the queuing network becomes reconfigurated. The reconfigurated network differs from the original one with respect to arrival rates, service time distributions, and the transition matrix. After that reconfiguration procedure, the usual quantities without asterisk will be used to describe the reconfigurated network.

It was shown by intensive simulations, that the reconfiguration step yields acceptable accuracy, whereas the analysis without that step results into considerable differences compared with simulations.

3.2.7 Mean Number of Visits at a Station

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

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

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

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

$$e_i(a) = \lambda_i(a)/\lambda_{oa}, \tag{18b}$$

$$e_i(a,b) = \lambda_i(a,b)/\lambda_{bo}(a), \text{where } \lambda_i(a,b) = \lambda_i(a)\cdot p_b(i),$$

$$i,a,b = 1,2,\ldots,N. \tag{18c}$$

It remains to calculate the various arrival rates in eq.(18a-c). The arrival rates λ_i are the solutions of eq.(2). The same procedure acc.to eq.(2) can be used to calculate the rates $\lambda_i(a)$ by setting $\lambda_{oj} = 0$ for $j \neq a$, cf. Fig.8b. The arrival rate $\lambda_i(a,b)$, finally, is that proportion of $\lambda_i(a)$ leaving the network via station b. This proportion is given by the probability $p_b(i)$ for customers entering at station i and leaving the network via station b, cf. eq.(19b).

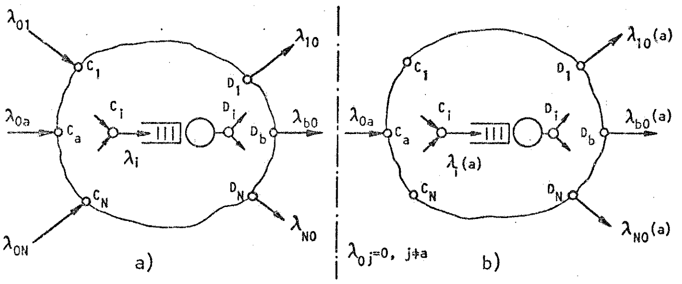


Fig.8. Total and partial arrival rates at station i
a) Network with complete exogenous arrivals
b) Network with exogenous arrivals only at input a

From Fig.8a,b the probabilities p_b and $p_b(a)$ are derived which define the probability that an arbitrary customer leaves the network via station b or that a customer of input a leaves the network via station b, respectively:

$$p_b = \lambda_{bo}/\lambda , \tag{19a}$$

$$p_b(a) = \lambda_{bo}(a)/\lambda_{oa}, \quad a,b = 1,2,\ldots,N. \tag{19b}$$

3.2.8 Mean Flow Times

The flow times T_F, $T_F(a)$, and $T_F(a,b)$ define the random life time of an arbitrary customer, a customer of input a, or a customer of input a who leaves the network via station b, respectively. The mean values are found directly by considering a "test customer" moving through the network:

$$f(a,b) = E\left[T_F(a,b)\right] = \sum_{i=1}^{N} e_i(a,b) f_i , \tag{20a}$$

$$f(a) = E\left[T_F(a)\right] = \sum_{i=1}^{N} e_i(a) f_i = \sum_{b=1}^{N} f(a,b) p_b(a), \tag{20b}$$

$$f = E\left[T_F\right] = \sum_{a=1}^{N} \frac{\lambda_{oa}}{\lambda} \cdot f(a) = \sum_{i=1}^{N} e_i f_i = \frac{1}{\lambda} \cdot \sum_{i=1}^{N} E\left[X_i\right] . \tag{20c}$$

The latter expressions in eq.(20c) are found by insertion of eqs.(20b),(18b),(18a),and (5c); the last expression is identical to Little's theorem applied to the total network.

3.3 QUEUING NETWORK ANALYSIS ALGORITHM

3.3.1 Standard Procedures

The algorithm is based on a number of procedures for standard operations as discussed in Section 3.2. These are:

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

RECONF Reconfiguration of the queuing network by substitution of stages with feedback through stages without feedback and transformation of the network parameters

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

DECOMP Decomposition of the departure process (λ_i,c_{Di}) of queuing station i into (N+1) component processes (λ_{ij},c_{ij}) at D_i of station i, $j=0,1,\dots,N$

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

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

FLOWTIME Calculation of mean flow times between arbitrary input and output ports of the network as well as means of total flow times with respect to all customers or all customers of a certain input, respectively.

3.3.2 Analysis Algorithm

In networks without feedback the analysis can be carried out straightforwardly. There exists always a sequence for analyzing the network stage by stage starting from one of those stages having only exogenous arrivals.

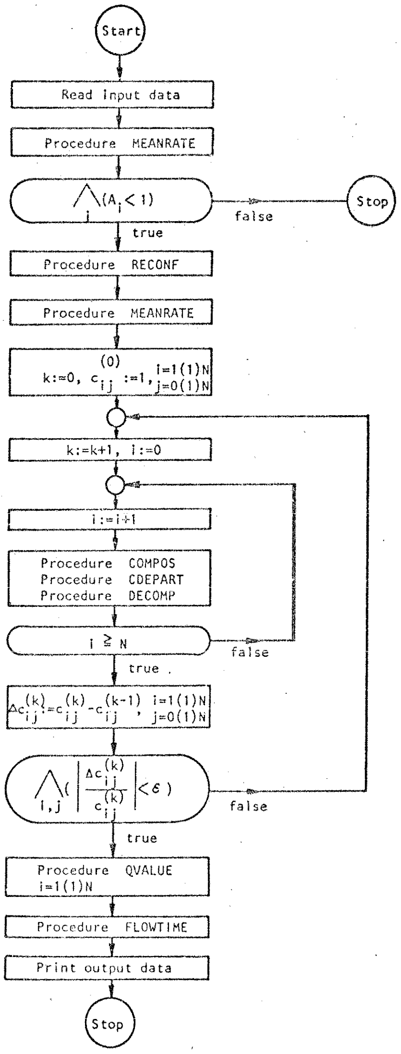


Fig.9. Principal flow chart of the decomposition analysis algorithm

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 variation coefficients c_{ji}, $i,j=1,2,\dots,N$. This problem is solved by iteration having the additional advantage to be 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 $5N^2$ storage capacity for data. It has been implemented by an ALGOL computer program [19], and its results were checked by a simulation program for general queuing networks [20]. Finally, we state that in the special case of pure Markovian networks the algorithm yields the exact results.

4. NUMERICAL RESULTS

In this chapter, several results are reported to show the accuracy of the algorithm for basic operations and whole networks, as well.

4.1 STANDARD OPERATIONS

4.1.1 Mean Values and Output Process of the Queuing System G/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. [11], and will not be reproduced here.

Concerning the output process, in Fig.10a,b the functions $c_D^2(A)$ are given for queuing systems of the type $E_2/G/1$ and $H_2/G/1$ acc.to eq.(9a), respectively. The simulation results with a 95% confidence interval show a reasonable accuracy of the approximate solution.

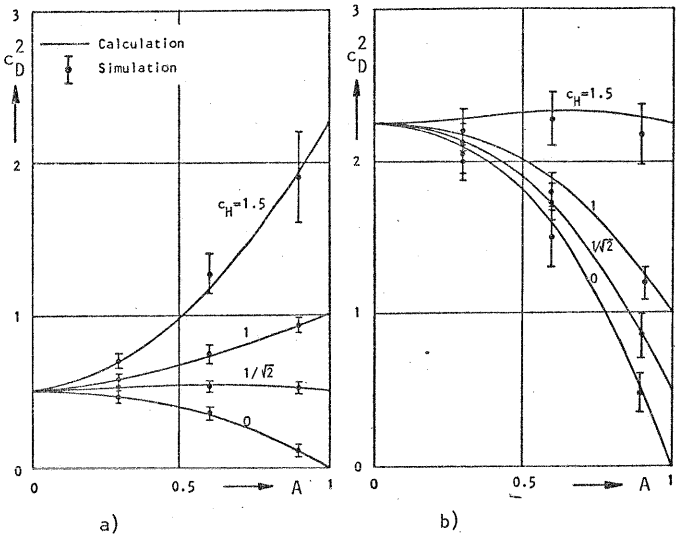


Fig.10. Variation coefficient c_D versus the offered traffic A (parameter c_H)
a) Queuing system $E_2/G/1$
b) Queuing system $H_2/G/1$

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 variation coefficient c and branching probability q_j as parameter acc.to Section 3.2.4, respectively. The results of the decomposition operation are shown in Fig.11a for renewal processes and in Fig.11b for nonrenewal processes. The nonrenewal processes were represented as output processes of queuing systems of the type M/G/1 for A = 0.6 (note, that the output process is renewal again for A → 0 and A → 1). The curves hold exactly only if the original process is renewal, cf. Fig.11a. For nonrenewal processes the calculated curves are still within the confidence intervals of the simulation, cf. Fig.11b. The accuracy does not depend remarkably on the parameter q_j.

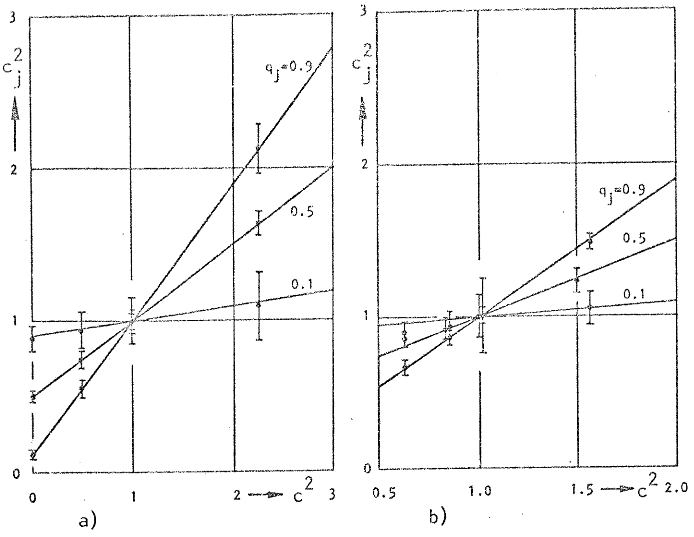


Fig.11. Variation coefficient c_j of decomposed processes versus variation coefficient c of the original process (parameter q_j)
a) Decomposition of renewal processes
b) Decomposition of nonrenewal processes

4.1.3 Composition of Point Processes

The results of the composition operation on two component processes acc.to Section 3.2.5 are shown in Fig.12. The figures represent the squared variation coefficient c^2 of the superposed process dependent on the variation coefficients c_1 (abscissa) and c_2 (parameter) for two renewal component processes (cf. Fig.12a) or nonrenewal processes (cf. Fig.12b), respectively. The nonrenewal processes were realized as output processes of M/G/1 at A = 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.

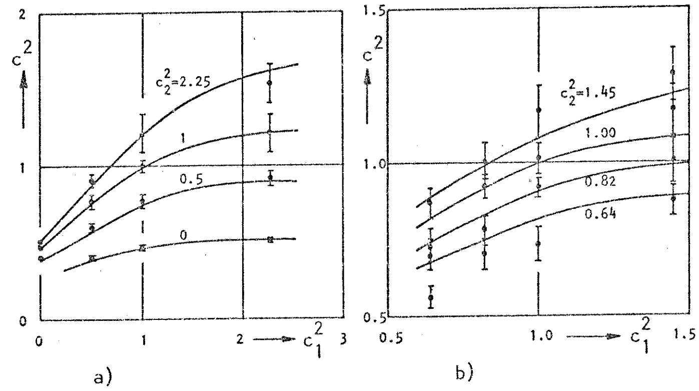


Fig.12. Variation coefficient c of the superposed process dependent on variation coefficients c_1 and c_2 of the component processes
a) Composition of renewal processes
b) Composition of nonrenewal processes

4.1.4 Substitution of Stage-Internal Feedbacks

In Fig.13 the calculated results of the reconfigurated queuing station without stage-internal feedback acc.to Section 3.2.6 are compared with simulations for the original system (the subscript i is omitted). The figures show the variation coefficient c_D^* of the output process of customers leaving the station (cf. Fig.13a) as well as the mean total flow time f of all outside arriving customers (cf. Fig.13b) versus the variation coefficient c_H of service times of the original system with parameter c_A^* of the arrival process of outside arriving customers. Both results are in good accordance with the simulation results. Similar results were obtained for the comparison of both systems by simulations only, although the equivalence is exact only in case of Poisson arrivals [18].

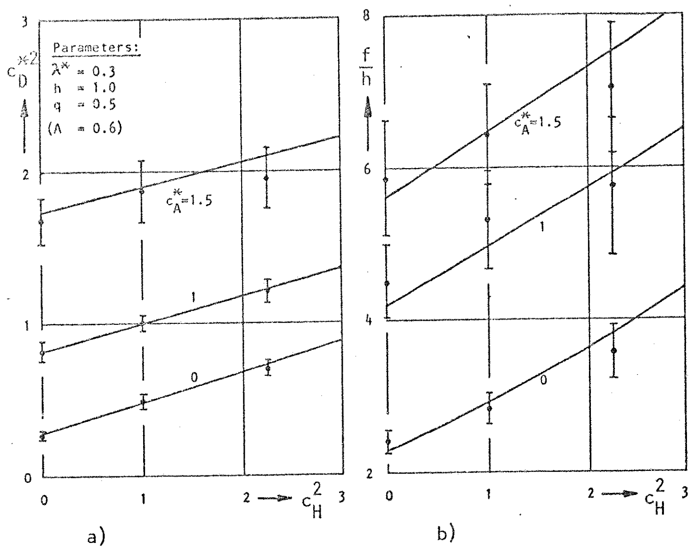


Fig.13. Equivalence of substitute systems without stage-internal feedback
a) Variation coefficient c_D^* of the output process versus c_H with parameter c_A^*
b) Mean total flow time f versus c_H with parameter c_A^*

4.2 ANALYSIS OF QUEUING NETWORKS

The accuracy of the decomposition method will be demonstrated by a network example consisting of N=9 queuing stations with some interconnections including several feedbacks, cf. Fig.14. Exogenous arrival processes are assumed to be Markovian whereas the service times are generally distributed. The figures at the transition paths represent the transition probabilities.

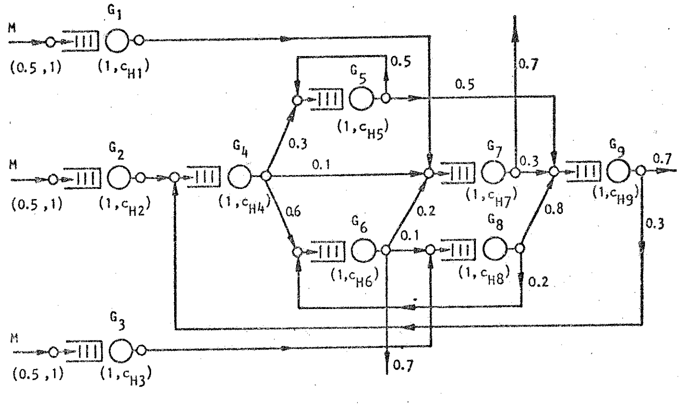


Fig.14. Queuing network example with 9 queuing stations

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}). In Fig.15 the mean total flow time f and the mean flow time f_4 of the interior station number 4 are drawn (solid curves) versus the variation coefficient c_{H1-9} in case of homogeneous servers (cf. Fig.15a) or versus c_{H4-9} with parameter c_{H1-3} in case of heterogeneous servers (cf. Fig.15b), respectively.

In Fig.15a two more curves are shown for comparison with different analysis methods when all arrival processes at each station are assumed to be Markovian (dashed curves), or when all arrival and service processes at each station are assumed to be Markovian (dotted curves), respectively.

Compared with the simulation results, the proposed two-moment method yields acceptable results wheras the neglection 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.

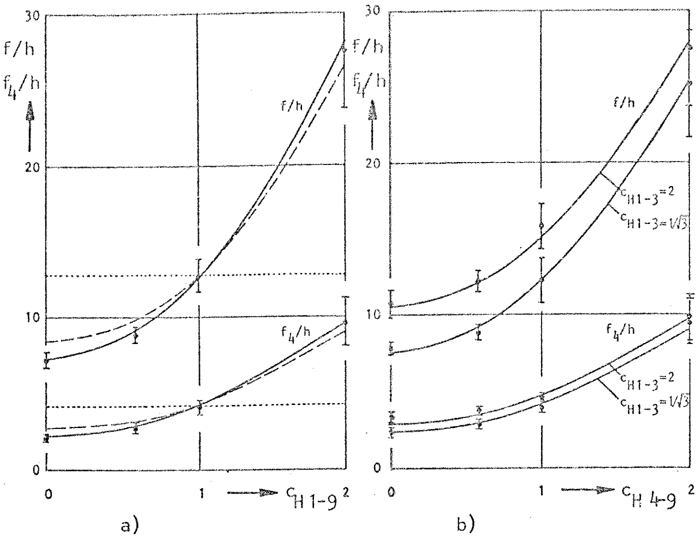


Fig.15. Mean flow times f and f_4 versus the variation coefficient c_H of network service stations
a) Homogeneous servers
b) Heterogeneous servers

5. 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 rates or closed networks with a fixed number of customers within the network as in the network theorems for Markovian arrival and service processes, cf.[2-3]. 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, allows easily 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 acc.to Fig.16a. The analysis is performed analogously by inserting the corresponding results for a $G/G/n_i$ queue with respect to the variance coefficient 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 > 1 classes of customers arriving from the outside world. Customers are classified acc.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 distinct queues acc.to their class index r, r=1,2,...,R, cf. Fig.16b. Waiting customers are selected for service by a scheduler with any nonpreemptive strategy, e.g., a nonpreemptive priority. A customer of class r leaving station i is branched to station j and changes into class s with probability ${}_{rs}q_{ij}$, r,s = 1,2,...,R, i = 1,2,...,N, j = 0,1,...,N.

For the analysis, in a first step the arrival rate ${}_r\lambda_i$ of class r customers at station i must be determined from a system of NR equations analogously to eq.(2). Herewith, the total arrival rate λ_i and the transition rates ${}_{rs}\lambda_{ij}$ are determined, too. Then, the general R-class queuing system G/G/1 must be analyzed considering only two moments of processes and the underlying scheduling discipline. The consistency of the variation coefficients 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 with one input port and one output port as shown in Fig.16c. 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 favourable in cases where the global subnet behaviour with respect to input and output processes is sufficient to describe its influence on the residual network behaviour, or in special cases when the decomposition method becomes worse if it is applied to the elements of the subnet. As an example for 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 stages as, e.g.,in case of closed loops or series stages with constant service times, cf. [21], 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.

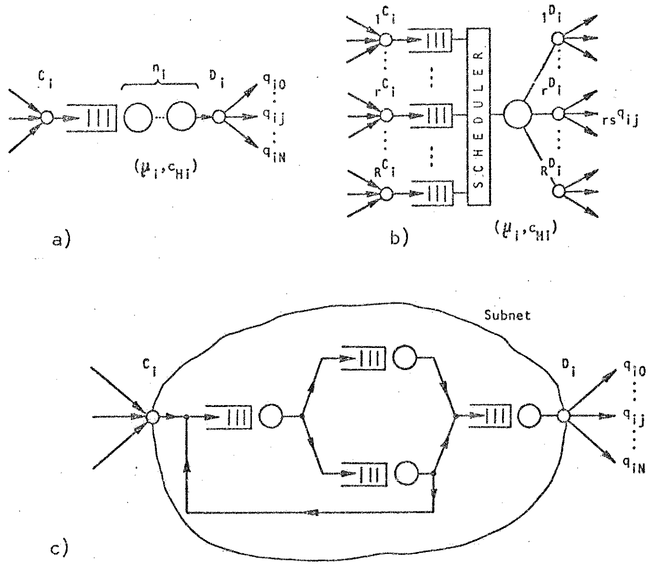


Fig.16. Generalized subsystems
a) Multi-server queuing station
b) Multi-queue (multi-class) station
c) Subnet

APPENDIX

APP.1. DECOMPOSITION OF RENEWAL PROCESSES

Given a stationary renewal point process with a random interarrival time T between successive events (arrivals) with df $F(t) = P\{T \leqq t\}$. At the decomposition point D an arriving customer is branched into direction j with probability q_j, cf. Fig.3. The random interarrival times T_j of the component process are constituted as sums of a random number X of successive realizations of the random interarrival time $T^{(\nu)}$ of the original process, i.e.

$$T_j = \sum_{\nu=1}^{X} T^{(\nu)} . \qquad (A.1)$$

Be

$$p_x = P\{X=x\} = \begin{cases} 0 & , \ x = 0 \\ q_j(1-q_j)^{x-1} & , \ x \geqq 1 \end{cases} \qquad (A.2a)$$

with generating function

$$G(z) = \sum_{x=0}^{\infty} p_x z^x = \frac{q_j z}{1 - (1-q_j)z} , \qquad (A.2b)$$

and

$$\phi(s) = \int_{0-}^{\infty} \exp(-st)F'(t)dt , \qquad (A.3a)$$

$$\phi_j(s) = \int_{0-}^{\infty} \exp(-st)F_j'(t)dt , \qquad (A.3b)$$

the LS-transforms of the df F(t) or $F_j(t)$ of the random variables T or T_j, respectively.

Then,

$$F_j(t) = \sum_{x=0}^{\infty} p_x \cdot P\{T_j \leq t | X = x\} \quad , \qquad (A.4a)$$

where $P\{T_j \leq t | X = x\}$ the conditional df of the sum of exactly x mutually independent random variables $T^{(\nu)}, \nu = 1,2,\ldots,x$, with identical df F(t). Thus,

$$\phi_j(s) = \sum_{x=0}^{\infty} p_x \cdot \left[\phi(s)\right]^x . \qquad (A.4b)$$

Compared with eq.(A.2b) we find:

$$\phi_j(s) = G(\phi(s)) = \frac{q_j \phi(s)}{1 - (1-q_j)\phi(s)} , \qquad (A.5)$$

with mean $E\left[T_j\right]$ and variance $VAR\left[T_j\right]$ as follows:

$$E\left[T_j\right] = E\left[T\right] \cdot E\left[X\right] \quad , \qquad (A.6a)$$

$$VAR\left[T_j\right] = E\left[T\right]^2 \cdot VAR\left[X\right] + VAR\left[T\right] \cdot E\left[X\right] . \qquad (A.6b)$$

Inserting

$$E\left[X\right] = 1/q_j, \quad VAR\left[X\right] = (1-q_j)/q_j^2 \quad , \qquad (A.7a)$$

$$E\left[T\right] = 1/\lambda \; , \quad VAR\left[T\right] = c^2/\lambda^2 \quad , \qquad (A.7b)$$

in eq.(A.6a,b), the result eq.(11a,b) is obtained.

APP.2. 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 be $F(t) = P\{T \leq t\}$, where T the interarrival time of the superposed process, cf. Fig.4.

Following the theory of renewal processes acc.to D.R.Cox and H.D.Miller [17], the forward recurrence times T_{V1}, T_{V2}, and T_V are introduced being the intervals between an arbitrary instant t_o 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_o and their density function is given by

$$V_j'(t) = \lambda_j \cdot F_j^c(t) \quad , \quad j = 1,2, \qquad (A.8a)$$

where $F_j^c(t) = 1 - F_j(t)$ the complementary df of $F_j(t)$. From eq.(A.8a) the complementary df of T_{Vj} follows by integration

$$V_j^c(t) = \int_{u=t}^{\infty} \lambda_j F_j^c(u)\,du \quad , \quad j = 1,2. \qquad (A.8b)$$

Since

$$P\{T_V > t\} = P\{T_{V1} > t\} \cdot P\{T_{V2} > t\} \; , \qquad (A.9)$$

the df of T_V is given by

$$V(t) = 1 - \{\int_{u=t}^{\infty} \lambda_1 F_1^c(u)\,du\} \cdot \{\int_{u=t}^{\infty} \lambda_2 F_2^c(u)\,du\} . \qquad (A.10)$$

Inserting the general relation

$$V'(t) = \lambda \cdot F^c(t) \qquad (A.11)$$

between V(t) and F(t) of the superposed point process into eq.(A.10), we find the final result acc.to eq.(12):

$$F(t) = 1 - \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \cdot \{F_1^c(t) \cdot \int_{u=t}^{\infty} F_2^c(u)\,du + F_2^c(t) \cdot \int_{u=t}^{\infty} F_1^c(u)\,du\} . \qquad (A.12)$$

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

$$E\left[T^k\right] = \int_{0-}^{\infty} t^k F'(t)\,dt = -\frac{1}{\lambda} \cdot \int_{0-}^{\infty} t^k V''(t)\,dt \quad , \qquad (A.13)$$

in terms of the moments of the component processes except in case of k=1. Then, we obtain from eq.(A.12),(A.13) the plausible result $\lambda = \lambda_1 + \lambda_2$, which reflects the law of the conservation of flow in node C.

The second moment $E\left[T^2\right]$ of the superposed process, and herewith its variation coefficient c, can be calculated from eq.(A.12),(A.13) straightforwardly using the concept of hypo- and hyperexponential substitute processes acc.to eq.(14a,b). For the algebraic manipulations, three subcases of hypo-and hyperexponential combinations must be distinguished. The explicit results are somewhat extensive and will be omitted here.

APP.3. SUBSTITUTION OF STAGE-INTERNAL FEEDBACKS

Given a queuing station i with stage-internal feedback acc. to Fig.7. Arriving customers from the outside of that station receive a geometrically distributed number of random service phases possibly interleaved by phases of other customers. A reconfigurated station without feedback is formed by giving each customer arriving from the outside its total service time continuously. Thus, the substitute service time T_{Hi}^* is the sum of a random number X of mutually independent service times $T_{Hi}^{(\nu)}$, $\nu = 1,2,\ldots,X$, with identical df $H_i(t)$. The service time df $H_i^*(t)$ is found by the same arguments as in case of the decomposition procedure, cf. App.1, considering the counterparts q_j and $(1-q_{ii})$, respectively.

The proof of the exact analogy between the stations with and without feedback was given by L.Takács [18] in case of Poisson arrivals. The analogy holds for the distribution of queue size and the mean flow time. The same argument has been adopted heuristically for the general case above.

ACKNOWLEDGEMENT

The author wishes to express his thanks to the Federal Ministry of Research and Technology (BMFT) of the Fed. Rep. of Germany for supporting this work.

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., Herzog,U., Woo,L.: Parametric analysis of queuing networks. IBM J. Res. Develop. 19(1975), pp. 36-42.

[5] Chandy,K.M., Herzog,U., Woo,L.: Approximate analysis of general queuing networks. IBM J. Res. Develop. 19(1975), pp. 43-49.

[6] 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.

[7] Courtois,P.J.: Decomposability, instabilities, and saturation in multiprogramming systems. C.ACM 18(1975), pp. 371-377.

[8] Kleinrock,L.: Communication nets, stochastic message flow and delay. McGraw-Hill Book Comp., New York/San Francisco/Toronto/London (1964).

[9] 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.

[10] Kleinrock,L.: Queuing systems. Vol.1: Theory. J.Wiley and Sons, New York/London/Sydney/Toronto (1975).

[11] Krämer,W., Langenbach-Belz,M.: Approximate formulae for the delay in the queuing system GI/G/1. Congressbook, 8th Internat. Teletraffic Congress, Melbourne (1976).

[12] Burke,P.J.: The output of a queuing system. Opns. Res. 4(1956), pp. 699-704.

[13] Pack,C.D.: The output of an M/D/1 queue. Opns. Res. 23(1975), pp. 750-760.

[14] Daley,D.J.: Notes on queuing output processes. In: Math. Methods in Queuing Theory. Springer-Verlag, Berlin/Heidelberg/New York (1974), pp. 351-354.

[15] Marshall,K.T.: Some inequalities in queuing. Opns. Res. 16(1968), pp. 651-665.

[16] Makino,T.: On a study of output distributions. J. of the Operations Res. Soc. of Japan 8(1966), pp. 109-133.

[17] Cox,D.R., Miller,H.D.: The theory of stochastic processes. Chapman and Hall Ltd., London (1965).

[18] Takács,L.: A single-server queue with feedback. BSTJ 42(1963), pp. 505-519.

[19] Ertelt,R., Kühn,P.: Analysis of complex queuing networks for computer systems. Monograph, Institute of Switching and Data Technics, Univ. of Stuttgart (1975).

[20] Bux,W., Krämer,W., Wucher,P.: Simulation of queuing networks. Monograph, Institute of Switching and Data Technics, Univ. of Stuttgart (1975).

[21] Krämer,W.: Investigations of systems with queues in series. 22nd Report on Studies in Congestion Theory, Institute of Switching and Data Technics, Univ. of Stuttgart (1975).