Properties and analysis of queueing network models with finite capacities

S.Balsamo
Dipartimento di Informatica
University of Pisa, Italy

Abstract
Queueing network models with finite capacity queues and blocking are used to represent systems with finite capacity resources and with resource constraints, such as production, communication and computer systems. Various blocking mechanisms have been defined in literature to represent the various behaviours of real systems with limited resources. Queueing networks with finite capacity queues and blocking, their properties and the exact and approximate analytical solution methods for their analysis are surveyed.
The class of product-form networks with finite capacities is described, including both homogeneous and non-homogeneous models, i.e., models of systems in which different resources work under either the same or different blocking mechanisms. Non-homogeneous network models can be used to represent complex systems, such as integrated computer-communication systems.
Exact solution algorithms to evaluate the passage time distribution of queueing networks with finite capacity are discussed, as well as some recent results on the arrival time queue length distribution and its relation to the random time queue length distribution. This result provides an extension of the arrival theorem to a class of product-form networks with finite capacity.
Properties of queueing network models with blocking are presented. These include insensitivity properties and equivalencies between models with and without blocking, between models with both homogeneous and non-homogeneous blocking types, and relationships between open and closed queueing network models with blocking.
Although properties of queueing networks with blocking have been mainly derived for the queue length distribution and average performance indices, we shall also present some equivalence properties in terms of passage time distribution in closed models.

1 Introduction

Queueing network models have been extensively applied to represent and analyze resource sharing systems, such as production, communication and computer systems. Queueing networks with finite capacity queues and blocking are used to represent systems with finite capacity resources and with resource constraints. Various blocking mechanisms have been defined in the literature to represent the various behaviours of real systems with limited resources.

We consider queueing network models with finite capacity queues and blocking, their properties, and analytical solution methods for their analysis.

The performance of systems with limited resources can be evaluated by considering both average performance indices, such as system throughput and resource utilization, and more detailed measures, such as queue length distribution and passage time distribution.
Most of the solutions proposed in literature concern exact or approximate evaluation

This work was partially supported by MURST and CNR Project Research Funds, by grant N. 92.00009.CT12.

of average performance indices and of the joint queue length distribution of the system in stationary conditions. Under special constraints, a product-form solution of the joint queue length distribution can be derived.
Identifying a class of product-form queueing networks with finite capacity queues is an important issue which allows efficient solution algorithms to be defined.
This paper introduces the class of product-form networks with finite capacities is introduced and the main key properties that lead to the closed form solution are discussed. This class of models has recently been extended to include both homogeneous and non-homogeneous models, i.e., models of systems in which several resources work under either the same or different blocking mechanisms. Non-homogeneous network models can be used to represent complex systems, such as integrated computer-communication systems.
A few results have been obtained for the evaluation of more detailed performance measures, such as passage time distribution and arrival queue length distribution. We survey algorithms to evaluate the passage time distribution exactly in closed network models with blocking. Some recent related results on the arrival time queue length distribution in queueing networks with finite capacity and its relationship with the random time queue length distribution are presented. This result provides an extension of the arrival theorem for a class of product-form queueing networks with finite capacity.
We briefly discuss the main approaches to approximate solution methods to evaluate performance indices of non product-form networks with finite capacity. These methods are based on the decomposition principle which is applied either to the network or to underlying Markov process. Approximate solutions with knowledge of the introduced error and bounded methods are discussed.

Finally, properties of queueing network models with blocking are presented, including insensitivity properties and equivalences between models with and without blocking, between models with both homogeneous and non-homogeneous blocking types, and relationships between open and closed queueing network models with blocking.
By using these equivalence relationships, it is possible both to extend product-form solution and solution methods defined for a given model to the corresponding one and to extend or to relate insensitivity properties for queueing networks with different blocking mechanisms.
Although properties of queueing networks with blocking have been mainly derived for the stationary joint queue length distribution and for average performance indices, we shall also present some equivalence properties in terms of passage time distribution in closed models.

This paper is organized as follows. Section 1 introduces the model and blocking type definition. In Section 2 analytical solution methods are surveyed. Exact analytical solutions are presented both for Markovian non product-form networks and for product-form networks with finite capacity in terms of different performance indices. Moreover we present an extension of the arrival theorem to this class of models. Section 3 deals with insensitivity and equivalence properties of networks with finite capacity and blocking, including equivalencies between networks with and without blocking, and between models with different blocking types. Finally, Section 4 presents the conclusions and open issues.

1.1 Queueing networks with finite capacity queues

We consider open and closed queueing networks with finite capacity queues. For the sake of simplicity we introduce model assumptions and notations for the single class

network. Multiclass queueing networks with finite capacities and their relationship with single class networks under different blocking types is discussed in [23].

Consider a queueing network formed by M finite capacity service centers (or nodes). The queueing network can be either open or closed. For a closed network, N denotes the number of customers in the network. For open networks an exogenous arrival process is defined at each node i, $1 \le i \le M$. The arrival rate can be either load independent or load dependent, denoted as λ and $\lambda\, a(n)$, $n \ge 0$, respectively, where $a(n)$ is an arbitrary non negative function of the total number of customers in the entire network. The arrival process is usually assumed to be Poisson.
An exogenous arrival tries to enter node i with probability p_{0i}, $1 \le i \le M$. In other words, the Poisson arrival process at node i has a parameter $\lambda\, p_{0i}$ for load independent arrivals, and $\lambda\, a(n)\, p_{0i}$ for load dependent arrivals.
The customers' behaviour between service centers of the network is described by the routing matrix $P = \| p_{ij} \|$, $1 \le i,j \le M$, where p_{ij} denotes the probability that a job leaving node i tries to enter node j. For open networks p_{i0} , $1 \le i \le M$ denotes the probability that a job which exits node i leaves the network. By definition, the following relation holds, for $1 \le i \le M$:

$$\sum_{j=1}^{M} p_{ij} + p_{i0} = 1$$

Let us introduce vector $\mathbf{x} = (x_1, \ldots, x_M)$ which can be obtained by solving the following linear system:

$$x_i = \lambda p_{0i} + \sum_{j=1}^{M} x_i p_{ji} \qquad (1)$$

For closed networks, by definition, $p_{0i} = p_{i0} = 0$ for $1 \le i \le M$, and system (1) does not provide a unique solution.
For queueing networks with infinite capacity queues, component x_i of the solution vector represents the throughput of node i for open networks, whereas it represents the relative throughput or mean number of visits of customers at node i for closed networks, $1 \le i \le M$ [16, 41, 47].
For queueing networks with finite capacity this meaning is not generally true.
Network routing matrix P is said to be reversible if $x_i\, p_{ij} = x_j\, p_{ji}$, and $\lambda\, p_{0i} = x_i\, p_{i0}$ for $1 \le i,j \le M$ [39].
Service center i, $1 \le i \le M$, is described by the number of servers, the service time distribution and the service discipline. Let $\mathbf{S}_i$ denote the state of node i, which includes the number of jobs in node i, denoted by n_i, and other components depending on both the node type (service discipline and service time distribution) and the blocking type.
The service time distribution of jobs at node i is denoted by $F_i(t)$, $t \ge 0$, $1 \le i \le M$, and its mean value by $1/\mu_i$ if it is load independent. The node i job service rate can be defined as dependent on the number of customers in node i, $n_i \ge 0$, and is denoted by $\mu_i f_i(n_i)$, where f_i is an arbitrary non-negative function, $1 \le i \le M$.
We consider a Markovian network, i.e., we assume that the queueing network model with finite capacity can be represented by a Markov process.
This is a common assumption in the performance analysis of computer and communication systems. From the modelling viewpoint, the queueing network model introduced above can be represented by a continuous time Markov process if service time distributions and inter-arrival time distribution in open networks have a

phase-type or coxian representation [41, 47]. This allows us to consider a large class of distributions and to represent, possibly by approximation, arbitrary distributions.

In queueing networks with finite capacity queues and blocking, additional constraints on the number of customers are included to represent different types of resource constraints in real systems, which correspond to definitions of different parameters. We consider two types of constraints, related to the capacity of a single resource and to a set of resources, respectively.
In the first case, let B_i denote the maximum number of customers admitted at node i (i.e., the maximum buffer size), $1 \le i \le M$. The total number of jobs in node i, n_i, is thus assumed to satisfy the constraint $n_i \le B_i$, $1 \le i \le M$. When the number of customers in a node reaches the finite capacity ($n_i = B_i$) the node is said to be full. Note that in multichain and multiclass networks one can also define a chain or a class dependent maximum queue length at node i.
In the second case, let B_W denote the maximum population admitted in a subnetwork W of the whole network. In certain cases, in order to represent particular system behaviours, a minimum population value L_W for subnetwork W it is also introduced. In other words, the total population in subnetwork W, $n_W = \sum_{i \in W} n_i$, is assumed to satisfy constraints $L_W \le n_W \le B_W$.

Finally, let $b_i(n_i)$ denote the blocking function, i.e., the probability that a job arriving at node i, is accepted when n_i is the state of the node, $1 \le i \le M$. For multiclass queueing networks the blocking function may also depend on the total number of jobs in the node and in the class or in the chain.
An example of a simple blocking function for single class queueing networks which allows us to define the maximum queue length B_i for each node i, $1 \le i \le M$, is defined as follows [36] :

$$b_i(n_i) = 1 \quad \text{for } 0 \le n_i < B_i, \qquad b_i(B_i) = 0 \qquad 1 \le i \le M$$

More generally, one can define

$$0 < b_i(n_i) \le 1 \quad \text{for } 0 \le n_i < B_i, \qquad b_i(B_i) = 0 \qquad 1 \le i \le M \qquad (2)$$

as an arbitrary non-negative load-dependent function which can be used to represent a flow control mechanism of node i input traffic.

1.2 Definition of blocking mechanisms

Various blocking mechanisms or types that describe different behaviours of customer arrivals at a full capacity node and the servers' activity in the network have been defined in literature. We now introduce the most commonly used five blocking mechanisms.

The first three blocking types, Blocking After Service, Blocking Before Service and Repetitive Service Blocking, have been named and classified in [5,51]. They are due to the finite capacity of service centres of the network [5,51].
The last two blocking mechanisms, Stop and Recirculate Blocking, which are very common in communication systems, have been named and compared in [64, 65]. They are due to the maximum queue length constraint for either a subnetwork or the total queueing network population.

Blocking After Service (BAS): if a job attempts to enter a full capacity queue j upon completion of a service at node i, it is forced to wait in node i server, until the destination node j can be entered. The server of source node i stops processing jobs (it

is blocked) until destination node j releases a job. Node i service will be resumed as soon as a departure occurs from node j. At that time the job waiting in node i immediately moves to node j.
If more than one node is blocked by the same node j, then a scheduling discipline must be considered to define the unblocking order of the blocked nodes when a departure occurs from node j. First Blocked First Unblocked is a possible discipline [7, 51] which states that the first node to be unblocked is the one which was first blocked.
This blocking mechanism, also called classical, transfer, manufacturing and production blocking [1, 5, 7, 9, 12, 14, 17, 26, 33, 35, 45, 49-57, 62], has been used to model production systems and disk I/O subsystems.

Blocking Before Service (BBS): a job declares its destination node j before it starts receiving service at node i. If at that time node j is full, the service at node i does not start and the server is blocked. If a destination node j becomes full during the service of a job at node i whose destination is j, node i service is interrupted and the server is blocked. The service of node i will be resumed as soon as a departure occurs from node j. The destination node of a blocked customer does not change.
Two different subcategories can be introduced [51] depending on whether the server can be used as a service centre buffer when the node is blocked:
BBS-SO (server occupied) when the server of the blocked node is used to hold a customer;
BBS-SNO (server is not occupied) when the server of the blocked node cannot be used to hold a customer. However note that BBS-SNO can only be defined for special topology networks, i.e., when the finite capacity node has only one possible sending node, i.e., if $n_i<B_i$, then there exists only one node j such that $p_{ji}>0$, and $p_{ki}=0$ for $k\neq j$, $1\leq i,j,k\leq M$.
A variant of the BBS type has been considered [9, 29, 33, 42] when the overall set of sending nodes is blocked. This variant is defined as follows :
BBS-O (Overall Blocking Before Service): when a destination node j becomes full, it blocks the service in each of its possible sending nodes i, regardless of the destination of the currently processed job. Note that a job which arrives at an empty node i cannot begin the service if one of the downstream nodes of i is full. Services will be resumed as soon as a departure occurs from node j. The destination node of a blocked customer does not change.
This blocking mechanism, also called service or immediate blocking [7, 9-11, 13-15, 17, 19, 28, 30-34, 50-56] has been used to model production, telecommunication, and computer systems.

Repetitive Service Blocking (RS): a job upon completion of its service at queue i attempts to enter destination queue j. If node j is full, the job is looped back into the sending queue i, where it receives a new independent service according to the service discipline.
Two different subcategories have been introduced depending on whether the job, after receiving a new service, chooses a new destination node independently of the one that it had selected previously:
RS-RD (random destination) if a job destination is randomly chosen at the end of each new service, whatever the previous choices;

RS-FD (fixed destination) if a job destination is determined after the first service and cannot be modified.
This blocking type, also called rejection, retransmission or repeat protocol [2-4, 8-10, 19, 27, 32, 36, 37, 40, 43, 50-56, 59, 64, 66, 68, 69] has been used to model telecommunication systems.

For the following two blocking types the population either of a subnetwork or of the total network is assumed to be in the range [L,U], where L and U are the minimum and maximum populations admitted, respectively. This constraint can be represented by an appropriate definition of both the load dependent arrival rate functions a(n) and of a (network) blocking function d(n), where $n \geq 0$ is the total network population. For multichain networks, arrival and blocking functions can also be defined for each chain, dependent on the total network population in the chain.

STOP Blocking: the service rate of each node is delayed by a factor $d(n) \geq 1$, when there is a network population $n \geq 0$. In other words, the actual job service rate of each node depends on the state n of the entire network according to function d(n). When $d(n) = 0$ then the service at each node in the network is stopped. Services will be resumed at each node as soon as an exogenous arrival occurs.
This blocking mechanism, also called delay blocking [64, 65, 67] has been used to model communication systems.

RECIRCULATE Blocking: a job upon completion of its service at queue i actually leaves the network with probability $p_{i0}\, d(n)$, when n is the total network population, whereas it is forced to stay in the network with probability $p_{i0}\, [1-d(n)]$, according to routing probabilities. Consequently, a job completing the service at node i actually enters node j with state dependent routing probability $p_{ij} + p_{i0}\, [1-d(n)]\, p_{0j}$, $1 \leq i,j \leq M$, $n \geq 0$.
This blocking type, also called triggering protocol [38, 46, 64, 65] has been used to model communication systems.

Closed queueing networks with finite capacity queues and blocking can deadlock, depending on the blocking type definition. If a deadlock occurs then either prevention techniques or detection and resolving techniques must be applied. Deadlock prevention for blocking types BAS, BBS and RS-FD is based on the condition that the overall network population N is less than the total buffer capacity of the nodes in each possible cycle in the network, whereas for RS-RD blocking it is sufficient that routing matrix P is irreducible and N is less than the total buffer capacity of the nodes in the network. Deadlock in queueing networks with blocking has been discussed in [45, 51]. Moreover, note that in order to avoid deadlocks for BAS and BBS blocking types we assume $p_{ii}=0$, $1 \leq i \leq M$.

Below we shall consider deadlock-free queueing networks in steady-state conditions.

1.3 Performance indices

The analysis and properties of queueing network models with finite capacity queues refer to a set of figures of merit of the system performance. These indices can be related to a single resource, corresponding to a service center of the queueing network, or to the overall system.

Specifically, for each resource i, $1 \le i \le M$, we consider the following average performance indices:

U_i utilization
X_i throughput
L_i mean queue length
T_i mean response time

and the following random variables whose distribution can be evaluated:

n_i number of customers in the resource
t_i customer passage time through the resource.

Random variable n_i is considered both at arbitrary times and at arrival times of a customer at the resource. The latter distribution is usually required in the evaluation of job residence time and passage time distributions in queueing networks.

Let $\pi_i(n_i)$ denote the stationary (marginal) queue length distribution of resource i, i.e., the stationary probability of n_i customers in node i at arbitrary time, $n_i \ge 0$, $1 \le i \le M$.
Let $\xi_i(n_i)$ denote the stationary queue length distribution of resource i at arrival times of a customer at that node, $n_i \ge 0$, $1 \le i \le M$.
Let $PB_i(n_i)$ denote the blocking probability of resource i, i.e., the probability that resource i is not empty and blocked by a full destination node, when there are n_i customers in node i. Let $PB_i = \sum_{n_i} PB_i(n_i)$ denote the overall blocking probability of node i, $1 \le i \le M$. The definition of these probabilities depends on the blocking type, as discussed in [9].

For open queueing networks with finite capacity another performance index of interest is the job loss probability, which can be computed by the stationary queue length distribution at arrival times.

Resource utilization and throughput in queueing networks with finite capacity depend on the blocking probability. For single server nodes this can be defined as follows:

$$U_i = 1 - \pi_i(0) - PB_i$$

$$X_i = \sum_{n_i} [\, \pi_i(n_i) - PB_i(n_i)] \, \mu_i(n_i)$$

which for constant service rate, i.e., when $\mu_i(n_i) = \mu_i$ for $n_i > 0$, reduces to

$$X_i = U_i \, \mu_i \, .$$

Mean queue length and mean response time for resource i can be computed as for queueing network models with infinite capacity queues as follows:

$$L_i = \sum_{n_i} n_i \, \pi_i(n_i)$$

$$T_i = L_i / X_i \, .$$

Performance indices of queueing networks with finite capacity can be evaluated through the analytical solution methods discussed in the next section. Insensitivity and equivalence properties of these models are expressed in terms of the performance measures and are presented in Section 3.

2 Analytical solution methods

In this section we overview analytical methods to analyze queueing network models with finite capacity queues and blocking.

Solutions have been proposed to evaluate both average performance indices and probability distribution of the number of customers in the nodes and of the passage time.
Exact solutions for the evaluation of average performance indices and of the stationary joint queue length probability distribution at arbitrary times of queueing networks with blocking have been derived in literature for different blocking mechanisms [1-4, 23, 27, 33, 36-39, 46, 50, 51, 59, 65, 66, 68-68].
Product-form solutions of the joint stationary queue length distribution have been obtained, under special constraints, for different blocking mechanisms. A survey of product-form solutions of queueing networks with blocking and equivalence properties among different blocking network models is presented in [10].
The exact evaluation of the arrival time queue length distribution in closed networks with different blocking types has been derived [11, 14] and a few results have been obtained for the passage time distribution for closed cyclic networks [11, 13].

In Section 2.1 we overview analytical solutions of Markovian networks with finite capacity in terms of average performance indices and stationary joint queue length distribution both at arbitrary and arrival times. Section 2.2 deals with product-form networks with blocking.
As regards the joint queue length distribution at arrival times, we discuss the conditions under which it can be related to the state distribution at arbitrary times for a class of non product-form networks with blocking. For the special case of product-form closed networks this result provides an extension of the arrival theorem for queueing networks with infinite capacity queues to networks with finite capacity queues and blocking.
Algorithms to evaluate the passage time distribution exactly in queueing networks with blocking are considered.

Several approximate solutions have been proposed for queueing networks with blocking, mostly to derive mean performance measures. A survey of exact and approximate methods for closed queueing networks with blocking is presented in [51]. Open queueing networks with blocking and a bibliography on networks with finite capacity queues are presented in [55, 56].
Since most of the approximation methods proposed in literature are based on the decomposition principle, in Section 2.3 we discuss the main approaches related to the decomposition applied either to the network model or to the underlying Markov process. The problem of the knowledge of the approximation error is discussed and bounded solution methods to evaluate queueing networks with blocking are considered.

2.1 Exact analysis of Markovian networks

The exact analysis of queueing networks with finite capacity and blocking concerns the evaluation of

1) mean performance indices and joint queue length distribution at arbitrary times;
2) stationary joint queue length distribution at arrival times;
3) passage time distribution.

In this section we deal with analytical solutions of Markovian networks which generally do not have a product-form solution. Product-form networks with finite capacity are considered in Section 2.2.

1) Mean performance indices and joint queue length distribution at arbitrary times

In order to evaluate the stationary joint queue length distribution at arbitrary times

and the average performance indices, the queueing network behaviour can be represented by a homogeneous continuous time Markov process *M* with discrete state space E.
The state of a queueing network with finite capacity can be defined as an M-vector $\mathbf{S}=(\mathbf{S}_1,\ldots,\mathbf{S}_M)$, where $\mathbf{S}_i$ is the state of node i which includes the number of customers in the node, n_i, $1\le i\le M$. The state space E of the network is the set of all feasible states. Queueing network evolution can be represented by a continuous time ergodic Markov chain *M* with discrete state space E and transition rate matrix Q. The stationary and transient behaviour of the network can be analyzed by the underlying Markov process.
Under the hypothesis of an irreducible routing matrix P, there exists a unique steady-state queue length probability distribution $\pi = \{\pi(\mathbf{S}), \mathbf{S}\in E\}$, which can be obtained by solving the following homogeneous linear system of the global balance equations [41]:

$$\pi\, Q = \mathbf{0} \qquad (3)$$

subject to the normalising condition $\Sigma_{\mathbf{S}\in E}\, \pi(\mathbf{S}) = 1$ and where $\mathbf{0}$ is the all zero vector.
The definition of state space E and transition rate matrix Q depends on the network characteristics and on the blocking type of each node.
For example, for an open exponential network with Poisson load independent arrivals, where each node i has finite capacity B_i and works under the RS-RD blocking type, the state of node i can be simply defined as $\mathbf{S}_i = n_i$, $1\le i\le M$, the state space is given by

$$E = \{(n_1,n_2,\ldots,n_M) \mid 0\le n_i\le B_i\ ,\ 1\le i\le M\}$$

and the transition rate matrix is defined as follows:
$Q = \|q(\mathbf{S},\mathbf{S}')\|$, for $\mathbf{S},\mathbf{S}'\in E$ and

$$q(\mathbf{S},\mathbf{S}') = \delta(n_j)\,\mu_j\, b_i(n_i)\, p_{ji} \qquad \text{if } \mathbf{S}' = \mathbf{S} + \mathbf{e}_i - \mathbf{e}_j$$

$$q(\mathbf{S},\mathbf{S}') = \delta(n_j)\,\mu_j\, p_{j0} \qquad \text{if } \mathbf{S}' = \mathbf{S} - \mathbf{e}_j$$

$$q(\mathbf{S},\mathbf{S}') = \lambda\, p_{0j}\, b_j(n_j) \qquad \text{if } \mathbf{S}' = \mathbf{S} + \mathbf{e}_j$$

$$q(\mathbf{S},\mathbf{S}) = -\,\Sigma_{\mathbf{S}'\in E,\ \mathbf{S}'\neq\mathbf{S}}\ q(\mathbf{S},\mathbf{S}')$$

where blocking functions $b_i(n_i)$ are given by formula (2), $\delta(n_i)$ is the following function: $\delta(n_i)=0$ if $n_i=0$, $\delta(n_i)=1$ otherwise, $1\le i\le M$, and $\mathbf{e}_i$ denotes the M-vector with all zero components except one in i-th position.

Note that system state S definition depends on the network characteristics and on the blocking type. Hence the state of node i definition may be more complex than the example above, including information such as the state of the server (whether it is active or blocked) and, for the BAS blocking type, the description of the set of nodes that are blocked by the finite capacity resource and the unblocking scheduling. A detailed definition of the system state for each blocking type is given in the Appendix.

An exact solution algorithm of queueing networks with finite capacity and blocking based on the Markov process approach can be summarised as follows:

1) Definition of the appropriate system state depending on the network characteristics (i.e., service time distributions, arrival distribution, service disciplines, network dimensions), and on the blocking type.
   Definition of the system state space E.
2) Definition of the transition rate matrix Q which describes the queueing network evolution, according to the blocking type of each node.

3) Solution of linear system (3) to derive the stationary state distribution π at arbitrary times.
4) Computation from the solution vector π of the joint and marginal queue length distributions and of the average performance indices, such as throughput, utilization and mean response time, for each resource i of the network, $1 \le i \le M$.

Note that state space E is finite for closed and open networks where all the nodes have finite capacity. For open networks which include at least one infinite capacity queue state space, E is infinite and the solution of the linear system (3) has to be approximated numerically.

Although the joint queue length distribution of the queueing network with finite capacity can be obtained by solving linear system (3) and the average performance measures can be derived from π, this approach becomes unfeasible as the state space E dimension grows, proportionally to the dimension of the model, i.e., the number of customers, nodes and chains.
Consequently for non product-form networks, approximation methods have to be considered.
Nevertheless, under certain constraints, which depend both on the network definition and the blocking mechanism, π has a product-form solution, as discussed in Section 2.2. Hence, steps 2 and 3 of the algorithm above can be substituted by the direct evaluation of the closed form solution for which computationally efficient exact solution algorithms can be defined.

Remark. A special case in which the computation step 3 can be drastically reduced concerns the so-called symmetrical networks. This class of networks was introduced in [29] and is defined as having the same blocking type, service rate and buffer capacity for each node. The routing probabilities out of each node are the same, and routing matrix P can be rewritten so that all rows are identical up to a rotation of the entries.
For symmetrical closed exponential networks with BBS-SO, BBS-O and BAS blocking types, a reduction technique has been introduced to efficiently compute solution π and average performance indices [29, 51]. Note that the reduction algorithm is related to the exact aggregation procedure applied to the Markov process, which can be easily computed due to the special characteristics of the network.

2) Arrival time distribution

When the performance index of interest is the joint queue length distribution at input times at a given node, a different solution method has to be applied, based on a new homogeneous discrete time Markov process M^e embedded in process M. Let A denote the discrete state space of M^e and $\mathbf{S}^e$ the system state as seen by an arriving job at input time at node i, where the state does not include the arriving job. Informally, each state $\mathbf{S}^e$ of the embedded process M^e is identical, except for one less job at node i, to a corresponding state of the process M just after the customer transition to node i denoted by $\mathbf{S}^a$. As in process M, also the state space definition of process M^e depends on the blocking type.
If the embedded Markov chain M^e is irreducible and recurrent then there exists the stationary state distribution ξ at arrival instants at node i. The direct evaluation of this distribution is not trivial. However, for a class of networks with finite capacity one can derive an expression of ξ in terms of the stationary state distribution at arbitrary times, π. By applying this result to some special cases of product-form networks with blocking, an extension of the arrival theorem for queueing networks

with infinite capacity queues to networks with finite capacity queues and blocking can be derived, as discussed in Section 2.2.
For Markovian non-product-form networks the following relationship between stationary distributions ξ and π for the same network holds. The following theorem has been proved for closed exponential networks with a general routing topology and blocking types BAS, BBS-SO and RS-RD [14].

Theorem 1
The stationary state probability distributions ξ and π of a closed exponential network with finite capacity queues and blocking of type BAS, BBS-SO or RS-RD, are related as follows:

$$\xi(\mathbf{S}^e) = \frac{1}{\eta} \sum_{\mathbf{S} \in I(\mathbf{S}^a)} \pi(\mathbf{S}) q(\mathbf{S}, \mathbf{S}^a)$$

where $\mathbf{S}^e \in A$, $\mathbf{S}^a \in E$ is the state corresponding to $\mathbf{S}^e$ and, according to the blocking type:

$I(\mathbf{S}^a)$	is the set of initial states of process M which occur just before a customer transition which leads to state $\mathbf{S}^a$,
$q(\mathbf{S},\mathbf{S}^a)$	is the transition rate from state $\mathbf{S}$ to $\mathbf{S}^a$ of process M and
η	is a normalising constant.

The proof of the theorem and the detailed definition of set $I(\mathbf{S}^a)$ and transition rates $q(\mathbf{S},\mathbf{S}^a)$ is given in [14]. When the finite capacity node i has only one upstream or sending node, say k (i.e., $p_{ki}>0$ and $p_{ji}=0$, $j \neq k$, $1 \leq j \leq M$) then the set $I(\mathbf{S}^a)$ is formed by a single state $\mathbf{S}=\mathbf{S}^a + \mathbf{e}_k - \mathbf{e}_i$ and, for blocking type RS and BBS-SO, this relationship can be simplified as follows [14]:

Corollary 1
If the node has only one sending node, then

$$\xi(\mathbf{S}^e) = \frac{1}{\eta} \pi(\mathbf{S}) \tag{4}$$

where $\mathbf{S}^e \in A$, $I(\mathbf{S}^a)=\{\mathbf{S}\}$ and η is a normalising constant.

This result is an extension of a similar relationship for queueing networks with infinite capacity to networks with finite capacity queues. As a consequence, the evaluation of the steady-state probability distribution at arrival times ξ can be reduced to the evaluation of the probability distribution at arbitrary times π. This arrival time distribution can be used in the analysis of job passage time distributions in the network. Moreover the theorem can be simplified for a class of product-form networks with finite capacity, leading to an extension of the arrival theorem for queueing networks with infinite capacity queues to networks with finite capacity and certain blocking types, as discussed in Section 2.2.

3) Passage time distribution

The time spent by a customer in the entire system or in a subsystem (the passage time) is an important performance measure which provides a more detailed performance evaluation of system behaviour than the average indices. The passage time distribution in queueing networks is generally difficult to obtain even for queueing networks with infinite capacity; a survey of sojourn time results in queueing networks is presented in [18]. For queueing networks with finite capacity queues and blocking, a few results have been obtained in terms of cycle time distribution for cyclic models [11-13] and for central server model or star topology networks [15].

A recursive algorithm to derive the cycle time distribution for cyclic closed exponential queueing networks with M≥2 finite capacity nodes and BBS-SO blocking is defined in [11, 13] and for M=2 nodes and BAS in [12]. The method is based on the definition of a transient Markov process which describes the evolution of a specific (tagged) customer in a complete walk through the network. Sets E_0 and F of the possible initial and final states of the network are defined, corresponding to the beginning and the end of the walk of the tagged customer, respectively.
The cycle time distribution is computed by evaluating the first hitting time probability distribution of the Markov process to the final states, starting from the initial states.
Let Z denote a state of the transient Markov process and let T(Z,s) denote the Laplace-Stieltjes transform (LST) of the passage time distribution from Z to the final states F.

The LST of the cycle time distribution, denoted by T(s) can be computed as follows:

$$T(s) = \sum_{Z \in E_0} Prob(Z)\, T(Z,s) \qquad (5)$$

where *Prob*(Z) is the probability of state Z at the cycle starting time and T(Z,s) can be computed by a recursive scheme. This recursive scheme can be reduced by taking into account the process structure and the blocking type definition, as described in [11, 13].
Since each state Z corresponds to a system state $S^e \in A$ of the embedded process M^e, as previously introduced to define the joint queue length distribution, then *Prob*(Z) can be evaluated as the stationary distribution at arrival times, $\xi(S^e)$ and, by applying Theorem 1, as a function of the stationary distribution at arbitrary times, $\pi(S)$, $S \in E$.

From the recursive scheme to evaluate the LST of the cycle time distribution one can derive an explicit expression for the cycle time distribution in the time domain, where coefficients are defined by recursion.
For a two-node exponential network with BBS-SO or BAS blocking this approach leads to the following closed-form expression in the time domain of the density function $f(t)$ of the cycle time:

$$f(t) = \sum_{j=1}^{3} e^{-\mu_j t} \sum_{i=1}^{k_j} c_{ji} \frac{t^{i-1}}{(i-1)!}$$

where $k_1=k_2=N$, $k_3=2N-3$, $\mu_3=\mu_1+\mu_2$ and coefficients c_{ij} are recursively computed for each i and j [11].
Note that for the special case of a two-node network with blocking, the stationary distribution π has a product-form solution and consequently ξ and *Prob*(Z) have a product-form solution. In this case an extension of the arrival theorem holds, as proved in [11, 12] and discussed in Section 2.2.
However, note that the algorithm sketched above to evaluate the passage time distribution applies to any Markovian non product-form network. For the class of product-form networks with finite capacity, the advantage consists in a more efficient computation of distribution π and consequently of *Prob*(Z) in formula (5).

In many practical applications it is sufficient to evaluate the first few moments of the cycle time. A recursive evaluation of the cycle time moments can be derived for cyclic closed exponential networks [11, 13] and for central server model networks [15].
For a two node exponential network with BBS-SO or BAS blocking, the k-th moment of the cycle time distribution, $E(k)$, for k=1,2,..., is given by [11, 12]:

$$E(k) = \sum_{j=1}^{3} \sum_{i=1}^{k_j+1} \frac{c_{ji}}{\mu_j^{i+k}} \frac{(i+k-1)!}{(i-1)!}$$

The evaluation of the passage time distribution in other classes of queueing network models with finite capacity, including different types of blocking and non-exponential service time distribution is an open issue.

2.2 Product-form networks

In this section we survey the class of product-form queueing networks with finite capacity and blocking. This class is a subset of the class of Markovian networks with finite capacity considered in the previous section, and hence the same analytical techniques can be applied to solve these networks.
However, an important consequence of the identification of the class of product-form networks is the definition or extension of efficient algorithms to evaluate performance indices. Specifically, one could extend basic algorithms for the class of product-form BCMP networks with infinite capacity queues [16], such as MVA and Convolution algorithms [47] to queueing networks with finite capacity queues.

First we summarise the cases of product-form networks with finite capacity and different blocking types. The extension of the arrival theorem to some cases of this class of network is then discussed.

• *Product-form solutions of the joint queue length distribution*

Product-form solutions of the joint queue length distribution π for single class open or closed networks under certain constraints, depending both on the network definition and the blocking mechanism, can be defined as follows:

$$\pi(S) = \frac{1}{G} V(n) \prod_{i=1}^{M} g_i(n_i) \qquad (6)$$

where G is a normalising constant, and n is the total network population. The functions V and g_i ,$1 \le i \le M$, are defined in terms of network parameters which include vector **x** defined by system (1) and service rates μ_i ,$1 \le i \le M$, and depend on the blocking type and additional constraints.
Similarly, for multichain open, closed or mixed queueing networks with blocking, formed by M nodes and R chains, product-form solutions can be defined as follows:

$$\pi(S) = \frac{1}{G} V_r(m_r) \prod_{i=1}^{M} g_i(n_i)$$

where G is a normalising constant, m_r is the total network population in chain r, $1 \le r \le R$, and the functions V_r , $1 \le r \le R$, and g_i, $1 \le i \le M$, are defined in terms of network parameters.

Table I summarises product-form networks with finite capacity and different blocking types.
Both homogeneous networks, where each node works under the same blocking type, and non-homogeneous ones, where different nodes work under different blocking mechanisms, are considered for five topologies.
The first three topologies concern closed networks and are the two-node network, the cyclic topology and the central server or star topology. For the central server topology networks, node 1 denotes the central node, i.e., routing matrix P is defined as follows: $p_{ij}>0$ for $i=1$, $2 \le j \le N$, $p_{i1}=1$ for $2 \le i \le N$ and $p_{ij}=0$ otherwise, $1 \le i,j \le N$.

The fourth case refers to queueing networks with reversible routing matrix P, as defined in the Section 2.1. The latter is the arbitrary topology network.
Table I shows the cases of product-form solution together with additional constraints, for each combination of blocking type and network topology, where:

- PFi denotes the corresponding product-form formula, $1 \le i \le 8$, defined below for i=3 and 6 and in Table II for all the other cases;
- an arrow denotes that the case is included in the more general class of arbitrary topology networks and, as far as we are aware, there are no special results which only hold for that specific topology;
- 'NO' means that, as far as we are aware, no product-form has been proved;
- 'NA' means that the blocking type is not applicable to the network topology;
- for non-homogeneous networks the allowed combination of blocking types is also given.

Some additional conditions are required in some cases:
Let $B = \sum_{i=1}^{M} B_i$ denote the total capacity of the network and let $B_{min} = \min \{ B_j , 1 \le j \le M \}$.
The *non-empty condition* for closed networks requires that at most one node can be empty, i.e., $N \ge B - B_{min}$.
This condition is said to be strictly verified when each node can never be empty, i.e., if the inequality strictly holds.
The *condition* which requires *at most one blocked node* is satisfied if $N = B_{min} + 1$.
In other words, if a node is full then at most one of its sending nodes is not empty and can be blocked.

Condition (A) refers to a particular model introduced in [4] of multiclass networks with parallel queues with interdependent blocking functions and service rates, and which satisfy a so-called invariant condition. See [4] for further details.

Condition (B) requires that each node i with finite capacity is the only destination node for each upstream node, i.e., it satisfies the following constraint:
if $p_{ji} > 0$ then $p_{ji} = 1$, $1 \le j \le M$.

To keep the presentation simple we only present formulas of product-form solutions for single class networks; the detailed expression of functions V_r, $1 \le r \le R$, and g_i, $1 \le i \le M$, in product-form solutions for multiclass networks is given in [10].
Table II shows the definitions of product-forms PFi, for i=1,2,4,5,7 and 8, in terms of conditions on the network model and expressions for functions V and g_i, $1 \le i \le M$, in product-form (6) for single class networks. In Table II, for product-form PF4, A-type nodes are defined as follows:

Definition. A node is said to be A-type if it has an arbitrary service time distribution and a symmetric scheduling discipline or exponential service times distributions, which are the same for each class at the same node, when the scheduling is arbitrary.

We shall now define the product-form solutions PF3 and PF6.

PF3:
Conditions:

- multiclass central server networks with the class type of a job fixed in the system,
- state-dependent routing depending on the class type,
- blocking functions dependent on node and class,
- A-type nodes.

blocking type	network topology				
	two-node	cyclic	central server	reversible routing	arbitrary
BAS	PF1	→	→	→	PF7 at most one blocked node
BBS-SNO	PF1 if $N \le B_1+B_2-2$	NO	NO	NO	NO
BBS-SO	PF1	PF2 non-empty condition	PF3 if only $B_1<\infty$	→	PF2 strictly non-empty cond. and cond. (B)
RS-RD	PF1	PF2 non-empty condition	PF3	PF4 PF6 and cond. (A)	PF2 strictly non-empty cond.
RS-FD	PF1	PF2 non-empty condition	PF3 if only $B_1<\infty$	→	PF2 strictly non-empty cond. and cond. (B)
Stop	PF5	NO	PF5	PF5	PF8
Recirculate	NA	NA	NA	→	PF8
Non-Homogeneous	BAS BBS-SO RS-RD RS-FD PF1	BBS-SO RS-RD RS-FD non-empty condition PF2	BBS-SO RS-RD RS-FD node 1 with RS PF3	RS-RD Stop PF5	BBS-SO RS-RD, RS-FD strictly non-empty cond. and cond. (B) for BBS-SO and RS-FD PF2

Table I - Product-form networks with blocking.

For single class exponential networks with load dependent service rates $\mu_i(n_i)=\mu_i f_i(n_i)$ and state-dependent routing $p_{1j}(n_j)= w_j(n_j)\, w(N-n_1)\ \forall n_j$, $p_{j1}=1$ for $2\le j\le N$, where N is the number of customers in the network, product form (6) holds with

$$V(N) = \prod_{l=1}^{N-n_1} w(l-1) \prod_{j=2}^{M} \prod_{l=1}^{n_j} w_j(l-1), \quad g_i(n_i) = \prod_{l=1}^{n_i} \frac{1}{\mu_i} \frac{b_i(l-1)}{f_i(l)} \ ,\forall n_i \ , \ 1\le i\le M$$

PF6:
Conditions:

- multiclass networks with the class type of a job fixed in the system,

	Conditions	V(n)	$g_i(n_i)$, $\forall n_i$, $1 \le i \le M$
PF1	multiclass networks BCMP type nodes class independent capacities	1	$(x_i / \mu_i)^{n_i}$
PF5	like PF4, but single class		
PF7	multiclass networks FCFS-exponential nodes class independent capacities		
PF8	multiclass open Jackson networks with class type fixed		
PF4	multiclass networks with class type fixed blocking functions dep. on node, class and chain A-type nodes load dependent service rates $\mu_i(n_i)=\mu_i f_i(n_i)$,	1	$(x_i / \mu_i)^{n_i} \prod_{l=1}^{n_i} \frac{b_i(l-1)}{f_i(l)}$
PF2	single class networks exponential nodes load independent service rates with $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_M)$ $\varepsilon = \varepsilon \mathbf{P'}$ $\mathbf{P'} = \| p'_{ij} \|$, $p'_{ij}=\mu_j p_{ji}$, $i \ne j$, $p'_{ii}=1-\sum_{j \ne i} p'_{ji}$, $1 \le i,j \le M$	1	$1 / \varepsilon_i^{n_i}$

Table II - Product-form formulas and conditions.

- interdependent blocking probability and service rates [4],
- A-type nodes.

Product-form (6) holds with

$$V(n) = 1,\ n \ge 0,\ g_i(n_i) = (x_i)^{n_i}\ h_i(\mu_i, n_i),\ n_i \ge 0,\ 1 \le i \le M,$$

where $h_i(\mu_i, n_i)$ is a product-form function dependent on the state and the service rate of node i and defined according to the scheduling of the interdependent parallel queues; for a complete definition see [4].

Product-form solutions can be proved by substituting the closed-form expression into the global balance equations of the underlying Markov process (linear system (3)).
Note that product-form expression (6) generalizes the closed-form expression for BCMP networks, and in certain cases, such as PFi, i=1,5,7 and 8, it corresponds to the same solution as the one for queueing networks with infinite capacity queues computed on the truncated state space of the network with finite capacities.
This relationship provides the basis for the equivalence between product-form networks with and without blocking discussed in Section 2.3.

Observations
Identification of necessary and sufficient conditions under which a queueing network model has a product-form solution is an open issue for networks with infinite capacity queues as well.
We observe that most of the product-form solutions for queueing networks with finite capacity queues have been derived by using the following properties:

- reversibility of the underlying Markov process,
- duality.

The first approach can be applied to networks with finite capacity whose underlying Markov process is shown to be obtained by truncating the reversible Markov process of the network with infinite capacity. Hence, a product-form solution immediately follows from the theorem for truncated Markov processes of reversible Markov processes. This theorem states that the truncated process shows the same equilibrium distribution as the whole process normalised on the truncated sub-space [39].
A product-form solution of both homogeneous and non-homogeneous two-node cyclic networks can be proved by using this property for exponential single class networks [1, 36, 40, 59] and for multiclass networks with BCMP nodes under additional constraints in [24, 50, 66].
Similarly, it has been proved that closed queueing networks with a reversible routing matrix P have a reversible underlying Markov process under RS-RD or Stop blocking and different types of nodes [36, 40, 59]. This class of product-form networks with RS-RD blocking for multiclass networks has been extended [3, 50, 68, 69] to include A-type nodes and more general blocking functions which may depend both on the total population, class population and routing chain population at the node.
The central server or star topology network is a special case of product-form networks with reversible routing. However, some product-form results have only been specifically proved for central server networks [2, 27, 44, 63, 68].

Remark. Although some of these results concern networks where routing probabilities are dependent on the state of the network, they are related to queueing networks with finite capacity and blocking. In fact, by using blocking functions, the actual routing probabilities of queueing networks with finite capacity can be interpreted as state dependent probabilities, and they are obtained by combining the routing probability matrix P with blocking functions $b_i(n_i)$, $1 \le i \le M$. Therefore, product-form solutions for networks with state dependent routing, such as the one proved in [63], can be interpreted in the same way as for blocking networks, as discussed and extended in [44] and [68] to multiclass networks. A generalization of these results obtained by combining state dependent routing and finite capacity queues is presented in [2].

Routing reversibility [36, 37, 39, 40, 59] which leads to the Markov process reversibility is related to the job-local-balance of the underlying Markov process introduced in [37]. This balance property, which is related to local balance and station balance for queueing networks with infinite capacity [16, 20, 21, 24, 40, 46], states that the rate outside a state due to any particular job in the system is equal to the rate inside that state which is due to that particular job [37]. Job-local-balance provides the basis for deriving equivalence and insensitivity properties, as discussed in Section 3.

The second approach to derive product-form solution of queueing networks with blocking is based on duality. Product-form PF2 has been obtained by adding the capacity constraint to the Gordon-Newell closed exponential networks and by defining a dual network which has the same stationary joint queue length distribution [33].

Consider a cyclic closed network with M nodes, N customers, node capacities B_i, $1 \le i \le M$ and BBS-SO or RS blocking. The dual network is obtained from the original one by reversing the connections between the nodes. It is formed by M centers and (B - N) customers which correspond to the 'holes' of the original (primal) network. When a customer moves from node i in the original network, a hole moves backward to node i in the dual one. When there are n_i customers in node i the original networks, the i-th center of the dual one contains $B_i - n_i$ holes, $1 \le i \le M$. It can be shown that the underlying Markov process which describes the evolution of customers in the network is equivalent to the one which describes the behaviour of holes in the dual network [33]. As a consequence, when the non-empty condition is satisfied, then the total number of holes in the dual network cannot exceed the minimum capacity, i.e., $(B - N) \le B_{min}$, and the dual network has a product-form solution like a network without blocking. Hence the product-form solution for the primal network is given by formula (6) with $V(N)=1$ and $g_i(n_i) = (1/\mu_{i-1})^{n_i}$, $1 \le i \le M$, (where if i=1 then i-1=M) [33], which corresponds to expression PF2. This solution can be extended to arbitrary topology networks with load independent service rates for RS-RD blocking [36]. This result has been extended to homogeneous networks with BBS-O blocking under condition (B) and to heterogeneous networks [9].
The concept of duality introduced in [33] has been applied to closed cyclic networks with phase-type service distributions and BBS-SO blocking for which the throughput of the network is shown to be symmetric with respect to its population [28].

• *Arrival theorem for product-form queueing networks with blocking*

The arrival theorem for product-form networks with infinite capacity [48, 60] provides the basic principle for the MVA computational algorithm. It states that the stationary state distribution at arrival instants of a customer at a particular node is equal to the stationary state distribution at arbitrary times of the same network, for open networks, and of the network with one less job, for closed networks. This result can also be applied for an efficient computation of the stationary state distribution at arrival times in the evaluation of passage time distribution, as discussed in Section 2.1.
Since the proof of the arrival theorem [48, 60] is based on the BCMP product-form solution [16] which does not allow blocking due to the finite capacity of the queues, the direct application of the arrival theorem to queueing networks with finite capacities does not hold. For example, the direct application of the arrival theorem to a product-form network with Stop protocol is shown to fail, as discussed in [67].
An extension of the arrival theorem has been proved for a special class of networks in which a particular type of blocking can be defined by using the 'loss' and 'trigger' functions, which allow a constraint on the overall network population of a chain in multichain queueing networks with infinite capacity queues [60] and is related to Recirculate blocking. A similar case is considered in [67].
A recent result related to product-form queueing networks with blocking is the extension of the arrival theorem to some finite capacity networks with either BBS-SO, BAS and RS-RD blocking [11, 12, 14].
In fact, from corollary 1 one can derive a relationship between the joint queue length distribution at arrival and arbitrary times of networks with different parameters for some closed exponential networks under BBS-SO, BAS and RS-RD blocking [11, 13, 14].

Consider closed networks with either a cyclic or central server topology. Let W denote the network model introduced above and let W^* denote a new network identical to W except for one less customer, and modified finite capacities denoted by B_j^*,

$1 \le j \le M$. Let π^* denote the steady-state probability distribution at arbitrary times of network W*. One can prove the following theorem [11, 14].

Theorem 2
The stationary state distribution at arrival instants at node i of network W is identical to the state distribution at arbitrary times of network W*, i.e., $\forall \mathbf{S}^e \in A$

$$\xi(\mathbf{S}^e) = \pi^*(\mathbf{S}^e)$$

i) for product-form networks with RS-RD or BBS-SO blocking and
for a cyclic topology with $M \ge 2$ nodes and
$B_j^* = B_j - 1$, for $j = i, i-1$ and $B_j^* = B_j$ for $j \ne i, i-1$, $1 \le j \le M$, $1 \le i \le M$,
and for a central server topology with
$B_j^* = B_j - 1$, for $j = 1, i$ and $B_j^* = B_j$ for $j \ne 1, i$, $1 \le j \le M$, $2 \le i \le M$,
where 1 denotes the central node;
ii) for the two-node product-form network with BAS blocking and $B_j^* = B_j - 1$, for $j = 1, 2$.

The extension of the arrival theorem to queueing network models with a more general topology and different blocking types is an open issue.

2.3 Approximate analysis

Many approximate solution methods to analyze queueing network models with finite capacity queues have been proposed in literature both for open and closed networks.
In this section we consider approximate solution techniques to solve queueing networks with finite capacity. We discuss the basic ideas and principles on which the approximations are based and the main results.

Approximate solution techniques have been proposed to evaluate the joint queue length distribution at arbitrary times and average performance indices, such as resource throughput and utilization [5, 51, 55].
Most of the approximations do not provide any bound on the introduced error and they are validated by comparing numerical results against either simulation results or exact solutions if the state space is small enough.
These approximate techniques are heuristics which are mainly based on:

- the decomposition principle applied to the underlying Markov process,
- the decomposition principle applied to the network,
- the forced solution of a non-blocking (product-form) network,
- special structural properties of a specific class of networks.

The decomposition principle applied to the Markov process consists in the identification of a partition of the state space E into K subsets E_k, $1 \le k \le K$, which leads to a decomposition of the rate matrix Q into K^2 submatrices. By referring to a decomposition-aggregation procedure, the solution of the entire system (3) is reduced to the solution of K subsystems of smaller dimensions, each related to a subset of E. These solutions are then combined to obtain the solution of the overall system.
This approach is based on the following relationship between the stationary state probability $\pi(\mathbf{S})$, the conditional probability of state $\mathbf{S}$ in E_k, $Prob(\mathbf{S} \mid E_k)$, and the aggregate probability of the subset E_k, $Prob(E_k)$, $\forall \mathbf{S} \in E_k$, $1 \le k \le K$:

$$\pi(\mathbf{S}) = Prob(\mathbf{S} \mid E_k)\, Prob(E_k)$$

Instead of the direct computation of $\pi(\mathbf{S})$, the decomposition technique requires the computation of $Prob(\mathbf{S} \mid E_k)$ and $Prob(E_k)$ for each $\mathbf{S}$ and E_k.

Unfortunately the exact computation of the decomposition-aggregation approach for Markov processes is comparable to the cost of solving the entire model and so soon becomes computationally intractable. However, exact aggregation can be performed efficiently for some classes of Markovian models such as symmetric networks.
Approximate solutions based on the decomposition of the Markov process provide an approximate evaluation of the conditional and aggregate probabilities $Prob(\mathbf{S} \mid E_k)$ and $Prob(E_k)$. Heuristics are defined by taking into account both the network model characteristics and the blocking type [5, 7, 17, 26, 27, 29, 31, 33, 35, 42, 43, 51, 52, 55-58, 62, 68, 69].
An important issue is the identification of an appropriate state space partition which affect both the accuracy and the time computational complexity of the approximate algorithm.
When the state space partition is related to a network partition into subnetworks, then the decomposition principle is applied to the queueing network and the subsystems can be solved in terms of solving (possibly modified) subnetworks in the original network. Various approaches have been proposed to determine the parameters of each subnetwork [26, 30, 31, 33, 43, 57, 58, 62, 68, 69].

Approximation methods are often based on the forced application of the exact aggregation technique to queueing networks with blocking for product-form queueing networks with infinite capacity [22]. This approach has a low computational cost and the accuracy observed by experimental results makes such approximate aggregation techniques suitable for many practical cases. However, the error introduced by the approximation is unknown.
Many approximation methods based on the decomposition approach require the iterative solution of subsystems or subnetworks to derive the approximate solution. Hence for such techniques, conditions and the speed of convergence should also be considered, as in [26].

Although few approximate solution techniques with known accuracy have been proposed, this is still an important issue which should be considered in the definition of approximations.
Another issue concerns bound solutions which can be used as approximate solution methods with known accuracy. A bounded aggregation technique has been defined for Markov processes and applied to queueing networks with blocking in [25] by exploiting the special structure of the underlying Markov process. Extending this work to more general classes of networks with finite capacities and different blocking mechanisms are challenging issues which are still open.

3 Properties of queueing networks with blocking

In this section we discuss some properties of queueing networks with finite capacity and blocking, which arise from the comparison of different models.
We consider insensitivity and equivalence properties in queueing network with blocking.
Insensitivity concerns how the characteristics of the service requirements affect the network performance.
Equivalence properties are the basis of problem reducibility. Equivalencies include both identity and reducibility relationships and can be defined between networks with and without blocking, between both homogeneous and non-homogeneous networks with different blocking types, and between open and closed networks.
Note that identifying these equivalencies depends on the performance indices involved.

Insensitivity and equivalence properties provide the basis for comparing the performance of system models with different parameters and with different blocking mechanisms.
These results can be applied, for example, in the study of the impact of the blocking type on system performance, by referring to a given set of performance indices and network parameters.
Another important consequence of these properties is that solution methods and algorithms already defined for a certain class of networks could be extended to other classes of network models with different blocking types and/or network parameters. For example, equivalence between networks with and without blocking immediately leads to the extension of efficient computational solution algorithms defined for BCMP networks such as MVA and Convolution algorithm to queueing networks with finite capacity queues.

3.1 Insensitivity

Insensitivity is the property which states that stationary characteristics of the stochastic process underlying the queueing network depends on the service requirements only in terms of their averages. Product-form queueing networks without blocking have been proved to be insensitive [16], i.e., the stationary joint queue length has been proved to depend on the service time distributions only in terms of their means.
Insensitivity can be extended to a certain class of queueing networks with finite capacity and blocking.
Referring to product-form networks with finite capacity, Tables I and II and product-form definitions show the cases where the stationary state distribution at arbitrary times depends on the distribution of the service time only in terms of the mean value (or the service rate μ_i).
Specifically, this insensitivity property holds for product form solutions PF1 which allow BCMP nodes, and for product-form PFi, i=3,4,5 and 6 which allow A-type nodes, as defined in Section 2.2.
Insensitivity for a two-node network with multiple class and RS blocking has been shown both for the joint stationary state distribution and for the call congestion of a job, i.e., the stationary probability that a job is blocked when requesting service at the next node [66].
Insensitivity of the joint queue length distribution for the central server and for reversible routing networks with A-type nodes and RS-RD and Stop blocking types has been discussed in [2-4, 50, 64, 68, 69].

3.2 Equivalence properties

Equivalence can be defined by referring to different performance indices. Most of the equivalence properties have been defined in terms of identity of the underlying Markov process of the queueing networks, which leads to an identical solution of the state probability vector π obtained by system (3).
However, note that network state **S** definition depends on the blocking type, as discussed in Section 2. Therefore although a bijective function between two state spaces of two networks can be identified such that the Markov process are identical, the meaning of corresponding states may be different and hence performance measures may be not equivalent. Moreover the identity of the joint queue length distribution between two networks does not necessarily imply that mean performance indices are identical as well.

Network with RS-RD blocking	Relationship	Network without blocking: parameters
reversible routing, solution PF4	$\pi \propto \pi^*$	$\mu_i^* = \mu_i$ $f_i^*(k) = f_i(k) / b_i(k-1)$ $1 \le k \le B_i$ $\mathbf{P}^* = \mathbf{P}$
arbitrary routing, solution PF2	$\pi \propto \pi^*$	$\mu_i^* = \mu_i h_i$ $f_i^*(k) = 1 / b_i(k-1)$ $1 \le k \le B_i$ $\mathbf{P}^* = \mathbf{P}$
	$\pi \propto 1 / \pi^*$	$\mu_i^* = \max_j \mu_j$ $f_i^*(k) = b_i(k-1)$ $1 \le k \le B_i$ $\mathbf{P}^* = \| p^*_{ij} \|$, $p^*_{ij} = \mu_j p_{ji} / \mu_i^*$ $i \ne j$, $p^*_{ii} = 1 - \sum_{j \ne i} p^*_{ji}$, $1 \le i,j \le M$

Table III - Equivalence between networks with and without blocking.

We shall now survey equivalence relationships expressed in terms of state probability π. These equivalence in some cases can be extended to average performance indices such as throughput, utilization, mean queue length and mean response time.
Then, we consider some equivalence properties in terms of passage time distribution.

• *Mean performance indices and joint queue length distribution at arbitrary times*

Some equivalence properties can be defined between networks with and without blocking. They allow us to analyse queueing networks with finite capacity by applying standard computational algorithms for queueing networks with infinite capacity, e.g., MVA and Convolution.
By comparing product-form solutions of queueing networks with and without blocking one can define a non-blocking network with appropriate parameters such that the stationary state distributions of the two networks are identical.
Let W denote the network with finite capacity, and W* the network identical to W except for infinite capacity queues and with the following different parameters: load dependent service rate $\mu^*_i f^*_i(n_i)$, and routing matrix P*. Let π^* denote the stationary state distribution of network W*. Single class exponential networks with RS-RD blocking have been shown to be equivalent, in terms of stationary state distribution, to a corresponding network without blocking, as defined in Table III [8].
The two cases of product-form networks with blocking considered refer to solutions PF4 and PF2, respectively, defined in Section 2.2. Table III shows the type of relationship between the two state distributions and the definition of the parameters of the network without blocking. Note that load dependent function $f^*_i(k)$ can be any positive arbitrary function for $k > B_i$, and in the second case h_i is defined as follows: $h_i = \varepsilon_i y_i$ where ε_i is given in PF2 definition in Table II and $\mathbf{y} = (y_1, \ldots, y_M)$ is obtained by the solution of $\mathbf{y} = \mathbf{y}\, A$, where $A = \| a_{ij} \|$, $a_{ij} = p_{ji}$, $j \ne i$, $a_{ii} = 1 - \sum_{j \ne i} a_{ij}$, $1 \le i,j \le M$.

network topology	performance indices	blocking types	assumptions
two-node	π	BBS-SO=BBS-O RS-RD=RS-FD	
		BBS-SO=RS-RD= =RS-FD=BBS-O	(I) : multiclass networks BCMP type nodes class independent capacities
		BBS-SO=BBS-SNO	assumption (I) and if $N \leq B_1+B_2-2$
		BBS-SO→BAS	assumption (I) and with $B_i^{BBS\text{-}SO}=B_i^{BAS}+1$, $1 \leq i \leq M$
	π U_i,X_i,L_i,T_i	BBS-SO=BBS-SO= =RS-RD	(II) : single class networks exponential nodes load independent service rates
	π U_i,X_i	BBS-SO→BAS	assumption (II) and with $B_i^{BBS\text{-}SO}=B_i^{BAS}+1$, $1 \leq i \leq M$
cyclic	π	BBS-SO=BBS-O RS-RD=RS-FD	
	π U_i,X_i,L_i,T_i	BBS-SO=RS-RD	assumption (II)
		BBS-SO=BBS-SNO	assumption (II), M>2 and $N \leq \min\{B_i+B_j : p_{ij}>0\} - 1$
	π U_i, X_i	BBS-SO→BAS	assumption (II) and with $B_i^{BBS\text{-}SO}=B_i^{BAS}+1$, $1 \leq i \leq M$
central server	π U_i,X_i,L_i,T_i	BBS-SO=BBS- SNO= =BBS-O= =RS-RD=RS-FD	assumption (II) and if only $B_1<\infty$ and $B_i=\infty$, $2 \leq i \leq M$
	π	BBS-SO=RS-RD= =BBS-SNO	assumption (II) and if $B_1=\infty$
		BBS-O→BAS	assumption (II) and if $B_1=\infty$ and $B_i^{BBS\text{-}O}=B_i^{BAS}+1$, $2 \leq i \leq M$

Table IV - Equivalence between closed networks with different blocking types.

Note that since the product-form for queueing networks with finite capacity has a similar structure to the product-form solution of networks with infinite capacity, this type of equivalence could be extended to other cases of product-form networks with blocking, including multiclass networks with different types of nodes. Equivalencies between networks with both homogeneous and non-homogeneous blocking types have been identified both for open and closed networks.
They include both identity relationship and reducibility. Identity states that the state distributions of the two networks are identical, while reducibility allows a correspondence between the two distributions to be defined. Most of the reducibility

network topology	performance indices	blocking types	assumptions
tandem	π	BBS-SO=BBS-O RS-RD=RS-FD	
		BBS-SO=RS-RD=RS-FD	assumption (II)
		BBS-SO=BBS-SNO	assumption (II), M=2 and if $B_1=\infty$
		BBS-SO$\rightarrow$BAS	assumption (II) and with B_iBBS-SO=B_iBAS+1, $2\le i\le M$
split	π	BBS-SO=RS-RD=RS-FD	assumption (II) and if only $B_1<\infty$ and $B_i=\infty$, $2\le i\le M$
		BBS-SO=BBS-SNO= =RS-FD	assumption (II) and if $B_1=\infty$
merge	π	BBS-SO=BBS-O RS-RD=RS-FD	
		BBS-SO=RS-RD=RS-FD	assumption (II) and if $B_1=\infty$
		BBS-SO=RS-RD= =RS-FD= =BBS-SNO=BBS-O	assumption (II) and if only $B_1<\infty$ and $B_i=\infty$, $2\le i\le M$

Table V - Equivalence between open networks with different blocking types.

network topology	performance indices	blocking types	assumptions
reversible routing	π	RS-RD=Stop	single class closed/open networks A-type nodes load independent service
arbitrary routing	π	BBS-SO=RS-FD	(II): single class networks exponential nodes load independent service rates
		Stop=Recirculate	multiclass open Jackson networks with class type fixed
		BBS-SO=RS-RD= =RS-FD=BBS-O	assumption (II) and condition (B)
		BBS-SO=BBS-SNO	assumption (II) and $N\le\min\{B_i+B_j : p_{ij}>0\} - 1$
		Stop$\rightarrow$BBS-O (open) (closed)	single class Jackson networks

Table VI - Equivalence between networks with different blocking types.

between networks can be defined by modifying the buffer capacities. A detailed definition of reducibility and of the definition of the correspondence function for equivalent exponential closed networks is given in [9]. Let $X=Y$ and $X \rightarrow Y$ respectively denote identity and reducibility of blocking types X and Y.
Let B_i^X denote the buffer capacity when node i works under blocking type X and π^X the stationary state distribution of the homogeneous network with blocking type X.
Tables IV, V and VI show equivalences of state distribution and, in some cases, of average performance indices between some blocking types for certain special topology networks.
Tables IV and V concern some closed and open networks, respectively, while Table VI refers to both open and closed networks.
Closed networks with two-node, cyclic and central server topologies are considered in Table IV. The central node in central server networks is denoted by 1.
Tandem open networks and two special cases of open networks denoted as split and merge topology are reported in Table V.
Split topology can be defined as follows: $p_{01}=1$, $p_{0i}=0$, $2\leq i\leq M$, $p_{ij}>0$ for $i=1$ and $2\leq j\leq M$, $p_{ij}=0$ otherwise, $p_{10}=0$, $p_{i0}=1$ for $2\leq i\leq M$,
i.e., an external arrival enters the network only at node 1, from which it can go to nodes 2,...,M and from which it eventually exits from the network.
Merge topology can be defined as:
$p_{01}=0$, $p_{0i}>0$, $2\leq i\leq M$, $p_{ij}>0$ for $2\leq i\leq M$ and $j=1$, $p_{ij}=0$ otherwise, $p_{10}=1$, $p_{i0}=0$ for $2\leq i\leq M$,
i.e., an external arrival enters in any of nodes 2,...,M and then it goes to node 1 from which it leaves the network.
Tables IV, V and VI show the conditions under which equivalence properties between networks with different blocking types hold, including network characteristics and special conditions on system parameters. More specifically, for reducible networks with different blocking types the relationships between finite capacities are shown.

Remark. Note that non-homogeneous networks where service centers work under different and equivalent blocking mechanisms are also equivalent to homogeneous networks with one of the considered blocking types.

The last equivalence reported in Table VI is a special case which relates an open network with M nodes and Stop blocking, which allows a total network population n in the range $L\leq n\leq U$, with a closed network, with an additional node with appropriate parameters and BBS-O blocking, as proved in [10]. Specifically, the closed network is defined by adding a node, denoted by 0, with service rate $\mu_0(n_0)=a(n)$, finite capacity $B_0=U-L$, blocking function $b_0(n_0)=d(n)$, where $n_0=U-n$, $0\leq n_0\leq B_0$ and a(n) and d(n) are the arrival rate and the network blocking functions of the open network with Stop blocking. Hence the following correspondence between state distributions holds: $\pi^{Stop}(\mathbf{S})=\pi^{BBS\text{-}O}(n_0,\mathbf{S})$ for each state $\mathbf{S}$ of the open network.

For the special class of symmetrical networks introduced in the previous section, some equivalence results have been obtained for BBS-SO blocking type in terms of throughput. Specifically, closed cyclic networks have the same throughput for N and N-B customers, as showed for exponential service times in [29] and generalised to phase-type distributions in [28]. Moreover, the relationship between this symmetry property and reversibility is discussed in [28]. Some monotonicity properties of the network throughput for this class of networks has been proved in [61] by considering increasing service rates or finite capacities or the overall network population.

• *Passage time distribution*

Equivalence between networks defined in terms of joint queue length distribution and average performance indices does not necessarily lead to equivalence in terms of passage time distributions.
Some equivalence results have been obtained in terms of cycle time distributions for cyclic networks with BAS and BBS-SO blocking types.
The extension of such results to other networks with different parameters including blocking type, service distribution and routing topology is an open issue.

Consider a two-node cyclic exponential network with N customers, finite capacities B_1 and B_2 and either BBS-SO or BAS blocking. Let $f_{N,B_1,B_2}(t)$ denote the density function of the cycle time. The following equivalence property can be proved [11, 12]:

Theorem 3
Consider two cyclic networks with two exponential nodes, N customers, service rates μ_i, i=1,2, BBS-SO or BAS blocking and finite capacities B_i and B_i', respectively, for i=1,2. If

$$B_1 + B_2 = B_1' + B_2'$$

then the two networks are equivalent in terms of cycle time distribution, i.e.:

$$f_{N,B_1,B_2}(t) = f_{N,B'_1,B'_2}(t).$$

In other words this equivalence states that the distribution of the cycle time does not depend on the single buffer size of each node, but on the total buffer capacity of the network. The extension of these equivalencies to queueing networks with a more general topology and different blocking types is an open issue.

Remark. Note that since the two networks have the same number of customers but different capacities they are also equivalent in terms of throughput, but they are not equivalent in terms of joint queue length distribution and other average performance indices (mean response time, utilization and mean queue length).

4 Conclusions

Performance evaluation of systems with finite capacity resources represented by queueing network models with finite capacity queues and different blocking mechanisms has been discussed.
The main analytical solution methods have been presented, by considering both the analysis of average performance indices and more detailed measures such as passage time distribution.
Properties of queueing networks with blocking have been discussed including equivalence between networks with and without blocking, between models with both homogeneous and non-homogeneous blocking types, and relationships between open and closed queueing network models with blocking.
Although product-form solutions have been proved for queueing networks with blocking under certain constraints, research must to be done to define efficient solution algorithms for general multiclass networks, and in particular approximate solutions with knowledge of the error and bounded algorithms for non product-form networks.
Other open research issues include the analysis of discrete-time queueing networks with finite capacity queues and blocking which can be used to represent discrete time

systems, such as for example ATM networks, and the performance comparison of queueing networks with different blocking types in order to identify optimal blocking mechanisms.

Acknowledgements

The author would like to thank Lorenzo Donatiello for helpful discussions and suggestions and Vittoria De Nitto for useful comments and careful reading of earlier draft of this paper.

Appendix

The system state definition of queueing networks with finite capacity and blocking depends both on network characteristics and on the blocking type. We shall now define system state for the single class network model introduced in Section 1.
For the sake of simplicity we consider exponential service time distribution and the First Come First Served discipline.
The extension to more general service time distributions and policies leads to the introduction of additional components to system states in a similar way as in queueing networks with infinite capacity queues. For this reason in order to define such additional state components which only depend on the node type and are independent of the blocking mechanism, it is sufficient to refer to state definitions introduced for networks with infinite capacity. For example, for queueing networks with BCMP-type nodes one can refer to the state definition introduced in [16] to complete the state definition of queueing networks with blocking defined below.

We consider the five blocking types introduced in Section 1.
$\mathbf{S}=(\mathbf{S}_1,\ldots,\mathbf{S}_M)$ denotes the system state, $\mathbf{S}_i$ the state of node i, and n_i the number of customers in node i, $1\le i\le M$. Due to the finite capacity of the queues $n_i\le B_i$, $1\le i\le M$, and if the network is closed with N customers the following condition holds:

$$\max\left\{0, N - \sum_{j=1, j\neq i}^{M} B_j\right\} \le n_i \qquad (A.1)$$

BAS
For BAS blocking, node i state can be defined as follows:
$\mathbf{S}_i = (n_i, s_i, \mathbf{m}_i)$, $s_i = 0,1$, $\mathbf{m}_i = (m_i,\ldots,m_{u(i)})$, $0\le u(i)\le M-1$
where s_i denotes the server state and $\mathbf{m}_i$ is the vector of the nodes indices blocked by node i. The server state indicates whether the server is active ($s_i = 1$) or blocked ($s_i = 0$). Vector $\mathbf{m}_i$ is non-empty only if node i is full, i.e., if $n_i = B_i$. When $\mathbf{m}_i$ is not empty, it contains the indices of the nodes which have attempted to send a job to node i and which are still blocked by node i (i.e., $p_{ji}>0$ and $s_j = 0$ for each $j=m_i,\ldots,m_{u(i)}$). The number of components of vector $\mathbf{m}_i$, u(i), is at most equal to the number of possible sending (upstream) nodes of node i: $u(i) \le \#\{ j : p_{ji}>0, 1\le j\le M, j\neq i\} \le M-1$.
Vector $\mathbf{m}_i$ is ordered according to the time at which the upstream nodes will be unblocked, that is according to the unblocking scheduling.

BBS-SO and BBS-SNO
In BBS blocking, node i state definition can be defined as follows:
$\mathbf{S}_i = (n_i, d_i)$, $1\le d_i\le M$
where d_i denotes the destination node of the next job that will exit from node i.
Note that d_i is the destination node of the next job currently in service if $n_i>0$, or of the next customer that will arrive at node i if $n_i=0$.

When node i is not empty and the destination node d_i is full, i.e., when $n_i>0$ and $n_{d_i}=B_{d_i}$, then by the blocking definition the server of node i is blocked and will be resumed as soon as a departure occurs from node d_i.
In addition to constraint (A.1) in BBS-SNO if node i is blocked ($n_{d_i}=B_{d_i}$) then $n_i<B_i$, because the server cannot be occupied by a job.

BBS-O
For BBS-O blocking, node i state can be defined as $S_i = n_i$. When at least one of the destination nodes of node i is full (i.e., there exists j : $p_{ij}>0$ and $n_j=B_j$), then the server of node i is blocked.

RS-RD
In RS-RD blocking node i state definition is simply $S_i = n_i$. Note that the server is always active and servicing a customer if $n_i>0$.

RS-FD
For RS-FD blocking the state of node i can be defined as for BBS blocking. Indeed in this case a customer that completes the service at node i and is not accepted by its destination node because of the full capacity does not change its destination as in RS-RD blocking. Hence the information on the destination node of the next customer that will exit from node i has to be included in state S_i.
However, note that when node i is not empty and the destination node d_i is full for RS-FD blocking the server of node i is not blocked as in BBS.
Like RS-RD the server of each node is always active and servicing a customer if $n_i>0$.

Stop and Recirculate
For Stop and Recirculate blocking node i state definition is $S_i = n_i$, like RS-RD and networks with infinite capacity.
Note that for Stop blocking all the servers are blocked when the total network population $n=n_1+\ldots+n_M$ reaches its minimum value for which the (network) blocking function d(n)=0.
For Recirculate blocking the servers are always active and the routing probabilities are state dependent.

Note that even though the system state for RS-RD, BBS-O, Stop and Recirculate blocking types may have the same definition, the underlying Markov processes are different, i.e., the process transition rate matrices Q are defined differently according to the blocking type.
For example for RS-RD blocking matrix Q is defined as presented in Section 2.1, while for Stop blocking $Q=\|q(\mathbf{S},\mathbf{S}')\|$ can be defined as follows for each pair of states $\mathbf{S},\mathbf{S}'$ with $\mathbf{S}\neq\mathbf{S}'$:

$$q(\mathbf{S},\mathbf{S}') = \delta(n_j)\,\mu_j\,d(n)\,p_{ji} \qquad \text{if } \mathbf{S}'=\mathbf{S}+\mathbf{e}_i-\mathbf{e}_j$$
$$q(\mathbf{S},\mathbf{S}') = \delta(n_j)\,\mu_j\,d(n)\,p_{j0} \qquad \text{if } \mathbf{S}'=\mathbf{S}-\mathbf{e}_j$$
$$q(\mathbf{S},\mathbf{S}') = \lambda\,p_{0j}\,b_j(n_j) \qquad \text{if } \mathbf{S}'=\mathbf{S}+\mathbf{e}_j$$

where d(n) is the network blocking function and δ has been defined in Section 2.

References

[1] I.F. Akyildiz "Exact product form solution for queueing networks with blocking" IEEE Trans. Computer, C-36-1 (1987) 122-125.

[2] I.F. Akyildiz and H. von Brand "Central Server Models with Multiple Job Classes, State Dependent Routing, and Rejection Blocking" IEEE Trans. on Softw. Eng., SE-15-10 (1989) 1305-1312.
[3] I.F. Akyildiz and H. von Brand "Exact solutions for open, closed and mixed queueing networks with rejection blocking" J. Theor. Computer Science, 64 (1989) 203-219.
[4] I.F. Akyildiz and N. van Dijk "Exact Solution for Networks of Parallel Queues with Finite Buffers" in: Proc. Performance '90 (P.J.B. 40, I. Mitrani and R.J. Pooley Eds.) North-Holland (1990) 35-49.
[5] I.F. Akyildiz and H.G. Perros, Special Issue on Queueing Networks with Finite Capacity Queues, Performance Evaluation, Vol. 10, 3 (1989).
[6] T. Altiok, S.S. Stidham "A note on Transfer Line with Unreliable Machines, Random Processing Times, and Finite Buffers" IIE Trans., Vol.14, 4 (1982) 125-127.
[7] T. Altiok and H.G. Perros "Approximate analysis of arbitrary configurations of queueing networks with blocking" Ann. Oper. Res. 9 (1987) 481-509
[8] S.Balsamo, G.Iazeolla "Some Equivalence Properties for Queueing Networks with and without Blocking" in *Performance '83* (A.K.Agrawala, S.K.Tripathi Eds.) North Holland.
[9] S. Balsamo and V. De Nitto "Closed queueing networks with finite capacities: blocking types, product-form solution and performance indices" Performance Evaluation, Vol.12, 4 (1991) 85-102.
[10] S. Balsamo, V.De Nitto "A survey of Product-form Queueing Networks with Blocking and their Equivalences" to appear on Annals of Operations Research.
[11] S.Balsamo, L. Donatiello "On the Cycle Time Distribution in a Two-stage Queueing Network with Blocking" IEEE Transactions on Software Engineering, Vol.13, 10, Oct.1989.
[12] S.Balsamo, L. Donatiello "Two-stage Queueing Networks with Blocking: Cycle Time Distribution and Equivalence Properties", in *Modelling Techniques and Tools for Computer Performance Evaluation* (R. Puigjaner, D.Potier Eds.) Plenum Press, 1989.
[13] S.Balsamo, M.C. Clò, L. Donatiello "Cycle Time Distribution of Cyclic Queueing Network with Blocking", in *Queueing Networks with Finite Capacities* (R.O.Onvural and I.F.Akyidiz Eds.), Elsevier, 1993, and Performance Evaluation, Vol.14, 3 (1993).
[14] S. Balsamo, M.C. Clò "State distribution at arrival times for closed queueing networks with blocking" Technical Report TR-35/92, Dipartimento di Informatica, University of Pisa, 1992.
[15] S. Balsamo, M.C. Clò "Delay distribution in a central server model with blocking", Technical Report TR-14/93, Dipartimento di Informatica, University of Pisa, 1993.
[16] F. Baskett , K.M. Chandy, R.R.Muntz, G. Palacios "Open, closed, and mixed networks of queues with different classes of customers" J. of ACM, 22 (1985) 248-260.
[17] O. Boxma and A.G. Konheim "Approximate analysis of exponential queueing systems with blocking" Acta Informatica, 15 (1981) 19-66.
[18] O. Boxma and H. Daduna "Sojourn time distribution in queueing networks" in 'Stochastic Analysis of computer and Communication Systems' (H.Takagi Ed.) North Holland (1990).
[19] P. Caseau and G. Pujolle "Throughput capacity of a sequence of transfer lines with blocking due to finite waiting room" IEEE Trans. on Softw. Eng. 5 (1979) 631-642.
[20] K.M. Chandy, A.J. Martin "A characterization of product-form queueing networks" J. ACM, Vol.30, 2 (1983) 286-299.

[21] K.M. Chandy, J.H.Howard and D. Towsley "Product form and local balance in queueing networks"J. ACM, Vol.24, 2 (1977) 250-263.
[22] K.M. Chandy, U. Herzog and L. Woo "Parametric analysis of queueing networks" IBM J. Res. Dev., 1 (1975) 36-42.
[23] T. Choukri "Exact Analysis of Multiple Job Classes and Different Types of Blocking" in *Queueing Networks with Finite Capacities* (R.O.Onvural and I.F.Akyidiz Eds.), Elsevier (1993).
[24] J.W. Cohen "The multiple phase service network with generalized processor sharing" Acta Informatica, Vol.12 (1979) 245-284.
[25] P.J. Courtois and P.Semal "Computable bounds for conditional steady-state probabilities in large Markov chains and queueing models" IEEE Journal on SAC 4, 6 (1986) 920-936.
[26] Y. Dallery and Y. Frein, A decomposition method for the approximate analysis of closed queueing networks with blocking, Proc. First Int. Workshop on Queueing Networks with Blocking, (H.G. Perros and T. Altiok Eds.) North Holland (1989).
[27] Y. Dallery and D.D. Yao "Modelling a system of flexible manufacturing cells" in: Modeling and Design of Flexible Manufacturing Systems (Kusiak Ed.) North-Holland (1986) 289-300.
[28] Y. Dallery and D.F. Towsley "Symmetry property of the throughput in closed tandem queueing networks with finite buffers" Op. Res. Letters, 10 (1991) 541-547.
[29] V. De Nitto and D. Grillo "Managing Blocking in Finite Capacity Simmetrical Ring Networks" Third Int. Conf. on Data Comm. Systems and their Performance, Rio de Janeiro, Brazil, June 22-25 (1987) 225-240.
[30] Y. Frein and Y. Dallery , Analysis of Cyclic Queueing Networks with Finite Buffers and Blocking Before Service, *Performance Evaluation*, Vol. 10 (1989) 197-210.
[31] S. Gershwin "An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking" Oper. Res., 35 (1987) 291-305.
[32] S. Gershwin and U. Berman "Analysis of transfer lines consisting of two unreliable machines with random processing times and finite storage buffers" AIIE Trans., 13, 1 (1981) 2-11.
[33] W.J. Gordon and G.F. Newell "Cyclic queueing systems with restricted queues" Oper. Res., 15 (1976) 286-302.
[34] L. Gün and A.M. Makowski "An approximation method for general tandem queueing systems subject to blocking" Proc. First Int. Workshop on Queueing Networks with Blocking, (H.G. Perros and T. Altiok Eds.) North Holland (1989) 147-171.
[35] F.S. Hillier and R.W. Boling "Finite queues in series with exponential or Erlang service times - a numerical approach" Oper.Res., 15 (1967) 286-303.
[36] A. Hordijk and N. van Dijk, Networks of queues with blocking, in: Performance '81 (K.J. Kylstra Ed.) North Holland (1981) 51-65.
[37] A. Hordijk and N. van Dijk, Networks of queues ; Part I: job-local-balance and the adjoint process; Part II : General routing and service characteristics, in: Lect. Notes in Control and Information Sciences (F.Baccelli and G.Fajolle Eds.) Springer-Verlag (1983) 158-205.
[38] J.R. Jackson "Jobshop-like queueing systems" Management Science, 10 (1963) 131-142.
[39] F.P. Kelly, Reversibility and Stochastic Networks, Wiley (1979).
[40] J.F.C. Kingman, Markovian population process, J. Appl. Prob., 6 (1969) 1-18.
[41] L. Kleinrock, Queueing Systems.Vol.1 :Theory, Wiley (1975).

[42] A.G. Konhein and M. Reiser "A queueing model with finite waiting room and blocking" SIAM J.Comput, 7 (1978) 210-229.
[43] D.Kouvatsos and N.P.Xenios "Maximum entropy analysis of general queueing networks with blocking" Proc. First Int. Workshop on Queueing Networks with Blocking, (H.G. Perros and T. Altiok Eds.) North Holland (1989).
[44] A.E. Krzesinski "Multiclass queueing networks with state-dependent routing" Performance Evaluation, Vol.7, 2 (1987) 125-145.
[45] S. Kundu and I.Akyildiz "Deadlock free buffer allocation in closed queueing networks" Queueing Systems Journal, 4, 47-56.
[46] S.S. Lam "Queueing networks with capacity constraints" IBM J. Res. Develop. 21 (1977) 370-378.
[47] S.S. Lavenberg, Computer Performance Modeling Handbook, (Prentice Hall, 1983).
[48] S. S. Lavenberg and M. Reiser "Stationary State Probabilities at Arrival Instants for Closed Queueing Networks with multiple Types of Customers" *J. Appl. Prob.*, Vol. 17 (1980) 1048-1061.
[49] M.F. Neuts "Two queues in series with a finite intermediate waiting room" J.Appl. Prob., 5 (1986) 123-142.
[50] R.O. Onvural "A Note on the Product Form Solutions of Multiclass Closed Queueing Networks with Blocking" Performance Evaluation, Vol.10, 3 (1989) 247-253.
[51] R.O. Onvural "Survey of Closed Queueing Networks with Blocking" ACM Computing Surveys, Vol. 22, 2 (1990) 83-121.
[52] R.O. Onvural Special Issue on Queueing Networks with Finite Capacity, Performance Evaluation, Vol. 17, 3 (1993).
[53] R.O. Onvural and H.G. Perros "On equivalences of blocking mechanisms in queueing networks with blocking" Oper. Res. Letters (1986) 293-297.
[54] R.O. Onvural and H.G. Perros "Some equivalencies on closed exponential queueing networks with blocking" Performance Evaluation, Vol.9 (1989) 111-118.
[55] H.G. Perros "Open queueing networks with blocking" in : Stochastic Analysis of Computer and Communications Systems (Takagi Ed.) North Holland (1989).
[56] H.G. Perros "A bibliography of papers on queueing networks with finite capacity queues" Performance Evaluation, Vol. 10, 3 (1989) 225-260.
[57] H.G. Perros, A. Nilsson and Y.G. Liu "Approximate analysis of product form type queueing networks with blocking and deadlock" Performance Evaluation, Vol. 8 (1988) 19-39.
[58] H.G. Perros and P.M. Snyder "A computationally efficient approximation algorithm for analyzing queueing networks with blocking" Performance Evaluation, Vol. 9 (1988/89) 217-224.
[59] B. Pittel "Closed exponential networks of queues with saturation: the Jackson-type stationary distribution and its asymptotic analysis" Math. Oper. Res. 4 (1979) 367-378.
[60] K. S. Sevcik and I. Mitrani "The Distribution of Queueing Network States at Input and Output Instants" *J. of ACM*, Vol. 28, 2 (1981) 358-371.
[61] G.J. Shantikumar and D.D. Yao "Monotonicity properties in cyclic networks with finite buffers" Proc. First Int. Workshop on Queueing Networks with Blocking, (H.G. Perros and T. Altiok Eds.) North Holland (1989).
[62] R. Suri and G.W. Diehl "A variable buffer size model and its use in analytical closed queueing networks with blocking" Management Sci. Vol.32, 2 (1986) 206-225.
[63] D.F. Towsley "Queueing network models with state-dependent routing" J. ACM 27 (1980) 323-337.

[64] N. van Dijk "On 'stop = repeat' servicing for non-exponential queueing networks with blocking" J. Appl. Prob., 28 (1991) 159-173.
[65] N. van Dijk "'Stop = recirculate' for exponential product form queueing networks with departure blocking" Oper. Res. Lett., 10 (1991) 343-351.
[66] N. van Dijk and H.G. Tijms "Insensitivity in two node blocking models with applications" in: Proc. Teletraffic Analysis and Computer Performance Evaluation, Eds. Boxma, Cohen and Tijms (North Holland, 1986) 329-340.
[67] N. van Dijk "On the Arrival Theorem for communication networks" Computer Networks and ISDN Systems, 25 (1993) 1135-1142.
[68] D.D. Yao and J.A. Buzacott "Modeling a class of state-dependent routing in flexible manufacturing systems" Ann. Oper. Res., 3 (1985) 153-167.
[69] D.D. Yao and J.A. Buzacott "Modeling a class of flexible manufacturing systems with reversible routing" Oper. Res., 35 (1987) 87-93.
[70] P. Whittle "Partial balance and insensitivity" J. Appl. Prob. 22 (1985) 168-175.