1er COLLOQUE AFCET–SMF DE MATHÉMATIQUES APPLIQUÉES

FIRST MEETING AFCET–SMF ON APPLIED MATHEMATICS

4-8 Septembre 1978

Ecole Polytechnique
PALAISEAU (FRANCE)

Tome I

CONFERENCES - INVITED PAPERS
COMMUNICATIONS THEMES I, II

AFCET : Association Française pour la Cybernétique Économique et Technique
156, boulevard Péreire
B.P. 571
75826 Paris Cédex 17

SMF : Société Mathématique de France
11, rue Pierre et Marie Curie
75231 Paris Cédex 05

First Meeting
AFCET - SMF
On applied Mathematics
4-8 September 1978
Ecole Polytechnique
PALAISEAU (FRANCE)
Tome I, 35-51.

EXACT AGGREGATION IN QUEUEING NETWORKS

P. J. COURTOIS

We give necessary and sufficient conditions under which an aggregation procedure yields the exact limiting probability distribution of a stochastic matrix. Vantilborgh's conditions for exact aggregation in exponential queueing networks [Vantilborgh, 1978] are extended to all closed networks with product form. Other results suggest that these conditions remain valid for a wider class of networks.

1. INTRODUCTION

Interest in queueing network models has been revived by the development of complex computer communication networks towards which data processing technology has been heading in the past decade. If realistic enough, these models could become a cheap and reliable tool to evaluate the cost-effectiveness of the many possible alternatives offered by the design of a communication network.

The study of queueing networks models has taken two main directions. A first approach consists in representing the whole network as an homogeneous Markov process with denumerable multidimensional state space [Jackson 1963, Gordon & Newell 1965, Baskett & al 1975, Gelenbe & Muntz 1976, Cohen, 1977] ; these models are adequate to obtain equilibrium solutions but break down if

---

MBLE Research Laboratory, 2, av. Van Becelaere, 1170-Brussels, Belgium.

the state space becomes too large or if transient or non-Markovian processes are involved. In the other approach [Disney & Cherry 1973, Courtois 1972, 1977] the network is decomposed into subnetworks, possibly single server systems, which can be analyzed in isolation; one hopes with this artifice to exploit the many results known in queueing theory on single server systems.

However, most of these results on single server systems assume that the input process is, like the service process, a renewal process; unfortunately, in queueing networks, input processes are generally the departure processes from other queues and, therefore, are not renewal, except for very special queues [see e.g. Daley 1976] . Approximate solutions can nevertheless be obtained by considering these input processes as being renewal [Kühn 1976]; this approximation is especially accurate if the network can be decomposed into subsystems such that interactions between subsystems are weak compared to the interactions within these subsystems [Courtois 1977] . But if exact solutions are sought by decomposition, exact models of subnetworks must be constructed.

Chandy & al [1975] proposed a method to construct a subsystem of a network with exponential server such that the queue length distributions within this subsystem are the same as in the original network. In a recent paper Vantilborgh [1978] gave the necessary and sufficient conditions under which the decomposition of an exponential network into subsystems yields the exact equilibrium state probability distribution of the network. The purpose of this paper is to extend these conditions to arbitrary stochastic systems and to a wider class of queueing network models.

2. BASIC MODEL

The first class of networks we shall consider consists of an arbitrary number of *servers* $R_1,\ldots,R_L$ and an arbitrary number K of *classes* of customers. The length of a service is an i.i.d. random variable with a given probability distribution for each class of customers at each server. Customers travel through the network and change classes according to *routing probabilities*; a customer of class k who completes service at R_i requires service from R_j in class r with probability $p_{i,k;j,r}$ ($\sum_{j=1}^{L} \sum_{r=1}^{K} p_{i,k;j,r} = 1$ for all pairs (i,k)). Customers are selected for service by a server according to the *service discipline* associated with this server.

The state of the network at time t is defined by the vector $(s_1,s_2,\ldots,s_L,t)$ where s_ℓ defines the conditions of service prevailing at server R_ℓ at time t; typically, the definition of s_ℓ consists of the numbers $n_{\ell,r}$ of customers of each class r in queue or in service at R_ℓ and, if necessary, of supplementary variables indicating the remaining service requirements and the positions in the queue of these customers at time t.

The network is *open* if customers may arrive in the network from outside sources, and may leave the network. Otherwise the network is *closed* and its total number of customers remains constant.

If $P(s_1,\ldots,s_L,t)$ is the probability that the network is, at time t, in state $(s_1,\ldots,s_L)$, which is assumed to be a feasible state, we are interested in the limiting probability distribution

$$P(s_1,\ldots,s_L) = \lim_{t\to\infty} P(s_1,\ldots,s_L,t) \; ; \qquad (1)$$

the conditions for this limit to exist are supposed to be satisfied. From this limiting distribution, it is easy to derive performance measures such as server utilization factors, queue length distributions, etc ...

A remarkable property of these models is that, under certain conditions,this limiting distribution exhibits what has been called the *product-form* :

$$P(s_1,\ldots,s_L) = G\, f_1(s_1) \ldots f_L(s_L) \qquad (2)$$

where G is a normalizing constant chosen so that the probabilities over all states of the network sum up to 1, and each $f_\ell(.)$ is a function wich depends on the type of server R_ℓ. This product form was introduced by Jackson [1963] for networks with exponential service times. With some restrictions on the service disciplines, it was generalized to multi-class networks with service distributions which have rational Laplace transforms in [Baskett & all 1975] ; Chandy & al [1977] extended these results to arbitrary differentiable service distributions through the technique of supplementary variables. It was recognized in [Baskett & al 1975] that it is sufficient for the product form to exist that at each server either the service distribution be exponential or the service discipline be *immediate*, i.e. customers begin to receive service immediately upon entering the queue; such immediate disciplines include first in-first out, processor sharing, infinite server, etc... but exclude first in-first out. This sufficient condition was proven in [Chandy & al 1977] to be also necessary when each server has a class independent service discipline, i.e. treats all classes of customers alike.

Another property of queueing networks is the *insensitivity property*: under certain conditions service distribution enter only through their means in the expression of the distribution $P(s_1,\ldots,s_L)$; in particular this is verified by all the networks mentioned above for which the product form has been demonstrated : each function $f_\ell(.)$ depends on the service distribution of server ℓ through its mean only. The general conditions for the insensitivity of steady-state distributions in generalized semi-Markov processes has been studied by several authors, see e.g. [Schassberger 1977].

So far, an explicit solution for the equilibrium distribution (1) has been found for networks with product form only; this is due to the method which has been followed, namely to assume the existence of the product form, to inject this product form into the network balance equations which, for each state, equate the probability of leaving and of entering this state, and to solve for the functions $f_\ell(.)$.

3. DECOMPOSITION AND AGGREGATION

Thus, despite of the progress which has been made, the family of networks models which can be explicitly solved is still rather limited in view of present practical needs. Another problem is the rapid growth of the number of states, and thus of the number of balance equations, with the network complexity.

As said in the introduction, these problems call for another direction of investigation : *decomposition methods*. The network is decomposed into subnetworks which are analyzed in isolation; each analysis focuses on the stochastic process which represents the flow of customers through a subnetwork; this process is used in place of the subnetwork as input to the remainder of the network to analyze the distribution of customers within this remainder.

Queueing networks have structural properties which are propitious to this type of approach. This is easier to demonstrate if we assume that the network can be represented by a Markov transition matrix; existing results on the decomposition of stochastic matrices [Courtois 1977] can then be used.

Consider a general stochastic system the time behavior of which is represented by the matrix equation

$$y(t+1) = y(t)\ Q\ , \qquad (3)$$

where an element $y_i(t)$, $i = 1,\ldots,n$, of the row vector $y(t)$ is the probability that the system is in state i at time t; Q is the stochastic matrix of the transition probabilities between these states; suppose that Q is regular so that at least one limiting probability vector ν exists which is independent of t (but not necessarily of the initial condition $y(0)$) :

$$\nu = \nu\, Q = \lim_{t\to\infty} y(t) \ . \qquad (3 \text{ bis})$$

The most general method to obtain ν by *decomposition* is to partition Q into A principal submatrices Q_{II}, $I = 1,\ldots,A$:

$$Q = \begin{bmatrix} Q_{11} & \cdots & Q_{1A} \\ \vdots & Q_{II} & \vdots \\ Q_{A1} & \cdots & Q_{AA} \end{bmatrix} \qquad (4)$$

and to construct from each principal submatrix Q_{II} a stochastic matrix Q_I^* by adding (in a way which will be discussed later) to each row i_I of Q_{II} the sum $\sum_{\substack{J=1 \\ J\neq I}}^{A} \sum_{j=1}^{n(J)} q_{i_I j_J}$, $q_{i_I j_J}$ denoting the (i,j) element of Q_{IJ} and n(J) being the order of Q_{JJ}, ($\sum_{J=1}^{A} n(J) = n$). Thus,

$$Q = Q^* + \varepsilon\, C$$

where Q^* is a completely decomposable matrix

$$Q^* = \begin{bmatrix} Q_1^* & & & \\ & Q_2^* & & \\ & & \ddots & \\ & & & Q_A^* \end{bmatrix}$$

with the elements not displayed being equal to 0; ε can be taken as the maximum probability of leaving a partition of states

$$\varepsilon = \max_{i,I} \Big(\sum_{J\neq I} \sum_{j} \sum_{J} q_{i_I j_J} \Big) \qquad (5)$$

which implies that

$$\max_{i,I} \sum_{J\neq I} \sum_{j} \sum_{J} |c_{i_I j_J}| = 1 \ ;$$

the matrices Q_I^* shall be referred to as the *aggregates* of Q.

A vector z which approximates ν can be obtained by the following *aggregation procedure*. Denoting $\nu^*(I)$ the equilibirum probability vector of Q_I^*, so that $\nu^*(I)Q_I^* = \nu^*(I)$, a matrix T of transitions between groups of states is constructed as a AxA matrix of elements

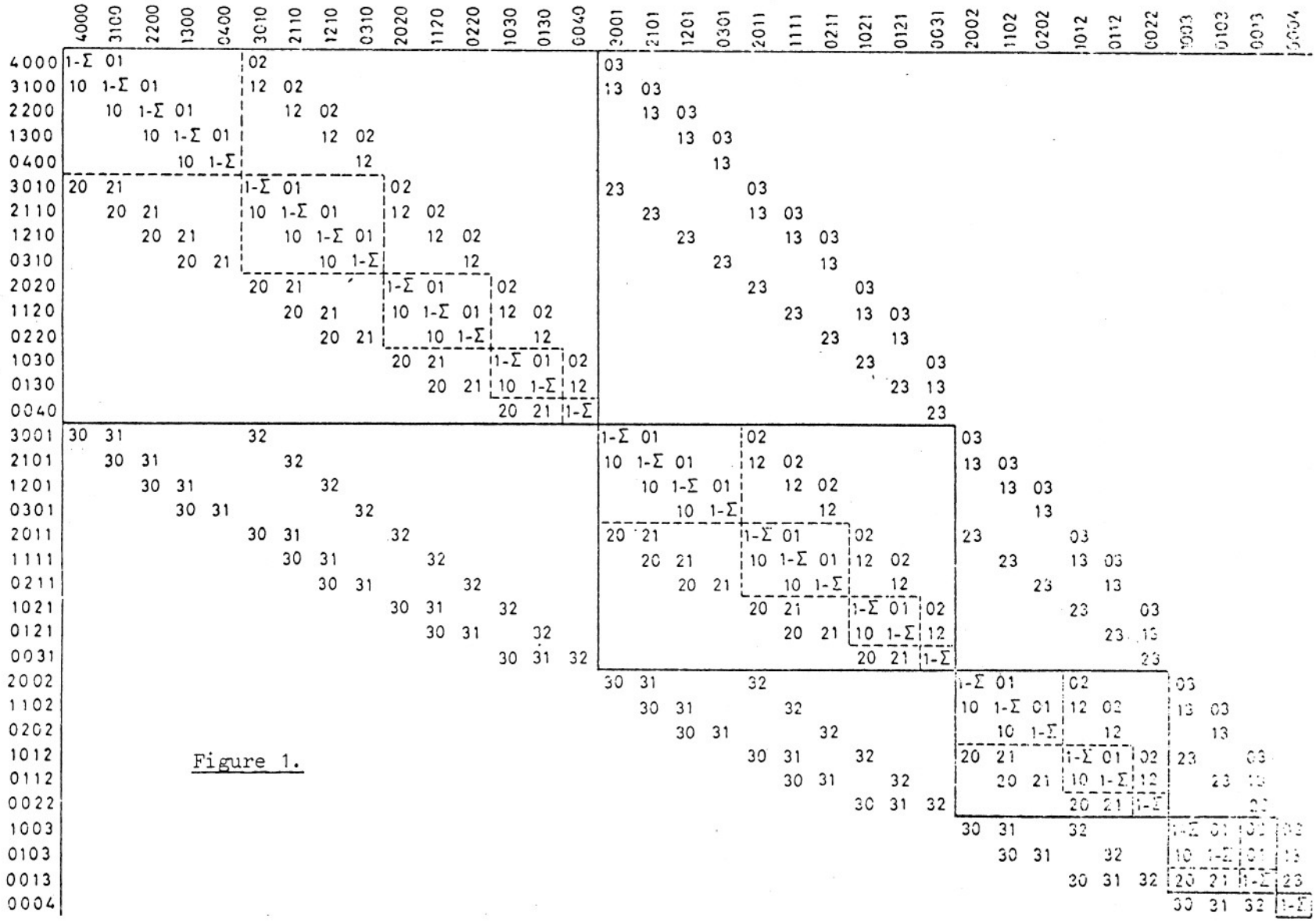

Figure 1.

$$T_{IJ} = \sum_{i \in I} \nu_i^*(I) \sum_{j \in J} q_{i_I j_J} ;$$

the equilibrium vector $Z = [Z_1 \ldots Z_I \ldots Z_A]$ of T, $Z = ZT$, is used to obtain $z_{i_I} = Z_I \nu_i^*(I)$ as an approximation to ν_{i_I}. If necessary, an aggregate Q_I^* can be decomposed into sub-aggregates, and the same procedure can be used to obtain the vector $\nu^*(I)$, and so forth (hierarchical aggregation).

The above aggregation procedure clearly has the advantage of reducing the analysis of a n-states system Q to the separate analysis of A smaller systems Q_I^*. The reduction is still more important in *closed* queueing network models. Consider a closed network with a single class of N customers, exponential service times with parameter μ_ℓ for server R_ℓ, and fixed routing probabilities p_{ij}. The state of the network is entirely specified by the L-tuple $(n_1,\ldots,n_L)$, n_ℓ being the number of customers present at resource R_ℓ. For a time unit small enough to make negligeable the probability of more than one customer completing service simultaneously, this model corresponds to a discrete time homogeneous Markov chain ; if the states $(n_1,\ldots,n_L)$, $\sum n_\ell = N$, are lexicographically ordered, the transition matrix takes the form displayed in figure 1 for a network with $N = L = 4$; elements not displayed are zero; a (i,j) element denoted by a pair ℓm is the probability $\mu_\ell p_{\ell m}$ of a transition from state i to state j, within a single time unit, caused by a customer completing service at server R_ℓ and applying to server R_m; the term $\sum$ in each diagonal element is equal to the sum of the off-diagonal elements of the corresponding row.

This matrix structure can be usefully exploited to analyze the equilibrium probability distribution of the network by the decomposition and aggregation procedure outlined above. The matrix is partitionable into (N+1) principal submatrices (separated by plain lines on fig. 1); each principal submatrix is, apart from the diagonal element, the matrix of a network of (L-1) servers $R_1,R_2,\ldots,R_{L-1}$ with a population of N,...,N-K,...,0 customers respectively. Moreover, each submatrix is again partitionable into (N-K+1) principal submatrices (separated by dotted lines on fig. 1) which, apart from the diagonal element, correspond to a network of (L-2) servers $R_1,R_2,\ldots,R_{L-2}$, with populations (N-K),(N-K-1),...,0 respectively; and so on for as many levels of decomposition as there are servers in the network.

By transforming all these principal submatrices into aggregates, each aggregate corresponding to a subnetwork with a given population, it is thus possible to apply the above aggregation procedure, into as many levels of decomposition as desired. An additional important simplification comes

from the fact that, at the lower levels of decomposition, more and more principal submatrices (see fig. 1) are, except for their diagonal, identical to each other; *identical* stochastic aggregates can thus be constructed from these submatrices by modifying only the diagonal element to obtain row sums equal to unity; it is not difficult to see on figure 1 that, as a result of this, the equilibrium probability distribution of only (N+1) distinct aggregates needs to be evaluated at each level. This approach was thoroughly investigated in [Courtois 1972,1977] where it was also demonstrated that, by taking as many decomposition levels as there are servers, each aggregate reduces to a two-server queueing system with finite population.

This type of decomposition works for more complex types of closed models. If service distributions are not exponential, each sever can be modelled by a network of exponential stages [Baskett & al 1975] , or each state must be complemented by supplementary variables [Chandy & al 1977] indicating the remaining service requirements of customers. With these extensions, just as in multi-class networks, the number of states of each aggregate is increased, but the network keeps the same number of decomposition levels and the same numbers and types of aggregates at each level.

Thus, closed network models can be decomposed into a restricted number of aggregates which represent the stochastic behavior of a given fixed population of customers applying to a subset of the servers.

When no special precautions are taken to decompose a stochastic matrix Q and to construct the aggregates Q_I^*, the degree of approximation of the above aggregation procedure is of order ε , defined by (5) [Courtois 1975,1977] . But, in queueing network models, exact results can be obtained if the decomposition and the aggregate construction obey certain conditions which, in practice, are not too restrictive. The determination of these conditions is the object of the remaining sections.

4. EXACT AGGREGATION IN STOCHASTIC SYSTEMS

In this section we establish the necessary and sufficient conditions under which the general aggregation procedure described in the preceding section yields exact results, i.e. under which the vector z obtained by this procedure is identical to the limiting probability distribution ν of the matrix Q given by (3 bis).

Let us use β_I to denote the probability of being in any state of the I^{th} set, I = 1,...,A, of the partition of the states of Q :

$$\beta_I = \sum_{i \in I} \nu_{i_I} .$$

If we set

$$b_{i_I} = \beta_I [\beta_I^{-1} \nu_{i_I} - \nu_i^*(I)] ,$$

we have, by definition of z :

$$\nu_{i_I} - x_{i_I} = \nu_{i_I} - Z_I \nu_i^*(I)$$

$$= (\beta_I - Z_I)\nu_i^*(I) + b_{i_I} . \qquad (6)$$

A theorem proved in [Courtois 1977] (theorem 2.1 page 32) states that

$$\beta = \beta P + \varepsilon K , \qquad (7)$$

where β is the A-element vector of probabilities β_I and K is an A-vector of elements

$$k_I = \sum_{J=1}^{A} \sum_{j \in J} b_{j_J} \sum_{i \in I} c_{j_J i_I} . \qquad (8)$$

From these relations, it results :

Theorem 1 : When $\beta_I \neq 0$ for all I, a necessary and sufficient condition for $z \equiv \nu$ is that $b_{i_I} = 0$ for all I and $i \in I$.

Proof. Necessity. If $z \equiv \nu$ then by (6)

$$b_{i_I} = (Z_I - \beta_I) \nu_i^*(I) ,$$

$$= (\sum_{i \in I} z_{i_I} - \sum_{i \in I} \nu_{i_I})\nu_i^*(I) = 0, \text{ for all } i_I .$$

Sufficiency. If $b_{i_I} = 0$ for all i_I, by (8) $k_I = 0$ for all I and by (7) β is the steady-state vector of T; since T is irreducible, β must be identical to Z; then, by (6) : $\nu_{i_I} = z_{i_I}$ for all i_I.

In plain words, an aggregation procedure yields the correct equilibrium probability vector of a stochastic matrix if and only if the aggregates are constructed in such a way that, for each aggregate, the equilibrium distribution $\nu^*(I)$ is equal to the true marginal probability equilibrium distribution $\beta_I^{-1} \nu_{i_I}$, $i \in I$.

This condition is, in practice, very restrictive; it will require in general the resolution of the whole system. Consider indeed only one subsystem, say Q_{11}, which, without loss of generality, can be taken as the first one :

$$Q = \begin{bmatrix} Q_{11} & E \\ F & G \end{bmatrix}$$

where the square submatrix G comprehends all the other submatrices $Q_{22},\ldots,Q_{AA}$.

Let $[\nu(1)\chi]$ be, with the same partition, the equilibrium probability vector of Q such that

$$[\nu(1)\chi]\, Q = [\nu(1)\chi]$$

$$\nu(1)\, Q_{11} + \chi\, F = \nu(1)$$

$$\nu(1)\, E + \chi\, G = \chi \quad ;$$

if Q is irreducible, 1 is not an eigenvalue of G and $(G-I)^{-1}$ exists; thus, solving the last equation for χ and replacing χ in the second equation yields

$$\nu(1)\, [Q_{11} - E(G-I)^{-1} F] = \nu(1) \ .$$

Thus, the marginal equilibrium distribution $\nu(1)$ and 1 are also eigenvector and eigenvalue of the matrix $[Q_{11} - E(G-I)^{-1}]F$. This result is simply a direct application of the Gauss-Aitken-Bodewig formula [Bodewig 1956] which gives the algebraic expression of the condensed matrix obtained by a triangular condensation procedure.

We note that the rowsums of $E(I-G)^{-1}F$ are equal to those of E; indeed, if $\underline{1}$ denotes a column vector of element 1, we have :

$$E(I-G)^{-1}\, F\, \underline{1} = E(I-G)^{-1}(I-G)\, \underline{1} = E\, \underline{1} \ .$$

Thus, a possible construction for a correct aggregate would be

$$Q_1^* = Q_{11} + E(I-G)^{-1}\, F$$

which requires the knowledge of $(I-G)^{-1}$. Q_1^* is not the only possible correct aggregate; but, if the original matrix has no particular structure, one can expect that other possible schemes of construction of correct aggregates will require an equivalent knowledge of the system. However, the structure of queueing networks is such that correct aggregates can be obtained more simply.

5. EXACT AGGREGATION . CLOSED NETWORKS WITH PRODUCT FORM

From the properties which have been established in the two preceding sections, we can derive the conditions under which correct aggregation is feasible in multi-class closed networks which have product form, arbitrary differentiable service distributions, and fixed routing matrix.

The routing matrix of probabilities $p_{i,k;j,r}$ is supposed to be irreducible so that there is a vector $[X_{i,k}]$, defined within a multiplicative constant which satisfies the set of equations

$$\sum_{j=1}^{L} \sum_{k=1}^{K} X_{j,k}\, p_{j,k;i,r} = X_{i,r} , \qquad 1 \leq i \leq L \qquad (9)$$
$$1 \leq r \leq K .$$

$X_{i,r}$ can be interpreted as the *relative departure rate* of customers of class r from server R_i. This set of balance equations expresses that, in the equilibrium, and over any given period of time, the average number of customers of a class departing from a server is equal to the average number arriving to this server. On this definition is based the

Theorem 2 : Exact aggregation of the equilibrium probability distribution of a network with product form is possible if and only if each aggregate model of servers yields the exact values of the relative departure rates of these servers.

Proof. Consider an aggregate $\mathcal{A}$ of servers $R_1,\ldots,R_\ell$, $1 \leq \ell < L$; and let $p(s_1,\ldots,s_\ell|\mathcal{A})$ be the marginal probability that this aggregate is in state $(s_1,\ldots,s_\ell)$ given that the network is in some state $S = (s_1,\ldots,s_\ell,s_{\ell+1},\ldots,s_L)$; thus

$$p(s_1,\ldots,s_\ell|\mathcal{A}) = \sum_{s_{\ell+1},\ldots,s_L} p(s_1,\ldots,s_\ell,s_{\ell+1},\ldots,s_L) ;$$

the summation being taken over all feasible states only; since the network has product form, we have also :

$$p(s_1,\ldots,s_L) = G\, f_1(s_1) f_2(s_2) \ldots f_L(s_L)$$

where G is a normalizing constant chosen to make the probability sum equal to 1. Thus

$$p(s_1,\ldots,s_\ell|\mathcal{A}) = G\, f_1(s_1)\ldots f_\ell(s_\ell)\Big[\sum_{s_{\ell+1},\ldots,s_L} f_{\ell+1}(s_{\ell+1})\ldots f_L(s_L)\Big] ,$$

the summation being over all feasible states only; or,

$$p(s_1,\ldots s_\ell|\mathcal{A}) = G_{\mathcal{A}}\, f_1(s_1)\ldots f_\ell(s_\ell) , \qquad (10)$$

$G_{\mathcal{A}}$ being a normalizing constant for the set of states of aggregate $\mathcal{A}$. In a network with product form, each function $f_i(s_i)$ is, for a given s_i, entirely defined by the service discipline, the service time distribution and the relative departure rate X_i of server R_i [Baskett & al 1975, Chandy & al 1977]. Hence, if and only if this departure rate is exactly obtained from the aggregate model, so can be the function $f_i(s_i)$ which, in all other respects, depends only on characteristics local to the server R_i; then,

also the marginal distribution (10) can be exactly obtained for aggregate $\mathcal{A}$. By theorem 1, this exact marginal distribution for each aggregate $\mathcal{A}$ is necessary and sufficient to obtain by decomposition and aggregation the network equilibrium distribution $p(s_1,\ldots,s_L)$. This completes the proof.

Exact Aggregation Procedure. There remains the problem of constructing correct aggregates which obey the conditions of theorem 2. In the present case of queueing networks, this is an easier problem than it is in the general case of stochastic matrices discussed in section 4.

From theorem 2 we can derive a theorem which generalizes Vantilborgh's [1978] conditions for exact aggregation to all closed networks with product form and fixed routing matrix.

Define a m-element vector u as being subparallel to a vector $[v_1,\ldots,v_n]$, $n \geq m$, iff there is a scalar $k \neq 0$ such that $u = k[v_1,\ldots,v_m]$. Then, we have

Theorem 3 : Exact aggregation of the equilibrium probability distribution of a closed network with product form is possible iff each aggregate has a routing matrix whose steady-state vector is subparallel to the steady state vector of the network routing matrix.

The proof of the necessity and sufficiency of this condition results directly from theorem 2 and from the fact that server relative departure rates are defined by the steady-state vector of the network routing matrix.

Thus, a correct aggregate is the model of a closed subnetwork with a given population of customers; each server has same server discipline and same service time distribution as in the original network; the routing matrix of this correct aggregate is derived from the network routing matrix so as to obey the conditions of theorem 3.

One particular way of deriving this correct aggregate routing matrix is to use the Gauss-Aitken-Bodewig formula to obtain a *condensed routing matrix*.

Indeed, assume that the original network routing matrix P, with steady-sate vector X, X = XP, is partitioned in the following way

$$P = \begin{bmatrix} P_{\ell x \ell} & E \\ F & G \end{bmatrix}$$

where $P_{\ell x \ell}$ is the submatrix of routing probabilities connecting the servers $R_1,\ldots,R_\ell$ of a given aggregate. As shown in section 4, the Gauss-Aitken-Bodewig formula ensures that the matrix $[P_{\ell x \ell} - E(G-I)^{-1}F]$ has a steady-state vector which is subparallel to the steady-state vector X of P.

It is thus remarkable that the conditions for exact aggregation which, in general, pertain to the whole state transition matrix of a system, reduce in the case of closed queueing networks with product form to conditions over the routing matrix only. There is, in this context, an intuitive interpretation for the Gauss-Aitken-Bodewig formula. Consider a network with a single class of customers and suppose an aggregate which groups all servers of this network, except one, say R_L. The condensed routing matrix of this aggregate yielded by the Gauss-Aitken-Bodewig formula is a matrix of probabilities p'_{ij} obtained from the network routing probabilities p_{ij} by

$$p'_{ij} = p_{ij} + p_{iL}(1 - p_{LL})^{-1} p_{Lj}$$

$$= p_{ij} + p_{iL}\left(\sum_{k=1}^{\infty} p_{LL}^{k}\right)p_{Lj} ,$$

assuming that $p_{LL} \neq 1$ (otherwise the network is reducible).

This is thus equivalent to a decomposition of the network into aggregates of servers $R_1,\ldots,R_{L-1}$, with given populations on the one hand, and on the other hand a dummy server with zero service times at which a customer cycles are arbitrary number k of times before returning to one of the aggregates. The marginal distribution of customers within each aggregate is, *in relative value*, the same as it is among the corresponding servers in the whole network.

This is analogous to Norton's theorem in electrical circuit analysis, an analogy which had been first established in [Chandy & al 1975]. Theorem 3 and the Gauss-Aitken-Bodewig formula situate this analogy in the more general context of *necessary and sufficient* conditions for the obtention of the exact equilibrium vector of the *whole* network; besides, as said earlier, the Gauss-Aitken-Bodewig formula is only one particular way of constructing a condensed matrix with the required property of subparallelism. For instance, Vantilborgh [1978] has shown that for particular types of networks such as central server-, balanced- or doubly stochastic networks, the subparallelism condition is also satisfied if the aggregates are constructed so as to be of the same type as the original network.

6. GENERALIZATIONS

Theorem 2 was established only (i) for closed networks which are representable by a Markov homogeneous process and (ii) which have the product form.

Networks were assumed to be closed in order to keep a finite state space and thus a finite number of aggregates, each with a finite population and a finite number of states. But, by arguments similar to those we have invoked, it is possible (see e.g. Chandy & al 1975) to show that an open network with Poisson exogeneous inputs can be analyzed as a set of separate subnetworks, open or closed. The correct aggregates are then constructed according to the conditions of theorem 3; but the relative departure rates are now solution of the system

$$\lambda_{i,r} + \sum_{j=1}^{L} \sum_{k=1}^{K} X_{j,k} p_{j,k;i,r} = X_{i,r} , \qquad 1 \leq i \leq L , \; 1 \leq r \leq K \qquad (11)$$

where $\lambda_{i,r}$ is the rate of exogeneous arrivals of class r customers to server i.

The Markov process representation, was required because we could establish the general necessary and sufficient condition of theorem 1 for stochastic matrices only. We can presume, however, that similar condition will prevail for other types of system representations.

The last assumption of product form was needed in the proof of theorem 2 as an easy means to ensure that the exact marginal equilibrium distributions of customers in an aggregate are entirely determined, if the servers relative departure rates are known, only by characteristics which are local to the aggregate such as the service disciplines, the service time distributions and the fixed routing probabilities. Again, this is presumably a rather general property of networks; the server departure rates are also input rates to other aggregates; if all these rates are correctly evaluated in relative value, the flows of customers through the aggregates are kept proportional to what they are in the original network, so that each aggregate marginal equilibrium distribution will be correctly obtained.

It is thus plausible that theorem 2 holds for a wider class of networks than those which have the product form; at least for networks in which the departure rates can be defined.

So far, we have defined the departure rates as the steady-state solution of a first order Markov chain routing matrix. The theorem we prove hereafter shows that limiting values for these departure rates exist also in networks which do not have necessarily a fixed Markov routing matrix; e.g. in networks where the routing probabilities are functions of the congestions at the server of departure and/or the server of arrival, or even at other servers. Hence, theorem 2 presumably holds also for these networks;

unfortunately, it is essentially a theorem of existence which gives little clue as how to calculate limiting departure rates. Consequently, there is not yet for these networks an equivalent of theorem 3 specifying how correct aggregates should be constructed.

We shall define a server as being *non-preemptible* if a customer who starts being served by such a server completes service before any other service is started.

Suppose then that R_i is a non-preemptible server of a network and let $\tau_{1,i}, \tau_{2,i}, \ldots, \tau_{k,i}, \ldots$ be the epochs at which a service is completed by R_i. Consider the class of events

$$\varepsilon_i = \{n_i(t) \neq 0 \ ; \ n_j(t)=0 \text{ for all } j\neq i \ ; \ t=(\tau_{k,i}+0) \text{ for some } k\} \ ;$$

an event of this class corresponds to the situation where all customers in the network are at server R_i and one customer has just completed service at this server. After such an event ε_i, the behavior of a closed network, or of an open network with Poissonian exogeneous input, is a probabilistic replica of its behavior after the first such event ; for a given non-preemtible server R_i, the sequence of such events is thus a sequence of *regenerative events* in these networks.

Define now $D_\ell(t)$ as the number of departures from server R_ℓ, $\ell=1,\ldots,L$, since t=0 up to time t ; and let us use $t_{1,i}, t_{2,i}, \ldots, t_{n,i}, \ldots$ to denote the successive epochs at which an event of class ε_i occurs. For each $\ell=1,\ldots,L$, the process $\{D_\ell(t)\}$ is a *cumulative process* [Smith 1955, 1958] since $D_\ell(t)$ is of bounded variation in every finite t-interval and $\{D_\ell(t_{n,i})-D_\ell(t_{n-1,i})\}$, for n=1,2,..., is a sequence of independent, identically distributed non-negative random variables. If we introduce the following expectations

$$\kappa_\ell = E\{D_\ell(t_{n,i})-D_\ell(t_{n-1,i})\}$$

and

$$\alpha_i = E\{t_{n,i}-t_{n-1,i}\} \ ,$$

we have the following theorem :

Theorem 4 : If, in a closed network or in an open network with Poissonian input, there is at least one non-preemptible server R_i for which ε_i is a class of positive recurrent events ($\alpha_i < \infty$), then

$$\lim_{t\to\infty} \frac{D_\ell(t)}{t} = \lim_{t\to\infty} \frac{E\{D_\ell(t)\}}{t} = \frac{\kappa_\ell}{\alpha_i} \ .$$

Proof. The proof results directly from Smith's ergodic theorems [1955,1958] ; all conditions are satisfied for these theorems since $D_\ell(t)$ is a random variable with positive increments only and since, with non-zero service times, $\alpha_i < \infty$ implies $\kappa_\ell < \infty$.

Thus, theorem 4 provides for the cumulative departure rate of the servers a limiting value which is independent from the routing process of the network. How this result can be used in the determination of correct aggregates is still an open question ; but Smith's central limit theorems for cumulative processes may prove useful in this context.

7. ACKNOWLEDGEMENTS.

Discussions with H. Vantilborgh during which he commented on an earlier draft, were very fruitful in the preparation of this paper.

8. References.

Baskett, F., Chandy, K. M., Muntz R. R. and Palacios F. G. (1975), Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, J. ACM, 22, 2, 248-260.

Bodewig, E. (1956), Matrix Calculus. North-Holland Pub. Co., Amsterdam.

Chandy, K. M., Herzog, U. and Woo L. (1975). Parametric Analysis of Queuing Networks, IBM J., Res. Develop., 19, 1, 43-49.

Chandy, K. M., Howard Jr., J. H. and Towsley D. F. (1977), Product form and Local Balance in Queueing Networks, J. ACM , 24, 2, 250-263.

Cohen, J. W. (1977), The Multiple Phase Service Network with Generalized Processor Sharing, Rep. 69, Dpt. Mathematics, Univ. of Utrecht.

Courtois, P. J. (1972), On the near-complete decomposability of networks of queues and of stochastic models of multiprogramming computing systems, Rep. CMU-CS-72, 111, Carnegie-Mellon Univ.

Courtois, P. J. (1975), Error Analysis in Nearly Completely Decomposable Systems, Econometrica, 43, 4, 691-709.

Courtois, P. J. (1977), Decomposability, Queueing and Computer System Applications, Academic Press, New-York.

Daley, D. J. (1976), Queueing Output Processes, Adv. Appl. Prob. 8, 395-415.

Disney, R. L. and Cherry W. P., Some Topics in Queueing Network Theory, Lecture Notes in Economics and Math. Syst. Vol. 98,Springer-Verlag, 23-44.

Gelenbe, E. and Muntz, R. R. (1976), Probabilistic Models of Computer Systems - Part I (Exact Results), Acta Informatica 7, 1, 35-60.

Gordon, W. J. and Newell, G. F. (1967), Closed Queueing Systems with Exponential Servers, Operations Research, 15, 2, 254-265.

Jackson, J. R. (1963), Jobshop-like Queueing Systems, Management Sci. 10, 131-142.

Kühn, P. (1976), Analysis of Complex Queueing Networks by Decomposition, 8th Internat. Teletraffic Congress, Melbourne, 236/1-236/8.

Schassberger, R. (1977), Insensitivity of Steady-State Distributions of Generalized Semi-Markov Processes. Part I, The annals of Probability, 5, 1, 87-99.

Smith, W. L. (1955), Regenerative Stochastic Processes, Proc. Roy. Soc. A, 232, 6-31.

Smith, W. L. (1958), Renewal Theory and Its Ramifications, J. R. Statist. Soc., B, 20, 243-302.

Vantilborgh, H. (1978), Exact Aggregation in Exponential Queueing Networks, to be published in the J. ACM.