Queueing Networks with Blocking: Analysis, Solution Algorithms and Properties

Simonetta Balsamo

Dipartimento di Informatica
Università Ca' Foscari di Venezia
via Torino, 155 Mestre-Venezia, Italy

Abstract. Queueing network models with finite capacity queues and blocking are used for modeling and performance evaluation of systems with finite resources and population constraints, such as communication and computer systems, traffic, production and manufacturing systems. Various blocking types can be defined to represent different system behaviors, network protocols and technologies. Queueing networks with blocking are difficult to analyze, except for the special class of product-form networks. Most of the analytical methods proposed in literature provide an approximate solution with a limited computational cost. We introduce queueing networks with finite capacity queues and blocking, the main solution techniques for their analysis, both exact and approximate algorithms, and some network properties. We discuss the conditions under which exact solutions can be derived, and criteria for the appropriate selection of approximate methods. We present equivalence properties among different types of blocking types, the analysis of heterogeneous networks, and some application examples.

Keywords: Queueing Networks, Blocking, Product-form models, Equivalence properties.

1 Introduction

Performance analysis of various systems, including communication and computer systems, as well as production and manufacturing systems can be carried out through queueing network models. System performance analysis consists of the derivation of a set of figures of merit, that typically includes queue length distribution and some average performance indices such as mean response time, throughput, and utilization. Queueing networks with finite capacity queues and blocking have been introduced to represent systems with finite capacity resources and population constraints. When a queue reaches its maximum capacity then the flow of customers into the service center is stopped, both from other service centers and from external sources in open networks, and the blocking phenomenon arises. Various blocking mechanisms have been defined and analyzed in the literature to represent distinct behaviors of real systems with limited resources [42, 47, 53, 55, 57, 58]. Some comparisons and equivalences among

D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 233–257, 2011.

blocking types have been presented for queueing networks with various topologies in [9, 10, 42, 45, 47, 57, 58]. Performance analysis of queueing networks with blocking can be exact or approximate. Exact solution algorithms have been proposed to evaluate both average performance indices, queue length distribution [10, 42, 47], and passage time distribution [7, 10, 11]. Under exponential assumption one can define and analyze the continuous-time Markov chain underlying the queueing network. In some special cases queueing networks with blocking show a product-form solution, under particular constraints, for various blocking types; a survey is presented in [10]. Some solution algorithms for product-form networks with finite capacities have been defined [8, 19, 50].

However, except for this special class of models, queueing networks with blocking do not have a product-from solution and a numerical solution of the associated Markov chain is seriously limited by the space and time computational complexity that grows exponentially with the model number of components. Hence recourse to approximate analytical methods or simulation is necessary. Several approximate solution methods for queueing networks with blocking have been proposed in literature both for open and closed models and surveys of some methods have been presented in [6, 42, 47]. Most of these methods provide an approximate solution with a limited computational cost. However, they do not provide any bound on the introduced approximation error. They are usually validated by comparing numerical results with either simulation results or exact solutions. Many approximation methods are heuristics based on the decomposition principle applied to the underlying Markov process or, more often, to the network itself. Some methods consider a forced solution of a product-form network and some approximations are based on the maximum entropy principle.

In this paper we focus on queueing networks with finite capacity queues and blocking, their exact and approximate analysis, their properties and applications. We consider various blocking mechanisms that represent different system behaviors. We review the main solution methods and algorithms to analyze queueing networks with blocking to evaluate a set performance indices, the conditions and some criteria for the appropriate selection of the solution method. We consider exact and approximate analytical methods for open and closed queueing networks with blocking. We recall some equivalence properties that allow the solution of heterogeneous models and some application examples.

The paper is structured as follows. Section 2 introduces the model definition of queueing networks with finite capacity queues, the various blocking mechanisms and the main performance indices. Section 3 describes the exact analysis of queueing networks with blocking based on analytical methods, that includes the approach based on the Markov chain definition and analysis, and the special cases of networks with product-form solutions. Section 4 recalls the main principles and approaches proposed for the approximate analytical solution of networks with blocking. Section 5 compare some approximation methods for closed and open networks with blocking, and for different network topologies. In Section 6 we recall some equivalence properties of network models with different blocking types, and we illustrate an application example.

2 Model Definition and Blocking Mechanisms

A queueing network consists of a set of service centers, each formed by a queue and a set of identical servers that provide service to a set of customers. Let us consider a network with N queues with finite capacity, one class of customers, and probabilistic routing. The network may be open or closed. For a closed network let K denote the number of customers. For an open network let λ_i be the external arrival rate (from outside the network) to station i, $1 \leq i \leq N$. For the sake of simplicity we usually assume exponential service time distribution and Poisson arrivals. The service rate of station i is denoted by μ_i, and K_i is the number of servers, usually just single servers, $1 \leq i \leq N$. The finite capacity of node i is denoted by B_i , $1 \leq i \leq N$. Let $\mathbf{P} = [p_{ij}]$ denote the routing probability matrix, where p_{ij} is the probability for a customer to go to station j after being served by station i, $1 \leq i, j \leq N$, and p_{i0} is the probability that a customer leaves the network after being served by station i. Let $\mathbf{e} = (e_1, \ldots, e_N)$ denote the solution of the traffic equations, defined as follows:

$$e_i = \lambda_i + \sum_{j=1}^{N} e_j p_{ji} , \quad 1 \leq i \leq N \tag{1}$$

If the routing probability matrix is irreducible, this system has a unique solution for open networks and an infinite number of solutions for a closed network, unique up to a multiplicative constant.

In queueing networks with finite capacities when a customer attempts to enter a finite capacity queue that is full, it can be blocked. We shall now introduce the definition of some blocking mechanism that describe the blocked customers behavior.

2.1 Blocking Types

Various blocking types have been defined to represent different system behaviors. We now recall three of the most commonly used blocking types defied for computer, communication, networks, and production systems [10, 42, 53].

- 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, and its 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.
- 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 subcategories distinguish whether the server can be used as a buffer when the node is blocked: BBS-SO (server occupied) and BBS-SNO (server not occupied). Hereafter we consider BBS-SO blocking, which is simply called BBS.
- Repetitive Service Blocking (RS): if, upon completion of its service at ode i, a job attempts to enter a destination queue j, which is full, the job is looped back into the sending queue i, whereupon it receives a new, independent and identically distributed service according to the service discipline. Two subcategories distinguish 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) and RS-FD (fixed destination).

Another kind of blocking, called *generalized blocking* or *kanban blocking*, is defined when the server continues processing customers in the queue even if the destination node is full [17, 18, 38, 39]. The customers that have completed service at node i, but cannot be sent to the next node, continue to share the buffer space of node i along with the other customers that are either waiting for service or being served upon. The customers arriving at a node when the queue is full are lost. This blocking is defined to model manufacturing systems. For particular values of the parameters that define this blocking type, it reduces to other blocking types, including BBS-SO and BAS.

Other types of blocking mechanisms model population constraints in a network, by assuming that the number of customers are in the range $[L, U]$, i.e., L and U are the minimum and maximum populations admitted, respectively. Let $a(n)$ denote a load dependent arrival rate function and $d(n)$ a non-negative departure blocking function, where $n \geq 0$ is the overall network population. Then we set $a(n) = 0$ for $n \geq U$ and $d(n) = 0$ for $n \leq L$. The blocking types defined for population constraints include the following types [36, 57, 58].

- Stop Blocking: the service rate at each node depends on the number n of customers in the network, according to function $d(n)$. When $d(n) = 0$ the service at each node is stopped. Service at a node is resumed upon arrival of a new customer to the network.
- Recirculate Blocking: a job upon completion of its service at node i leaves the network with probability $p_{i0}d(n)$, when n is the total network population and it is forced to stay in the network with probability $p_{i0}[1 - d(n)]$, where p_{i0} is the routing probability. Hence, a job completing the service at node i enters node j with state dependent routing probability $p_{ij} + p_{i0}[1 - d(n)]p_{0j}$, $1 \leq i, j \leq N$, $n \geq 0$.

Closed queueing networks with finite capacity queues and blocking can deadlock, depending on the blocking type. Deadlock prevention or detection and resolving techniques must be applied. Deadlock prevention for blocking types BAS, BBS and RS-FD requires that the overall network population K 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 $\mathbf{P}$ is irreducible and K is less than the total buffer capacity of the nodes in the network [10, 42]. Moreover, to avoid deadlocks for BAS and BBS blocking types we assume $p_{ii} = 0$, $1 \leq i \leq N$. In the following we shall consider deadlock-free queueing networks in steady-state conditions.

2.2 Performance Indices

Queueing networks with blocking are used to model real life systems with finite capacities and to estimate various performance indices. These performance metrics may be defined for each node, and, in multiclass networks, for each chain and/or class. Performance indices are defined in terms of distributions of various random variables, or in terms of average or mean rate of a performance measure. The most commonly used average performance indices are, for each node i: the utilization U_i, the throughput X_i, the average queue length L_i and the mean response time T_i. Performance indices evaluated in terms of random variable distribution are the queue length n_i, i.e., the number of customers at node i, and the number of active servers at node i, that is servers that are neither empty nor blocked. More complex analysis can be carried out to derive more detailed performance indices, such as the customer passage time distribution through the node, and the cycle time distribution for closed network [7, 11].

The definition of the performance indices, both probabilities and average values, depends on the blocking type. Let $PB_i(n_i)$ denote the probability that node i is not empty and blocked when there are n_i customers in node i, and let $PB_i = \sum_{n_i} PB_i(n_i)$ the overall blocking probability. They depend on the blocking type [10].

Then the performance indices for each node i can be defined as follows:

- queue length distribution $\pi_i(n_i)$, $max(0, K - \sum_{j \neq i} B_j) \leq n_i \leq B_i$
- utilization $U_i = 1 - \pi_i(0) - PB_i$
- throughput $X_i = \sum_{n_i} [\pi_i(n_i) - PB_i(n_i)]\mu_i(n_i)$, for load dependent service rate and $X_i = U_i \mu_i$ for constant service rate
- mean queue length $L_i = \sum_{n_i} n_i \pi_i(n_i)$
- mean response time $T_i = L_i / X_i$
- mean cycle time $\sum_j x_j T_j / x_i$.

In networks with blocking we can further define a specific performance index called *effective utilization*. It is defined as the fraction of time that the node is neither empty nor blocked, that is a measure of the useful work of the node. Similarly, for BBS blocking where the interrupted service is repeated and for RS blocking where the job can be looped back, we can also define the *effective throughput* as a measure of the useful work of the node. It is given by the fraction of throughput that is not due to the service repetition because of blocking.

In this paper we mainly focus on the evaluation of the queue length distribution and the average performance indices. We now recall the main method used for exact analysis of queuing networks with blocking.

3 Exact Analysis of Networks with Finite Capacities

Exact analysis of queueing networks with finite capacities and blocking can be obtained by representing the model with a stochastic Markov process, and specifically with a continuous-time Markov chain. By using exact analysis of queueing networks with blocking one can evaluate of a set of average performance indices, the joint queue length distribution at arbitrary times and at arrival times, and possibly the passage time and cycle time distributions. We shall now recall the exact solution based on the Markov process associated to queueing networks with blocking to evaluate the queue length distribution and average performance indices. Then we present the special class of product-form networks with blocking that can be solved by more efficient techniques.

3.1 Markov Process Analysis

Under the assumptions of exponential delays and independence between service times and inter-arrival times, the network can be represented by a continuous-time, homogeneous Markov chain. Let $\mathbf{S} = (S_1, \ldots, S_N)$ denote the state of the network with N nodes, and let E be the state space, i.e., the set of all feasible states. The network model evolution can be represented by a continuous-time ergodic Markov chain with discrete state space E and transition rate matrix $\mathbf{Q}$. The stationary and transient behaviour of the network can be analyzed by the underlying Markov process. Under the hypothesis of an irreducible network routing matrix $\mathbf{P}$, there exists a unique steady-state queue length probability distribution, denoted by $\pi = [\pi(\mathbf{S})], \forall \mathbf{S} \in E$. It can be obtained by solving the homogeneous linear system of the global balance equations

$$\pi \mathbf{Q} = \mathbf{0}, \tag{2}$$

subject to the normalising condition $\sum_{\mathbf{S} \in E} \pi(\mathbf{S}) = 1$ and where $\mathbf{0}$ is the all zero vector.

The definition of state $\mathbf{S}$, state space E and the transition rate matrix $\mathbf{Q}$ depends on the network characteristics and on the blocking type of each node [9, 10, 42, 47, 57, 58]. Each process state transition corresponds to a particular set of events on the network model, such as a job service completion at a node and the simultaneous transition towards another node or an external arrival at a node. This correspondence depends on the blocking type.

For example for RS-RD blocking, by definition the servers cannot be blocked. That is the server is always active and servicing a customer, if $n_i \geq 0$. Therefore, under exponential assumptions, node i state definition is simply $S_i = n_i$. For BAS blocking we have to consider the server activity and the scheduling of the nodes that are blocked by a full destination node. Then, under exponential assumptions, the process state of node i can be defined as $S_i = (n_i, s_i, \mathbf{m}_i)$, where n_i is the number of jobs in node i, s_i is the number of servers of node i blocked by a full destination node and therefore containing a served job, $0 \leq s_i \leq min(n_i, K_i)$, for K_i servers of node i, and $\mathbf{m}_i$ is the list of nodes blocked by node i. For BBS blocking, since a job declares its destination node j before it starts receiving service, and it can be blocked when node j is full, then the state can be defined as $S_i = (n_i, \mathbf{NS}_i)$, where n_i is the number of jobs in node i, and $\mathbf{NS}_i$ is a vector defined only for nodes that can be blocked. The j-th component $NS_i(j)$ denotes the number of node i servers that are servicing jobs destined to node j, and for an open network $NS_i(0)$ denotes the number of node i servers with jobs that will leave the network.

For each blocking type one can define the corresponding transition rate matrix $\mathbf{Q}$ and solve the liner system (2) to derive the steady-state distribution π. From vector π one can derive the queue length distribution of node i, π_i, and other average performance indices of node i, such as throughput (X_i), utilization (U_i), average queue length (L_i) and mean response time (T_i).

By summarizing, exact analysis of queueing network model with finite capacities based on a continuous-time Markov process requires:

1. Definition of system state $\mathbf{S}$ and state space E according to the blocking type.
2. Definition of transition rate matrix $\mathbf{Q}$ according to the blocking type and the network topology.
3. Solution of global balance equations (2) to derive the steady-state distribution $\pi(\mathbf{S}), \forall \mathbf{S} \in E$.
4. Computation, from the steady-state distribution π, of the average performance indices for each node of the network.

The numerical solution based of the Markov chain analysis is seriously limited by the space and time computational complexity that grows exponentially with the model number of components. For open network the Markov chain is infinite and, unless a special regular structure of matrix $\mathbf{Q}$ allows to derive closed form expression of the solution π, one has to approximate the solution on a truncated state space. For closed networks the time computational complexity of liner system (2) is determined by the space state E cardinality that grows exponentially with the buffer sizes ($B_i \leq K, 1 \leq i \leq N$) and N. Although the state space cardinality of the process can be much smaller than that of the process underlying the same network with infinite capacity queues (which is exponential in K and N), it still remains numerically untractable as the number of model components grows.

When special constraints are satisfied we can apply exact analysis based on product-form, that we now introduce, or, in many practical cases, it is necessary to apply approximate solution methods.

3.2 Product-Form Networks

In some special cases, queueing networks with blocking have a product-form solution, under certain constraints on network parameters and for various blocking types. Various product-form networks with finite capacities have been defined [1, 3, 9, 10, 25, 27, 40, 41, 58, 59]. A detailed description of product-form solutions of networks with blocking and equivalence properties among different blocking network models is presented in [9] and in [10, 59]. Some efficient algorithms for some closed product-form networks with blocking have been defined [7, 19, 50] and can be applied to derive the performance indices, under some constraints.

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

$$\pi(\mathbf{S}) = \frac{1}{G} V(n) \prod_{i=i}^{N} g_i^{n_i}, \quad \forall \mathbf{S} \in E \tag{3}$$

where G is a normalising constant and $n = \sum_{i=1}^{N} n_i$ is the total network population, n_i is the number of customers in node i, defined in the node state S_i. The definition of functions V and g_i depends on some network parameters, which include the solution e_i of the traffic balance equations (1) and the service rates $\mu_i, 1 \leq i \leq N$, on the blocking type and some additional constraints.

We shall now recall the main product-form results. For the sake of simplicity we provide the product-form definition for single class networks.

Consider the following five network topologies: two-node networks, cyclic topology, central server (or star topology), reversible routing networks, and arbitrary topology. The first three are special cases of closed networks, the last two apply to closed and open networks.

Reversible routing. A routing matrix $\mathbf{P}$ is said to be reversible if the following conditions hold:

$$e_i p_{ij} = e_j p_{ji} \quad , \quad \lambda_i = e_i p_{i0} \qquad \forall 1 \leq i, j \leq N \tag{4}$$

where $\mathbf{e} = [e_1, \ldots, e_N]$ is the solution of system (1).

Note that for closed networks, only the first condition of this definition has to be verified. Two-node networks are a special case of reversible routing.

In order to define some cases of product-form we introduce the following definitions.

Condition 1. (Non-empty condition). The non-empty condition for closed networks requires that at most one node can be empty, i.e., $K \geq B - B_{min}$, where $B = \sum_{1 \leq i \leq N} B_i$ and $B_{min} = \min_{1 \leq i \leq N} B_i$.

Condition 2. (Strictly non-empty condition). This condition is said to hold strictly when each node can never be empty, i.e., the inequality is strict: $K > B - B_{min}$.

Condition 3. (Single destination node). Each node i with finite capacity is the only destination node for each upstream node, i.e., if $B_i < K$ and $p_{ji} > 0$ then $p_{ji} = 1$, $1 \leq i, j \leq N$.

Condition 4. (Only one node blocked). At most one node can be blocked, i.e., if $K = B_{min} + 1$.

Definition: A-type node. An A-type node has arbitrary service time distribution, symmetric scheduling discipline or exponential service time, identical for each class at the same node, when the scheduling is arbitrary [3].

Product-form solution (3) has been derived for networks with different blocking types and with different topologies. Some product-forms hold for both homogeneous networks, that is where each node operates with the same blocking mechanism, and non-homogeneous ones, where different nodes in the networks work under different blocking mechanisms. Table 1 summarizes the main cases of allowed combination of blocking types for each network topology, under some additional constraints, i.e., conditions 1 through 4 defined above, and where product-form (3) is defined by formulas $F1$ through $F5$ as follows.

Table 1. Product-form heterogeneous networks with blocking

Network topology	Blocking types	Product-form formulas
Two nodes	BAS, BBS, RS	F1
Cyclic topology	BBS, RS	F2 and Condition 1
Central server (star)	BBS, RS (central node with RS)	F3
Reversible routing	RS-RD, Stop	F4
Arbitrary routing	BBS, RS-FD	F2 and Conditions 2 and 3
Arbitrary routing	BAS	F5 and Condition 4

Let us define $\rho_i = e_i/\mu_i$, where μ_i is the service rate of node i, and e_i the solution of the system of traffic equations (1).

Product-form F1. For multiclass networks with BCMP-type nodes [13] and class independent capacities, formula $F1$ defines: $V(n) = 1$ and $g_i(n_i) = \rho_i^{n_i}$.

Product-form F2. For single class network and nodes with exponential service time distribution, and load independent service rates μ_i, formula $F2$ defines: $V(n) = 1$ and $g_i(n_i) = 1/y_i$, where $\mathbf{y} = (y_1, \ldots, y_N)$ is the solution of the equations $\mathbf{y} = \mathbf{y}\mathbf{P}'$, and matrix $\mathbf{P}' = [p'_{ij}]$ is defined in terms of the routing probability matrix $\mathbf{P}$ and the service rates as follows: $p'_{ij} = \mu_j p_{ji}, p'_{ii} = 1 - \sum_{j \neq i} p'_{ji}, 1 \leq i, j \leq N$.

Product-form F3. It applies to multiclass central server networks with A-type nodes, the class type of a job fixed in the system, state-dependent routing depending on the class type, and blocking functions dependent on node. Let 1 denote the central node. Let $b_i(n_i)$ denote the blocking function of node i, that is the probability that a job arriving at node i, is accepted when there are n_i customers. For single class exponential networks, load dependent service rates $\mu_i(n_i) = \mu_i f_i(n_i)$, and the state-dependent routing defined as $p_{1j}(n_j) = w_j(n_j)w(K - n_1), \forall n_j, p_{j1} = 1$, and $2 \leq j \leq N$, formula $F3$ defines:

$$V(n) = \prod_{l=i}^{K-n_1} w(l-1) \prod_{j=2}^{N} \prod_{l=i}^{n_j} w_j(l-1), \;\; g_i(n_i) = \prod_{l=i}^{n_i} \frac{1}{\mu_i} \frac{b_i(l-1)}{f_i(l)}, \forall i. \qquad (5)$$

For the definition of formula $F3$ for multiclass central server networks expression refer to [3].

Product-form F4. It applies to single class networks with A-type nodes. For the case load dependent service rates $\mu_i(n_i) = \mu_i f_i(n_i)$, and blocking function $b_i(n_i)$ for each node i, formula $F4$ defines: $V(n) = 1$ and $g_i(n_i) = \rho_i^{n_i} \prod_{l=i}^{n_i} \frac{b_i(l-1)}{f_i(l)}$.

Product-form F5. For multiclass networks and nodes with FCFS service discipline, exponential service time, and class independent capacities. Formula $F5$, like $F1$, defines: $V(n) = 1$ and $g_i(n_i) = \rho_i^{n_i}$.

Note that product-form formula (3) generalizes the closed-form expression for BCMP networks [13], and in certain cases, corresponds to the same solution as for queueing networks with infinite capacity queues computed on the truncated state space defined by the network with finite capacities. Product-forms for queueing networks with finite capacities are proved mostly by applying two approaches: i) reversibility of the underlying Markov process, ii) duality.

The former approach applies to reversible routing 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. This allows us to immediately derive a product-form solution 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. For example networks with RS blocking, BCMP-type networks with finite capacity and reversible routing $\mathbf{P}$ have product-form steady-state distribution given by formula $F1$ defined above [3, 27, 41]. Note that this solution is the BCMP product-form, renormalised over the reduced state space.

The latter approach, duality, applies to networks with arbitrary topology (possibly non-reversible) routing, for which the product-form solution is derived by the definition of a dual network that has the same equilibrium probability distribution. The dual network is proved to be in product-form under the *non-empty condition* (condition 1). For example, consider a cyclic closed network with single class, load independent exponential service rates and BBS or RS blocking. We can define a dual network which has the same steady-state joint queue length distribution [25]. It is obtained from the original one by reversing the connections between the nodes. It is formed by N nodes and $(B-K)$ customers, which correspond to the 'holes' of the original (primal) network, where $B = \sum_{i=1}^{N} B_i$ is the total capacity of the network. When a customer moves from node i in the original network, a hole moves backward to node i in the dual one. The state of n_i customers in node i of the original network, corresponds to $B_i - n_i$ holes in node i of the dual one contains. The underlying Markov process that describes the evolution of customers in the network is equivalent to the one that describes the evolution of the holes in the dual network. Hence, when the non-empty condition is satisfied, the total number of holes in the dual network cannot exceed the minimum capacity, i.e., $(B-K) \leq B_{min}$, and the dual network has a product-form solution like a network without blocking. Then the product-form solution for the primal network is given by equation (3) with formula $F2$ defined above [25]. This solution can be extended to arbitrary topology networks with load independent service rates for RS blocking, as proved in [27]. Another remarkable example of duality is for closed cyclic networks with phase-type (general) service distributions and BBS blocking for which the throughput of the network is shown to be symmetric with respect to its population, i.e., $X(B-K) = X(B)$ [22].

3.3 Algorithm for Closed Networks with Blocking

Product-form closed networks with blocking can be analyzed by some efficient algorithms [8, 19, 50]. They can be applied if some additional constraints are verified. They provide the model solution with a time computational complexity linear in the number of network components, i.e., they require $O(NK)$ operations, for a network with N service centers and K customers. There are two types of algorithms for product-form closed networks with blocking: Convolution and MVA (Mean Value Analysis). Note that we cannot directly apply the algorithms already known for BCMP networks, such as convolution algorithm and MVA [49], because of the different state space definition. However, the main idea of the two algorithms is similar to the non blocking case. Convolution algorithm aims to evaluating the normalizing constant G in formula (3) and average performance indices. MVA provides a direct computation of a set of average performance indices (mean response time, throughput, and mean queue length).

Convolution algorithm. We shall now briefly recall a Convolution algorithm for product-form queueing networks with blocking, whose computational complexity has a linear time computational complexity in the number of network components. With respect to the algorithm for BCMP networks, a Convolution algorithm for queuing networks with finite capacities takes into account the set of constraints on the queue lengths. This corresponds to a state space limitation that leads to a new definition of recursive equations to compute the normalizing constant.

The Convolution algorithm applies to networks with RS and BBS blocking, arbitrary topology, load independent service rates, and product-form solution given by formula $F1$ or $F2$. The algorithm computes the normalizing constant G in formula (3). This is obtained by on a set of recursive equations to evaluate functions $G_j(n)$, that can be interpreted as the normalizing constant of the network with finite capacity queues and with the first j nodes and n customers, $1 \leq j \leq N$, $\forall$ feasible $n \leq K$. The algorithm eventually computes $G = G_N(K)$. It defines a set of different recursive equations depending on the network population and the finite capacities. Once the last function $G_N(n)$, for each feasible n, has been computed, one can derive for each node i, the marginal queue length distribution $\pi_i(n_i)$, $\forall n_i$, and the average performance indices, i.e., the mean queue length L_i, the mean response time T_i, the node throughput X_i and utilization U_i, the mean busy period, and the blocking probabilities.

The time computational complexity of the algorithm depends on the network parameters and is $O(NC)$, where $C = max_{1\leq i\leq N}(B_i - a_i)$, and $a_i = max(0, K - \sum_{j\neq i} B_j)$ is minimum feasible queue length of node i. A detailed description of the algorithm is given in [8].

MVA algorithm. The MVA algorithm directly computes a set of average performance indices, without evaluating the normalizing constant. The algorithm recursively evaluates the mean queue length, mean response time, and throughput. Other performance indices that can be derived are utilization, mean busy period and blocking probabilities for each node.

An MVA algorithm is defined for the class of product-form networks with cyclic topology and with BBS-SO and RS blocking [19]. In this case product-form $F2$ holds when the non-empty condition is satisfied, and we can define a dual network without blocking with identical product-form state distribution. Hence, by duality, this algorithm simply applies the standard MVA algorithm for networks without blocking to the dual network (see [19] for details). Note that such a MVA algorithm is not a direct application of the arrival theorem, as we have in MVA for queueing networks without blocking [49], since it is based on the dual network that is without blocking. The arrival theorem for network with blocking is discussed in [7, 10, 14].

Another MVA algorithm has been extended to a class of product-form networks with RS blocking, load independent service rates, and with $F2$ or $F3$ product-form [50]. The MVA algorithm has a time computational complexity of $O(B_{max}NK)$ operations, where $B_{max} = \max_{1\leq i\leq N} B_i$.

4 Approximate Analysis of Networks with Finite Capacities

General queueing networks with blocking that have not a product-from solution can be analyzed by approximate analytical methods or by simulation. Several approximate techniques for open or closed queueing networks with finite capacity queues have been proposed to evaluate average performance indices and queue length distributions [10, 47, 54]. Most of the methods provide an approximate solution with a limited computational cost, but they do not give any bound on the introduced approximation error. The accuracy of the methods is usually validated by comparing numerical results with either simulation results or exact solutions.

Various heuristics have been defined by taking into account both the network model characteristics and the blocking type [4, 15, 16, 20, 23–26, 28–35, 37, 42, 48, 53–56, 60, 61]. Approximate methods for queuing networks with finite capacities are defined on the basis of the following principles:

- decomposition applied to the Markov process or to the network,
- forced product-form solution,
- structural properties for special cases,
- maximum entropy.

The various approaches can be applied under some constraints and for some blocking type, and they show different accuracy and time computational complexity. The methods based on forced product-form solution try to apply the product-form results to networks that do not satisfy the required constraints, possibly making some iterative check to appropriately select the approximation parameters. They usually are quite efficient from the computational viewpoint, but with unknown approximation error. Networks with particular topologies can be solved by special approximation methods that take advantage of their structure. Maximum entropy approximations apply the maximum entropy method (ME) to match the performance indices, which leads to a closed-form solution of the queue length distribution. ME approximation can be applied under quite general conditions and provide a good accuracy [29–33, 35]. We now discuss the decomposition approach that is widely used.

Network and process decomposition. Many approximate methods are heuristics based on the *decomposition* principle applied to the underlying Markov process or directly to the network.

Decomposing a Markov process consists in identifying a state space E partition of into H subsets E_h, $1 \leq h \leq H$, which leads to a decomposition of the rate matrix $\mathbf{Q}$ into H^2 submatrices. Hence the solution of the entire system of global balance equations (2) is reduced to the solution of H subsystems of smaller dimension. Each subsystem is related to a subset E_h, so obtaining the conditioned state probability denoted by $Prob(\mathbf{S} \mid E_h)$, $\forall$ state $\mathbf{S} \in E_h$, $\forall h$. Then these solutions are combined to obtain the overall process solution, i.e., the state distribution as

$$\pi(\mathbf{S}) = Prob(\mathbf{S} \mid E_h) Prob(E_h) \tag{6}$$

where $Prob(E_h)$ is the aggregated probability of subset E_h, $\forall h$. Then the decomposition technique substitutes the direct computation of $\pi(\mathbf{S})$ with the computation of probabilities $Prob(\mathbf{S} \mid E_h)$ and $Prob(E_h)$, $\forall S, \forall E_h$. Exact process decomposition in general cannot be efficiently applied, except for special cases.

Approximate methods based on the decomposition of the Markov process provide an approximate evaluation of these probabilities. They require to:

- identify a partition of E into H subsets, so decomposing state space E and transition rate matrix $\mathbf{Q}$,
- compute the conditional state probabilities $Prob(\mathbf{S} \mid E_h)$ and the aggregate probabilities $Prob(E_h)$for each subset $E_h, \forall h$, and compute state probability π by formula (6).

A critical issue is the definition of the state space partition that affects both the accuracy and the time computational complexity of the approximate algorithm. If the partition of E corresponds to a network partition into subnetworks then subsystems are (possibly modified) subnetworks.

The decomposition principle applied to the queueing network is based on the aggregation theorem for queueing networks. It performs in three steps: 1) network decomposition into a set of subnetworks, 2) analysis of each subnetwork in isolation to define an aggregate component, 3) definition and analysis the new aggregated network. Step 1 is a NP-complete problem, so it is the most critical issue. One should then choose simple subnetworks to apply efficient solution methods at step 2. At step 3 aggregation can be exactly applied only for product-form networks, and it is approximated otherwise, in general with unknown error. Network decomposition can be very efficient when the isolated subnetworks at step 2 and the aggregated network at step 3 are simple to analyze. The various approaches determine the subnetwork parameters. Many approximate methods use iterative aggregation-disaggregation procedures, for which conditions and speed of convergence should also be considered. Few approximations have known accuracy. An open issue is the definition of approximate methods with known error, such as bound solutions.

Approximate method comparison. We now present a review and comparison of some approximate methods by considering their accuracy, efficiency and the class of models to which they can be applied. Specifically, we consider the algorithm rationale and the model assumptions, i.e., constraints on the network parameters such as topology, types of service distributions, queue capacities, and blocking type. The approximation accuracy is evaluated by comparing numerical results with either simulation or exact results [6, 10].

We shall now consider some significant approximations for the two classes of closed and open networks. Table 2 summarizes the conditions under which the methods for closed and open networks can be applied, i.e., the constraints on network topology, service centers (service time distribution, number of servers and queue capacity, and blocking type).

Table 2. Approximate methods for queuing networks with blocking

	Network costraints	
Methods for closed networks	topology - node type -	blocking types
Throughput Approximation (TA)	cyclic - G/M/1/B	BAS - BBS
Network Decomposition (ND)	cyclic - G/M/1/B	BBS
Variable Queue Capacity Decomp. (VQD)	cyclic[1] - G/M/1/B	BBS
Matching State Space (MSS)	general - G/M/1/B	BAS
Approximate MVA (AMVA)	general - G/M/1/B	BAS
Maximum Entropy Algorithm (ME)	general - G/GE/1/B	RS-RD
Methods for open networks		
Tandem Exponential Network Decomp. (TED)	tandem - G/M/1/B	BAS
Tandem Phase-Type Network Decomp. (TPD)	tandem - G/M/1/B	BAS
Acyclic Network Decomposition (AND)	acyclic - G/M/1/B	BAS
Maximum Entropy Algorithm (ME-O)	general - G/GE/1/B	RS-RD

4.1 Approximate Methods for Closed Networks with Finite Capacities

We consider the following six algorithms for closed queuing networks, based on various principles:

- Throughput Approximation (TA) [46]
- Network Decomposition (ND) [23]
- Variable Queue Capacity Decomposition (VQD) [56]
- Matching State Space (MSS) [1]
- Approximate MVA (AMVA) [2]
- Maximum Entropy Algorithm (ME) [32, 35]

They applied to homogeneous networks, i.e., each node has the same blocking type. We assume FCFS service discipline at each node. Table 3 reports the key idea of each approximation method.

Cyclic networks. The first three methods (TA, ND and VCD) evaluate the throughput of cyclic networks with exponential service time distribution. TA and VCD algorithm compute the network throughput $X(K)$ as a function of network population K.

Throughput Approximation (TA) applies to cyclic networks with BBS or BAS blocking and exponential service times [46]. It evaluates the network throughput, assuming that it is a symmetrical function, that is $X(K) = X(B - K)$, where $B = \sum_i B_i$. This property holds for BBS blocking as proved under the more general assumption of phase-type service distribution in [22], and it reaches its maximum value for $K = K^* = \lfloor \frac{B}{2} \rfloor$. The algorithm directly computes few values of function $X(K)$ with exact analytical methods and computes the other values by fitting the curve through those known points. For BAS blocking the symmetry property of the throughput does not hold, but a similar shape of the curve as for BBS blocking is conjectured, supported by experimental results, where K^* is approximated by an iterative scheme that depends on the queue capacities and the service rates. The main drawback of this method is the cumbersome computational complexity required to evaluate the exact throughput. Hence, it can be used for parametric analysis of the throughput by varying the network population and only for networks with a limited number of nodes and customers.

Network Decomposition (ND) approximates the throughput of the cyclic network with BBS blocking by a network decomposition method [23]. At step 1 the network is partitioned into N one-node subnetworks. At step 2 each subnetwork is analyzed in isolation as an $M/M/1/B_i$ network with arrival rate λ_i^* and load dependent service rate $\mu_i^*(n)$, $\forall n$, to derive the marginal queue length distribution $\pi_i^*(n)$, $\forall n$, $\forall i$. Parameters λ_i^* and $\mu_i^*(n)$ are defined by a set of equations and are approximated for each subnetwork. The isolated queue is approximated by taking into account the blocking of customers due to the finite capacity of the downstream nodes. The authors consider two cases depending on whether all the nodes have finite capacity or there is one infinite capacity node.

Table 3. Approximate methods for closed networks with blocking: main idea

Method	Key idea
TA	Exact model analysis for some network population and throughput interpolation by varying network population K.
ND	Network decomposition into nodes analyzed in isolation as M/M/1/B.
VCD	Network decomposition and aggregation into a single composite node with state dependent service rate and variable buffer size.
MSS	Analysis of the QN without blocking by choosing the network population to approximately match the state space cardinality.
AMVA	Modified and forced MVA algorithm to consider blocking.
ME	Approximate product-form for the queue length distribution based on maximum entropy.

They define the parameters by a fixed-point equation for λ_i^* and an iterative algorithm. It starts with a throughput approximate interval $[X_{min}(0), X_{max}(0)]$, computes new parameters λ_i^* and $\mu_i^*(n)$ at each step and appropriately updates the k-th throughput approximation $[X_{min}(k), X_{max}(k)]$, until a convergence condition is satisfied. Such conditions check the approximate throughput interval width, and some consistency control on the network. If all nodes have finite capacity an additional iteration cycle is required to compute probabilities $\pi_i^*(B_i)$ (see [23] for details). Convergence has not been proved, but it has been observed. The time computational complexity is of $O(kN^4B_{max}^3)$ operations, for k iteration steps.

Variable Queue Capacity Decomposition (VQD) method can be applied to cyclic[1] networks with BBS blocking [56], and we assume that node 1 has infinite capacity ($B_1 = \infty$). The algorithm is based on the network decomposition principle applied to nested subnetworks. The key idea is that given a node i, all the downstream nodes $(i + 1, \ldots, N)$ are aggregated in a single composite node C_{i+1} with load dependent service rate and a variable queue capacity. The approximation evaluates the composite node C_{i+1} parameters (load dependent service rate, and the fraction of time in which the queue capacity is n, given the network population). The algorithm starts with the analysis of the two-node subnetwork formed by the last two nodes $(N - 1, N)$ to define the composite aggregate node C_{N-1}, that is seen by node $N-2$. Then the algorithm goes backward from node $i = N - 2$ to node 1 eventually to the two-node network formed by $(1, C_2)$ that represents the entire aggregated network, and from which one obtains the approximated throughput. The analysis of each two-node network where the composite node has variable queue capacity (or variable buffer) is carried out by considering two corresponding two-node networks where a composite node has fixed buffer and infinite buffer, respectively (see [56] for details). The algorithm is very simple, non-iterative and its time computational complexity is of $O(NK^3)$ operations.

[1] With a node with unlimited capacity.

Arbitrary topology networks. The three methods (MSS, AMVA and ME) apply to arbitrary topology networks. MSS and AMVA methods assume networks with BAS blocking, exponential service time, and evaluate the network throughput. ME algorithm assumes RS-RD blocking, generalized exponential service time and evaluates the queue length distribution and average performance indices.

The basic idea of Matching State Space (MSS) method [1] is to approximate the behavior of the network with blocking with that of a network without blocking by choosing the population to approximately match the state space cardinality of the underlying Markov chain. The assumption is that the two networks with nearly the same state space cardinality should have similar throughputs. The algorithm defines a new network with infinite capacity queues and K' customers so that the state space cardinality of the underlying Markov process, say $C'(K')$, is nearly equal to that of the Markov process of the original network with K customers, $C(K)$. The algorithm determines K' to approximate the state space matching, that is to minimize the difference function $|C'(K') - C(K)|$. Then the network without blocking is analysed (see [1] for details). The algorithm implementation is simple and the time computational complexity is of $O(N^3 + NK^2)$ operations.

Approximate MVA (AMVA) [2] analyzes networks with BAS blocking and exponential service times by a modification of the MVA algorithm originally defined for product-form networks with unlimited queue capacities [49]. The MVA algorithm is based on Little's theorem and the arrival theorem. Note that the arrival theorem and the MVA algorithm, as defined for networks without blocking, cannot be immediately applied to networks with blocking. Let $T_i(n)$, $L_i(n)$ and $X_i(n)$ denote the average response time, mean queue length and throughput of node i when there are n customers in the network. For load independent service center the MVA is based on the following recursive scheme, for $1 \leq n \leq K$:

- $T_i(n) = \frac{1}{\mu_i}[1 + L_i(n-1)]$, $\forall i$
- $X_i(n) = ne_i/[\sum_j e_j T_j(n)]$, $\forall i$
- $L_i(n) = X_i(n) T_i(n)$, $\forall i$.

The approximation algorithm modifies the first equation trying to take into account blocking. In particular if node i is full, it cannot accept new customers and there is at least one node j blocked by node i, then approximation defines:

- $T_i(n) = \frac{1}{\mu_i} L_i(n-1)$
- $T_j(n) = \frac{1}{\mu_j}[1 + L_j(n-1)] + \frac{1}{\mu_i}(e_j p_{ji})/e_i$

For node i only the customers already in the node contribute to the average response time, while for the blocked node j the response time increases of a blocking time due to node i (see [2] for further details). The algorithm can be simply implemented and the time computational complexity is of $O(N^3 + kNK)$ operations where k is the number of iterations of the approximate iterative computation at step n.

Maximum Entropy Algorithm (ME) [32, 35] evaluates the queue length distribution and average performance indices of a network with RS-RD blocking and generalized exponential service time. The approximation is based on the principle of maximum entropy and is an extension of the algorithm defined for open networks and more general cases, such as multiclass networks and priorities [29–31, 33]. Let $a_i = \max(0, K - \sum_{j \neq i} B_j)$ be the minimum number of customers in node i. The algorithm approximates the joint queue length distribution $\pi(\mathbf{S})$ for each network state S by maximizing the entropy function

$H(\pi) = -\sum_{\mathbf{S}} \pi(\mathbf{S}) log(\pi(\mathbf{S}))$

subject to the following constraints

- normalization: $\sum_{\mathbf{S}} \pi(\mathbf{S}) = 1$
- u_i is the probability of more than a_i customers in i: $\sum_{n_i > a_i} \pi(n_i) = u_i$
- L_i is the mean queue length: $\sum_{n_i = a_i}^{B_i} h_i(n_i)\pi_i(n_i) = L_i$
- Φ_i is the probability that node i is full: $\sum_{n_i = a_i}^{B_i} f_i(n_i)\pi_i(n_i) = \Phi_i$

where $h_i(n_i) {=} min(0, n_i - a_i - 1)$ and $f(n_i) = max(0, n_i - B_i + 1)$. By the Lagrange's method of undetermined multipliers the algorithm determines an approximation of $\pi(\mathbf{S})$ that has the following product-form expression:

$$\pi(\mathbf{S}) = \frac{1}{Z} \prod_{i=1}^{N} x_i(n_i) y_i^{h_i(n_i)} z_i^{f_i} \tag{7}$$

where Z is a normalizing constant, $x_i(n_i) = 1$ if $n_i = a_i$, and $x_i(n_i) = x_i$ if $a_i < n_i \leq B_i$, and x_i, y_i and z_i are the Lagrangian coefficients corresponding to constraints above. The network cannot be decomposed into single nodes and the coefficients do not have a closed form expression. The algorithm approximates the closed network with a pseudo open network without exogenous departures and arrivals. This open network is analysed by the approximation based on the same principle applied to open networks, introducing an additional constraint on the average queue lengths $K = \sum_i L_i$ and slight modifications to derive the coefficients of formula (7). Then the coefficients are iteratively approximated. The algorithm details are given in [32, 35]. The time computational complexity of the algorithm depends on the algorithm for open networks and for the iterative approximation, with k iteration, is of $O(kN^2K^2)$ operations.

4.2 Approximate Methods for Open Networks with Finite Capacities

We consider the following algorithms for open queuing networks, as reported in Table 2 that shows the corresponding constraints on the network topology, the type of service centers and the blocking type:

- Tandem Exponential Network Decomposition (TED) [20]
- Tandem Phase-Type Network Decomposition (TPD) [48]
- Acyclic Network Decomposition (AND) [37]
- Maximum Entropy Algorithm for Open networks (ME-O) [35, 51]

All the algorithms are based on network decomposition and ME-O method on the maximum entropy. Decomposition define one-node subnetworks as $M/M/1/B$ queues by TED and AND, $M/Cox/1/B$ queue by the other algorithms.

Tandem networks. The TED [20] and TPD [48] algorithms approximate the throughput of the tandem network with BAS blocking by network decomposition. The network is partitioned into N one-node subnetworks $T(i)$, $1 \leq i \leq N$. Subnetwork $T(i)$ represents the isolated node i and is analyzed as an $M/M/1/B_i$ queue by TED and as an $M/PH_n/1/B_i$ queue by TPD (with phase-type service distribution). The method define appropriate parameters to derive marginal probability $\pi_i(n)$, $\forall n$ of each subnetwork $T(i)$. Since $T(1)$ and $T(N)$ correspond to the first and last node of the tandem network the first has arrival rate λ (exogenous arrival rate) and the last service rate μ_N. The remaining $2(N-1)$ unknowns have to be determined. The approximation is based on an iterative scheme to approximate the subnetworks unknown parameters(see [20] for details). TED algorithm requires $O(kNB_{max}^2)$ operations, where k is the number of iterations. The authors proved the algorithm convergence, and numerical results show that it is fast. TPD method solve subsystems $T(i)$ with a matrix-geometric technique and distinguish two cases depending on whether the first node has finite capacity. When all the nodes have finite capacity it has an additional iterative cycle to estimate the effective arrival rates (see [48]). Convergence has not been proved. the algorithm requires $O(k_1 \sum_{2 \leq i \leq N} k_i (N-i+1)^3 B_i^2)$ operations where k_i is the number of iterations to compute the arrival rate of system $T(i)$.

Acyclic and arbitrary topology networks. The last two methods AND and ME-O apply to more general topology networks and evaluate the queue length distribution and are respectively based on network decomposition and the maximum entropy principle.

The Acyclic Network Decomposition (AND) method [37] extends TED approximation to acyclic networks with exponential service time distribution and BAS blocking. Like TED, the approximation is based on a network decomposition into N single node subsystems $T(i)$. Each subsystem is analyzed as an $M/M/1/B_i$ system, but AND method defines a new set of equations to determine the subsystems parameters (service and arrival rates). If node i has U_i predecessor nodes, then each subsystem $T(i)$ receives arrivals from U_i exponential sources with unknown rates, one source from each predecessor j (i.e., any node j such that $p_{ij} > 0$). These rates are approximated by an iterative procedure. To this aim AND algorithm evaluates the probability that at arrival time at $T(i)$ from the j-th source there are n other nodes blocked by node i, and the probability that at the end of a service system $T(i)$ is empty. These probabilities appear in the new formulas defined for the unknown rates (see [37] for details). The $T(i)$ subsystems are eventually analyzed to derive marginal probabilities $\pi_i(n)$, $\forall n$, $\forall i$. The time computational complexity of AND is bounded by $O(kN[(U + B_{max})^2 + U^3 + 2^{U+1}])$, where k is the iteration number and $U = max_i U_i$.

Maximum Entropy Algorithm for Open networks (ME-O) approximation [35, 51] analyses a more general classes of networks with arbitrary topology, generalized exponential service time distribution, and RS-RD blocking. It is similar to the ME method by the same authors for closed networks, and the approximation is based on the maximum entropy. The open networks is decomposed into N subsystems $T(i)$, each analysed as $GE/GE/1/B$ nodes with appropriate parameters by considering blocking. The analysis of the i-th $GE/GE/1/B$ systems is based on an iterative scheme that computes: 1) the arrival rate by the traffic equations, 2) the probability that, at service completion time at i, node j is full, $\forall j$, 3) the queue length probability π_i defined by a product-form whose coefficient are the Lagrange multipliers corresponding to the constraints of the maximum entropy problem, and 4) the coefficient of variation of the interarrival time at $T(i)$. The iterative scheme is repeated until convergence of the coefficient of variations at step 4. The probability computation at step 2 requires the solution of non-linear system that can lead to numerical instability and problems of convergence, which, however, is rarely observed. There is no proof of convergence and uniqueness of the solution. See [35, 51] for details. The time computational complexity is of $O(\Omega^3)$, where Ω is the cardinality of the set of probabilities computed at step 2.

5 Algorithms Comparison

Table 4 shows a comparison of approximate methods for closed and open networks. It shows the performance indices evaluated by every method, their accuracy and efficiency.

For closed networks, approximation ND is more accurate than VCD and the difference increases with the number of network nodes. TA is more accurate than ND and its accuracy is more stable than that of ND as the number of network nodes increases. However, ND is more efficient than TA, which is limited to small networks. If $K < NB_{max}$ then VCD approximation is better than ND, while the opposite is true otherwise. VCD approximation is less efficient than ND for large network population K. Note that VCD and TA provide the throughput for all the network population from 1 to K. ND is based on a fixed-point iteration and can show some numerical instability. ND and AT apply to a more general class than VCD approximation.

By comparing methods MSS and AMVA, we observe that the former is more accurate and more efficient. The approximations are quite different, since their rationales are not related. They are stable and their accuracy seems to be independent of network parameters (N, μ_i and B_i), but dependent on the topology. Specifically they provide better results for central server networks and worse results for cyclic networks.

For open tandem exponential networks with BAS blocking the two approximation algorithms TED and TPD have nearly the same accuracy, with quite similar approximations, for sign and value. Their accuracy increases for small blocking probabilities, i.e., for networks with large B_i or large μ_i with respect to the arrival rate. The approximation accuracy of TPD is influenced by capacity queue unbalancing, while that of TED is affected by service rate unbalancing. TPD is slightly better than TED for high blocking probabilities. TED is certainly more efficient and has a simpler implementation than TPD, which can show numerical instability that can affect the algorithm convergence.

Table 4. Comparison of approximate methods for networks with blocking

Method	Index	Accuracy	Efficiency
TA	$X(K)$	Very good	Poor for $N > 5$
ND	X	Good	Good
VQD	$X(K)$	Good for $N \leq 4$	Fair
MSS	X_i	Fair	Good
AMVA	L_i, X_i, T_i	Fair for X	Very good
ME	L_i, X_i, T_i	Fair	Fair
TED	L_i, X_i, T_i, π_i	Very good	Very good
TPD	L_i, X_i, T_i, π_i	Very good	Slow for networks with all finite capacity nodes, fair otherwise.
AND	L_i, X_i, T_i, π_i	Very good	Very good
ME-O	L_i, X_i, T_i, π_i	Good	Fair

Finally, the maximum entropy methods, ME and ME-O, for closed open networks apply to the more general class of networks with arbitrary topology, generalized exponential service time distribution, and RS-RD blocking. The throughput accuracy of ME is not affected by the topology and the symmetry of network parameters (μ_i and B_i, $\forall i$), but it depends on the coefficient of variation of the service distributions. The approximation error grows with these coefficients of variation. The accuracy of the ME-O method decreases with the presence of cycles in the networks.

6 Application Examples of Networks with Blocking

Some equivalence, insensitivity and monotonicity properties of queueing networks with finite capacities have been proved [10, 12, 21, 22, 43, 44, 52, 57, 58].

Insensitivity properties lead to the identification of the factors that affect system performance. Monotonicity provides insights in the system behavior. It can be applied in parametric analysis to study the impact of various parameters (e.g., system load, buffer dimension) on system performance, to solve optimization problems or for bounding analysis. Equivalencies are defined in terms of state probability distribution π, average performance indices, or passage time distribution. Most of the equivalencies derive from the identity of the network processes. However, even if two networks have identical Markov processes, the meaning of corresponding states may be different. Then performance measures may be not equivalent, because the equivalence in terms of π does not necessarily lead to equivalence in terms of average performance indices.

Examples of equivalences are between networks with and without blocking that immediately leads to the extension of efficient computational solution algorithms defined for BCMP networks. Such equivalences hold for exponential networks with RS-RD blocking with reversible routing and product-form $F4$, and with arbitrary routing and product-form $F2$, for which an equivalent product-form network without blocking can be defined (see [10, 12]). Several equivalences can be defined between networks with different blocking types, and between homogeneous and non-homogeneous networks. Some examples are:

- BBS and RS types are equivalent for multiclass two-node networks with BCMP type nodes and class independent capacities,
- BAS is reducible to BBS for cyclic networks provided that node capacities are augmented by 1.
- for central server topology networks, BAS is reducible to BBS with node capacities B_i, $2 \leq i \leq M$, augmented by 1.
- BBS and RS-FD types are equivalent for networks with arbitrary routing, single class, with exponential nodes, load independent service rates and if condition 3 holds (single destination node) defined in Section. 3.2.
- Stop and Recirculate blocking are equivalent for multiclass open Jackson networks with class type fixed.

These results can be applied, for example, to define more efficient methods or to extend solution algorithms to more general classes of models of networks with different blocking types or network parameters. A detailed description of equivalence properties can e found in [10, 12, 21, 43].

A simple application. A simple application example is a store-and-forward packet switching network with virtual circuits modeled at level 3 in OSI reference model. Under independence assumptions, the window flow control can be represented by a closed cyclic queueing network with finite capacities and RS blocking. Under exponential assumptions and if the non empty conditions (condition 1) is satisfied, then product form solution (3) with formula $F2$ holds and we can apply Convolution algorithm or MVA to derive the performance indices, such as network throughput, delay, and buffer occupancy.

Another simple application of finite capacity networks to model communications and computer systems is shown in Figure 1 that represents an heterogeneous network with blocking modelling two computer systems connected through a communication link. We assume that nodes $C1$ and $C2$ represent computer CPU subsystem with RS-RD blocking, nodes $D1$ and $D2$ are computer disk subsystem with BAS blocking, nodes $N1$ and $N2$ are computer network access with BAS blocking, and nodes $N2$ and $N4$ communication links with BBS

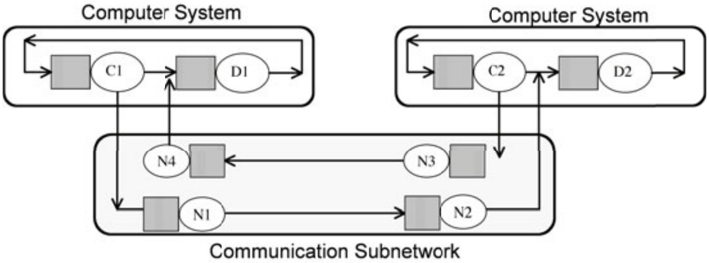

Fig. 1. Example

blocking. The customers of the model represent jobs (in computer systems) and packets (in communication subnetwork). Under exponential assumptions a heterogeneous network reducible to a homogeneous queing network with RS-RD blocking can represent the system. The network has arbitrary topology and we can can apply the ME approximate solution algorithm to derive the performance indices, such as the average response time and system throughput.

Moreover if nodes $D1$ and $D2$ have RS-RD blocking, then the network has product-form solution $F2$ and we can apply the convolution algorithm to evaluate the system performance.

References

[1] Akyildiz, I.F.: On the Exact and Approximate Throughput Analysis of Closed Queueing Networks with Blocking. IEEE Trans. Soft. Eng. 14, 62–71 (1988)

[2] Akyildiz, I.F.: Mean value analysis of blocking queueing networks. IEEE Trans. Soft. Eng 14, 418–429 (1988)

[3] Akyildiz, I.F., Von Brand, H.: Exact solutions for open, closed and mixed queueing networks with rejection blocking. J. Theor. Comp. Sci. 64, 203–219 (1989)

[4] Altiok, T., Perros, H.G.: Approximate analysis of arbitrary configurations of queueing networks with blocking. Ann. Oper. Res. 9, 481–509 (1987)

[5] Awan, I.U., Kouvatsos, D.D.: Approximate analysis of QNMs with space and service priorities. In: Kouvatsos, D.D. (ed.) Performance Analysis of ATM Networks, ch. 25, pp. 497–521. Kluwer, IFIP Publication (1999)

[6] Balsamo, S.: Closed Queueing Networks with Finite Capacity Queues: Approximate analysis. In: Proc. ESM 2000, SCS, Europ. Sim. Multiconf. Ghent, May 23-26 (2000)

[7] Balsamo, S., Clo', C., Donatiello, L.: Cycle Time Distribution of Cyclic Queueing Network with Blocking. Performance Evaluation 14(3) (1993)

[8] Balsamo, S., Clo', C.: A Convolution Algorithm for Product Form Queueing Networks with Blocking. Annals of Operations Research 79, 97–117 (1998)

[9] Balsamo, S., De Nitto, V.: A survey of Product-form Queueing Networks with Blocking and their Equivalences. Annals of Operations Research 48 (1994)

[10] Balsamo, S., De Nitto, V., Onvural, R.: Analysis of Queueing Networks with Blocking. Kluwer Academic Publishers, Dordrecht (2001)

[11] Balsamo, S., Donatiello, L.: On the Cycle Time Distribution in a Two-stage Queueing Network with Blocking. IEEE Trans. on Soft. Eng. 13, 1206–1216 (1989)

[12] Balsamo, S., Iazeolla, G.: Some Equivalence Properties for Queueing Networks with and without Blocking. In: Agrawala, Tripathi (eds.) Performance 1983. North-Holland, Amsterdam (1983)

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

[14] Boucherie, R., Van Dijk, N.: On the arrival theorem for product form queueing networks with blocking. Performance Evaluation 29, 155–176 (1997)

[15] Boxma, O., Konheim, A.G.: Approximate analysis of exponential queueing systems with blocking. Acta Informatica 15, 19–66 (1981)

[16] Brandwajn, A., Jow, Y.L.: An approximation method for tandem queueing systems with blocking. Operations Research 1, 73–83 (1988)

[17] Buzacott, J.A., Shanthikumar, J.G.: Design of Manufacturing Systems using Queueing Models. Queueing Systems: Theory and Applications (1992)

[18] Cheng, D.W.: Analysis of a tandem queue with state dependent general blocking: a GSMP perspective. Performance Evaluation 17, 169–173 (1993)

[19] Clo', C.: MVA for Product-Form Cyclic Queueing Networks with RS Blocking. Annals of Operations Research 79 (1998)

[20] Dallery, Y., Frein, Y.: On decomposition methods for tandem queueing networks with blocking. Operations Research 14, 386–399 (1993)

[21] Dallery, Y., Liu, Z., Towsley, D.F.: Equivalence, reversibility, symmetry and concavity properties in fork/join queueing networks with blocking. J. of the ACM 41, 903–942 (1994)

[22] Dallery, Y., Towsley, D.F.: Symmetry property of the throughput in closed tandem queueing networks with finite buffers. Op. Res. Letters 10, 541–547 (1991)

[23] Frein, Y., Dallery, Y.: Analysis of Cyclic Queueing Networks with Finite Buffers and Blocking Before Service. Performance Evaluation 10, 197–210 (1989)

[24] Gershwin, S.B.: An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. Oper. Res. 35, 291–305 (1987)

[25] Gordon, W.J., Newell, G.F.: Cyclic queueing systems with restricted queues. Oper. Res. 15, 286–302 (1967)

[26] Hillier, F.S., Boling, W.: Finite queues in series with exponential or Erlang service times - a numerical approach. Oper. Res. 15, 286–303 (1967)

[27] Hordijk, A., Van Dijk, N.: Networks of queues with blocking. In: Kylstra, K.J. (ed.) Performance 1981, pp. 51–65. North Holland, Amsterdam (1981)

[28] Jun, K.P., Perros, H.G.: An approximate analysis of open tandem queueing networks with blocking and general service times. Europ. Journal of Operations Research 46, 123–135 (1990)

[29] Kouvatsos, D.D.: Maximum Entropy Methods for General Queueing Networks. In: Potier (ed.) Proc. Modeling Tech. and Tools for Perf. Analysis, pp. 589–608. North-Holland, Amsterdam (1983)

[30] Kouvatsos, D.D.: A Universal Maximum Entropy Solution for Complex Queueing Systems and Networks. In: Karmeshu (ed.) Entropy Measures, maximum Entropy Principles and Emerging Applications, pp. 137–162. Springer, Heidelberg (2003)

[31] Kouvatsos, D., Awan, I.U.: Arbitrary closed queueing networks with blocking and multiple job classes. In: Proc. Third Int. Work. on Queueing Networks with Finite Capacity, Bradford, UK, July 6-7 (1995)

[32] Kouvatsos, D.D., Awan, I.U.: MEM for arbitrary closed queueing networks with RS blocking and multiple job classes. Annals of Oper. Res. 79, 231–269 (1998)

[33] Kouvatsos, D., Awan, I.U.: Entropy maximization and open queueing networks with priorities and blocking. Performance Evaluation 51, 191–227 (2003)

[34] Kouvatsos, D., Denazis, S.G.: Entropy maximized queueing networks with blocking and multiple job classes. Performance Evaluation 17, 189–205 (1993)

[35] Kouvatsos, D.D., Xenios, N.P.: MEM for arbitrary queueing networks with multiple general servers and repetitive-service blocking. Perf. Ev. 10, 106–195 (1989)

[36] Lam, S.S.: Queueing networks with capacity constraints. IBM J. Res. Develop. 21, 370–378 (1977)

[37] Lee, H.S., Bouhchouch, A., Dallery, Y., Frein, Y.: Performance Evaluation of open queueing networks with arbitrary configurations and finite buffers. In: Proc. Third Int. Work. on Queueing Networks with Finite Capacity, Bradford, UK, July 6-7 (1995)

[38] Mishra, S., Fang, S.C.: A maximum entropy optimization approach to tandem queues with generalized blocking. Perf. Evaluation 30, 217–241 (1997)

[39] Mitra, D., Mitrani, I.: Analysis of a Kanban discipline for cell coordination in production lines I. Management Science 36, 1548–1566 (1990)

[40] Onvural, R.O.: Some Product Form Solutions of Multi-Class Queueing Networks with Blocking. Perf. Evaluation 10(3) (1989)

[41] Onvural, R.O.: A Note on the Product Form Solutions of Multiclass Closed Queueing Networks with Blocking. Performance Evaluation 10, 247–253 (1989)

[42] Onvural, R.O.: Survey of Closed Queueing Networks with Blocking. ACM Computing Surveys 22(2), 83–121 (1990)

[43] Onvural, R.O., Perros, H.G.: On Equivalencies of Blocking Mechanisms in Queueing Networks with Blocking. Oper. Res. Letters 5, 293–298 (1986)

[44] Onvural, R.O., Perros, H.G.: Equivalencies Between Open and Closed Queueing Networks with Finite Buffers. Performance Evaluation (1988)

[45] Onvural, R.O., Perros, H.G.: Some equivalencies on closed exponential queueing networks with blocking. Performance Evaluation 9, 111–118 (1989)

[46] Onvural, R.O., Perros, H.G.: Throughput Analysis in Cyclic Queueing Networks with Blocking. IEEE Trans. Software Engineering 15, 800–808 (1989)

[47] Perros, H.G.: Queueing networks with blocking. Oxford University Press, Oxford (1994)

[48] Perros, H.G., Altiok, T.: Approximate analysis of open networks of queues with blocking: tandem configurations. IEEE Trans. Soft. Eng. 12, 450–461 (1986)

[49] Raiser, M., Lavenberg, S.S.: Mean Value Analysis of closed multi-chain queueing networks. Journal of ACM 27, 217–224 (1989)

[50] Sereno, M.: Mean Value Analysis of product form solution queueing networks with repetitive service blocking. Performance Evaluation 36-37, 19–33 (1999)

[51] Skianis, C.A., Kouvatsos, D.D.: Arbitrary open queueing networks with service vacation periods and blocking. Annals of Operations Research 79, 143–180 (1998)

[52] Shanthikumar, G.J., Yao, D.D.: Monotonicity Properties in Cyclic Queueing Networks with Finite Buffers. In: Perros, Altiok (eds.) First Int. Work. on Queueing Networks with Blocking. North Holland, Amsterdam (1989)

[53] Akyildiz, Perros (eds.): Special Issue on Queueing Networks with Finite Capacity Queues. Performance Evaluation, vol. 10(3). North Holland, Amsterdam (1989)

[54] Onvural, R.O. (ed.): Special Issue on Queueing Networks with Finite Capacity. Performance Evaluation, vol. 17(3). North-Holland, Amsterdam (1993)

[55] Balsamo, S., Kouvatsos, D.: Special Issue on Queueing Networks with Blocking Performance Evaluation Journal, vol. 51(2-4). North Holland, Amsterdam (2003)

[56] Suri, R., Diehl, G.W.: A variable buffer size model and its use in analytical closed queueing networks with blocking. Management Sci. 32(2), 206–225 (1986)

[57] van Dijk, N.: On stop = repeat servicing for non-exponential queueing networks with blocking. J. Appl. Prob. 28, 159–173 (1991)

[58] van Dijk, N.: Stop = recirculate for exponential product form queueing networks with departure blocking. Oper. Res. Lett. 10, 343–351 (1991)

[59] Van Dijk, N.: Queueing networks and product form. John Wiley, Chichester (1993)

[60] Yao, D.D., Buzacott, J.A.: Modeling a Class of State Dependent Routing in Flexible Manufacturing Systems. Annals of Oper. Research 3, 153–167 (1985)

[61] Yao, D.D., Buzacott, J.A.: Modeling a class of flexible manufacturing systems with reversible routing. Oper. Res. 35, 87–93 (1987)