CLOSED QUEUEING NETWORKS WITH FINITE CAPACITY QUEUES: APPROXIMATE ANALYSIS

Simonetta Balsamo

Department of Computer Science, University of Venice
Via Torino 155, Mestre - Venezia, Italy
balsamo@dsi.unive.it

ABSTRACT

Queueing networks with finite capacity queues and blocking are applied to model systems with finite resources and population constraints, such as computer and communication systems, as well as traffic, production and manufacturing systems. Various blocking types can be defined to represent different system behaviors. When a customer attempts to enter a full capacity queue blocking occurs. The analysis of queueing networks with finite capacity queues is often based on approximate methods and simulation, since exact analytical techniques cannot be applied because of the synchronization constraint, except for a few special cases. Various approximate analytical methods have been proposed in literature and provide a solution in terms of average performance indices such as throughput and mean response time. These methods have different characteristics including model assumptions and constraints, type of blocking, algorithm accuracy and efficiency. We present a comparison of some significant approximate methods to analyze closed queueing networks with finite capacity queues. By experimental results we identify the condition under which one can appropriately select a given solution method. Experimental comparisons have been performed by the Queueing Networks with Blocking Analyzer (QNBA), a software tool developed for modelling and analysis of queueing network models with finite capacity queues and blocking.

1. INTRODUCTION

Queueing networks with finite capacity queues are used to model systems with finite capacity resources and population constraints, such as computer and communication systems, as well as traffic, production and manufacturing systems. System performance analysis can be carried out through such queueing network models to derive of a set of figures of merit, such as queue length distribution, mean response time and throughput. In queueing networks with finite capacity queues when a queue reaches its maximum capacity then the flow of customers from other service centers into this queue is stopped, and blocking occurs. Various blocking types have been defined and analyzed in the literature to represent distinct behaviors of real systems with limited resources (Akyildiz and Perros 1989, Onvural 1990, Perros 1994, Van Dijk 1991a, Van Dijk 1991b). Some comparisons and equivalences among blocking types have been presented for queueing networks with various network topologies in (Balsamo and De Nitto Personè 1991, Balsamo and De Nitto Personè 1994, Balsamo 1994, Onvural and Perros 1989a, Onvural 1990, Perros 1994, Van Dijk 1991a, Van Dijk 1991b). Exact analytical methods cannot be generally applied to analyze queueing networks with finite capacity queues, except for a few special cases, because of the synchronization constraints between the service centers of the network. Hence recourse to approximate analytical methods or simulation is necessary. In this paper we consider approximate analytical method for closed queueing networks with finite capacity queues and blocking.

Various approximate analytical methods have been proposed in literature and provide a solution in terms of average performance indices such as throughput and mean response time. However these methods have different characteristics including model assumptions and constraints, type of blocking, algorithm accuracy and efficiency. Most of these approximate methods provide a solution with a limited computational cost and without any bound on the introduced approximation error. The methods are usually validated by comparing numerical results with either simulation results or exact solutions. Many 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 exact analytical product-form solution.

We present a comparison of some significant approximate methods to analyze closed queueing networks with finite capacity queues. By experimental results we identify the conditions under which one can appropriately select an approximate solution method. Experimental comparisons have been performed by the Queueing Networks with Blocking Analyzer (QNBA) (Rainero 1997), a software tool developed for modelling and analysis of queueing network models with finite capacity queues and blocking.

In the next section we introduce closed queueing networks with finite capacity queues, the blocking types and the model analysis. We present the approximate methods for closed networks with blocking in Sections 3. The experimental results and the comparison of approximate methods are discussed in Section 4.

2. CLOSED QUEUING NETWORK WITH FINITE CAPACITY QUEUES AND BLOCKING

Consider a single class closed queueing network model formed by M service centers with constant number of customers N circulating in it. Let $P=[p_{ij}]$ $(1\leq i,j\leq M)$ denote the routing matrix where p_{ij} is the probability that a job leaving node i tries to enter node j. For each service center we define the number of servers, the service time distribution, the queue capacity and the service discipline. We consider the following service time distributions: exponential (M), phase-type with n exponential stages (PH_n), general (G) and generalized exponential distribution (GE). Let $\mathbf{S}_i$ denote the state of node i, which includes the number of customers in node i, denoted by n_i, and other components depending both on the node type (service discipline and service time distribution) and the blocking type. Let μ_i denote the service rate of node i. Let B_i denote the maximum number of customers admitted at node i, that is in the queue and in the servers (B_i=c+s), $1\leq i\leq M$. Thus the total number of jobs in node i satisfies the constraint $n_i\leq B_i$. When the queue reaches the finite capacity ($n_i=B_i$) the node is said to be full and blocking occurs.

Various blocking types have been defined to represent different system behaviors. We now recall three of the most commonly used blocking types (Akyildiz and Perros 1989, Onvural 1990, Onvural 1993, Perros 1994).

• *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.

• *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,

This work has been partially supported by MURST Research Fundsv.

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 can be defined depending on whether the server can be used as a buffer when the node is blocked: *BBS-SO* (server occupied) and *BBS-SNO* (server is not occupied) Hereafter we consider *BBS-SO* blocking that is simply called *BBS*.

• *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 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) and *RS-FD* (fixed destination).

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. In the following we shall consider deadlock-free queueing networks in steady-state conditions.

Under general assumptions the queueing network model with finite capacity can be represented by a Markov process. Let $\mathbf{S}=(\mathbf{S}_1,\ldots,\mathbf{S}_M)$ denote the state of the network 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 Q. Let $\pi = \{\pi(\mathbf{S}), \mathbf{S}\in E\}$ denote the steady-state queue length probability distribution, that can be obtained by solving the homogeneous linear system of the global balance equations

$$\pi Q = \mathbf{0} \qquad (1)$$

subject to the normalising condition $\Sigma_{\mathbf{S}\in E}\, \pi(\mathbf{S}) =1$. The definition of state space E and transition rate matrix Q depends on the network definition and on the blocking type of each node (Balsamo and De Nitto Personè 1991, Balsamo and De Nitto Personè 1994, Onvural 1990, Onvural 1993, Perros 1994, Van Dijk 1991a, Van Dijk 1991b). From vector π one can derive π_i, the queue length distribution of node i and other average performance indices of node i, such as throughput (X_i), the average queue length (L_i) and the mean response time (R_i). 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. Indeed the time computational complexity of liner system (1) is determined by the space state E cardinality that grows exponentially with the buffer sizes ($B_i\le N$, $1\le i\le M$) and M. Hence the problem is numerically untractable as the number of model components grows.

In a few special cases, queueing networks with finite capacity queues can be solved by exact analytical methods, when they have a product-form solution of vector π, under particular constraints and for various blocking types. A survey of product-form solutions of networks with finite capacity queues is presented in (Balsamo and De Nitto Personè 1994) and an efficient algorithms for some closed product-form networks with blocking is given in (Balsamo and Clò 1995).

However, general queueing networks with blocking do not have a product-from solution and recourse to approximate analytical methods or simulation is necessary.

3. APPROXIMATE ANALYSIS

Several approximate techniques to analyze closed queueing networks with finite capacity queues have been proposed to evaluate average performance indices and queue length distributions (Onvural 1993, Perros 1989). Most of the methods provide an approximate solution with a limited computational cost, but 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.

Most approximate methods are heuristics based on the *decomposition principle* applied to the underlying Markov process or, more often, directly to the network. 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.

Once the network partition is defined, the analysis of the original network reduces to the analysis of each subnetwork at step 2 and of the aggregated one at step 3. Network decomposition can be very efficient when the isolated subnetworks at step 2 and the aggregated network are simple to analyze. For product-form networks the aggregation theorem provides exact results, i.e., the aggregated network is equivalent to the original model in terms of queue length probability and average performance indices. For non product-form networks the aggregated network only approximates the original model, and no bound is known on the approximation error. Various heuristics have been defined by taking into account both the network model characteristics and the blocking type (Akyildiz and Perros 1989, Dallery and Frein 1989, Frein and Dallery 1989, Gordon and Newell 1976, Kouvatsos and Xenios 1989, Kouvatsos and Denazis 1993, Kouvatsos and Awan 1995, Onvural 1990, Onvural 1993, Perros 1994, Suri and Diehl 1986, Yao and Buzacott 1987). A bounded aggregation technique has been defined for Markov processes and applied to queueing networks with blocking in (Courtois and Semal 1986) by exploiting the special structure of the underlying Markov process.

The definition of network decomposition and the evaluation of the approximation error are two critical issues for the approximate methods based on decomposition. Some techniques apply an iterative aggregation-disaggregation procedure, for which conditions and speed of convergence should also be considered, as in (Dallery and Frein 1989). Some approximations are obtained by forcing exact aggregation for product-form queueing networks with infinite capacity (Onvural and Perros 1989b). Other approximation algorithms are based on a product-form solution defined by the maximum entropy principle (Kouvatsos and Xenios 1989, Kouvatsos and Denazis 1993, Kouvatsos and Awan 95).

We shall now survey some significant approximation methods for closed networks with finite capacity queues. We classify the approximate methods by considering model assumptions, i.e. constraints on the network parameters such as topology, types of service distributions and blocking type. We briefly recall each algorithm and discuss its rationale. In the next section we present a performance comparison of some approximate methods by considering their accuracy, efficiency and the class of models to which they can be applied. We consider the following six algorithms based on various principles:

- Throughput Approximation (Onvural and Perros 1989b)
- Network Decomposition (Frein and Dallery 1989)
- Variable Queue Capacity Decomposition (Suri and Diehl D86)
- Matching State Space (Akyildiz 1988a)
- Approximate MVA (Akyildiz 1988b)
- Maximum Entropy Algorithm (Kouvatsos and Xenios 1989).

Table I shows the conditions under which the methods can be applied, and in particular the constraints on network topology, service centers (i.e. service time distribution, number of servers and queue capacity) and blocking type. Note that all algorithms are applied to homogeneous networks, i.e. each node has the same blocking type. We assume FCFS service discipline at each node. Last column of Table I gives the key idea of each approximation method.

Let $B=\Sigma_i B_i$ denote the total network capacity and let $B^+=\max_{1\le i\le M} B_i$ and $B^-=\min_{1\le i\le M} B_i$ respectively denote the maximum and minimum node capacity of the network.

3.1 Approximate Methods For Cyclic Networks

The first three methods (Onvural and Perros 1989b, Frein and Dallery 1989, Suri and Diehl 1986) evaluate the throughput of cyclic networks with exponential service time distribution. The first and the third algorithm compute the throughput as a function of network population.

Throughput Approximation
This method applies to cyclic networks with BBS or BAS blocking. It is based on the assumption that the throughput is a symmetrical function of the population network. Let X(N) denote the network throughput when there are N customers in it. Throughput is symmetrical if X(N)=X(B-N). This property holds for BBS blocking as proved under the more general assumption of phase-type service distributions and in (Dallery and Towsley 1991). Moreover the throughput reaches its maximum value for $N=N^*$, where $N^*=\lfloor B/2 \rfloor$ if B is even and $N^*=\lfloor B/2 \rfloor, \lfloor B/2 \rfloor+1$ if B is odd. Function X(N) is non-decreasing for $1 \le N \le N^*$ and non-increasing function for $N^* \le N < B$.
Hence the algorithm directly computes few values of function X(N) with exact analytical methods and computes the other values by fitting the curve through those known points, by the following function

$$X(N)=X(N+1)-y\, x^{N^*-N} \qquad (3.1)$$

where

$$y= [X(N^*)-X(B^-)]/\Sigma_{i=1,\dots,(N^*-B^-)}\, x^i \qquad (3.2)$$

and x is the fixed-point of the following equation:

$$X(B^- -1)=$$

$$X(B^-)-[X(N^*)-X(B^-)][x^{N^*-B^-+1}(1-x)]/[x-x^{N^*-B^-+1}] \qquad (3.3)$$

To overcome a throughput underestimation, a correction to formula (3.1) is proposed by adding a factor $c_N y$, where coefficient c_N is simply defined (see (Onvural and Perros 1989 b, Perros 1994) for details).
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. However, N^* depends on queue capacities and service rates and is approximated by $\lceil \Sigma_i (B_i+1)/2 \rceil -1$. In this case it is necessary to directly evaluate more values of X(N) that for BBS and to compute the approximation also for $N^* \le N \le B-2$. This affects the algorithm efficiency and accuracy with respect to the case of BBS blocking.
The algorithm has the following structure:

1. Exact computation of X(N) for $N=B^- -1$, B^-, N^*. Since the network for $1 \le N \le B^-$ is without blocking, we can apply an algorithm for product-form networks without blocking. The exact evaluation of $X(N^*)$ requires the solution of the associated Markov chain, i.e. of linear system (1).
2. Approximate computation of X(N) for $B^- +1 \le N \le N^*-1$. These values are approximated by formulas (3.1), (3.2) and the solution of the fixed-point problem (3.3).

For BAS blocking X(N) is not symmetrical and the algorithm has two additional steps:

3. Exact computation of X(N) for N=B-1, B. X(B) is evaluated as the average time between two successive deadlocks which are immediately detected and resolved; this requires a numerical integration; X(B-1) is approximated by a function of X(B) or it can be directly computed (see (Onvural and Perros 1989b, Perros 1994)).
4. Approximate computation of X(N) for $N^*+1 \le N \le B-2$, as at step 2.

The algorithm is based on an iterative scheme for the fixed-point problem (3.3). Although convergence has not been proved, it has been observed. Space and time computational complexity of the algorithm mainly depends on the direct computation of $X(N^*)$ at step 1 obtained by the Markov chain analysis. As discussed in Section 2, the process state space cardinality grows exponentially with N and M and seriously affects the applicability of this method.

Network Decomposition
The throughput of the cyclic network with BBS blocking is approximated by a network decomposition method. In the first step the network is partitioned into M 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 $\mu^*_i(n)$, $0 \le n \le B_i$ to derive the marginal queue length distribution $p^*_i(n)$, $0 \le n \le B_i$, $1 \le i \le M$. This aggregation procedure does not provide exact results for this blocking network and the analysis of 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, denoted by 1, i.e. $B_1=\infty$. In the former case they propose the following definition of the parameters:

$$\mu^*_i(n)= \{(1/\mu_i) + \sum_{i=1}^{M} b_{ij}(n)[\sum_{k=i+1}^{j}(1/\mu_k)]\}^{-1}$$

$$1 \le i \le M-1,\ 1 \le B_i \le M-1 \qquad (3.4)$$

$$\mu^*_M(n)=\mu_M,\ 1 \le n \le B_M \qquad (3.5)$$

$$\lambda_i^*= X/(1-p^*_i(B_i)) \qquad (3.6)$$

where X is the network throughput and $b_{ij}(n)$ denotes the probability that nodes i+2,…,j are full and node j+1 is not full, given n customers in node i, $1 \le i,j \le M$. This probability is expressed in terms of probabilities $p^*_k(B_k)$ for $i+1 \le k \le j+1$. Moreover $p^*_i(B_i)$ in formula (3.6) is a function of λ_i^*. Hence given the throughput X the arrival rate λ_i^* can be computed as the solution of the fixed-point equation (3.6).
These formulae are the basis of the iterative algorithm that starts with a throughput approximate interval $[X_{min}^{(0)}, X_{max}^{(0)}]$. At the k-th step ($k \ge 1$) it computes new parameters λ_i^* and $\mu^*_i(n)$, $0 \le n \le B_i$, $1 \le i \le M$, by formulas (3.4)-(3.6) and appropriately updates the k-th throughput approximation $[X_{min}^{(k)}, X_{max}^{(k)}]$. The iterative scheme continues until a convergence condition is satisfied. These conditions include a control of the approximate interval width, i.e. $(X_{max}^{(k)}-X_{min}^{(k)})<\_$ for a small_, a consistency control, that is the summation of all the average node population is close to N, and $\lambda_i^*<\mu^*_{i-1}$,_i for the convergence of the fixed-point equation (3.6). If all nodes have finite capacity ($B_1<\infty$), formula (3.5) does not hold and $\mu^*_M(n)$ is computed by an expression similar to (3.4) and an additional iteration cycle is required to compute probabilities $p^*_i(B_i)$, $\forall i$ (see (FD89) for details). Like the previous algorithm, convergence has not been proved, but it has been observed. One can show that the time computational complexity is of $O(kM^4(B^+)^3)$ operations for k iteration steps.

Variable Queue Capacity Decomposition
This method can be applied to cyclic networks with BBS blocking and where one node 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,…,M} are aggregated in a single composite node C_{i+1} with load dependent service rate and a *variable queue capacity*,. This special definition of variable queue capacity allows to overcome the classical definition of composite node with constant capacity, so improving the approximation. The approximation evaluates the composite node C_{i+1} parameters, that are the load dependent service rate denoted by $\nu_{i+1}(n)$ and the fraction of time in which the queue capacity is n, given N customers in the network denoted by $f_{i+1}(n|N)$, $1 \le n \le N$.
The algorithm starts with the analysis of the two-node subnetwork formed by {M-1,M} to define the composite aggregate node C_{M-1}, that is seen by node M-2. Then the algorithm goes backward from node i=M-2 to node 1 to the analysis of the two-node subnetwork formed by {i, C_{i+1}} to define the composite aggregate node C_i. At the last step the two-node network formed by {1, C_2} represents the entire aggregated network and one obtains the approximated throughput. Each two-node network with a composite node with

variable queue capacity (or variable buffer) (VB) is analyzed by considering two corresponding two-node networks with a composite node with fixed buffer (FB) and with infinite buffer (IB), respectively. Parameters $\nu_{i+1}(n)$ and $f_{i+1}(n|N)$ of the VB network are derived by the solution of the two corresponding FB and IB networks, whose analysis is well-known and quite simple. In particular the VB network parameters and state probabilities are defined as a weighted sum of the FB parameters and state probabilities; these are in turn approximated by using the IB model solution. The algorithm details are given in (Suri and Diehl 1989). The algorithm is very simple, non iterative and its time computational complexity is of $O(MN^3)$ operations.

3.2 Approximate Methods For Arbitrary Topology Networks

The three methods (Akyildiz 1988a, Akyildiz 1988b, Kouvatsos and Xenios 1989) apply to arbitrary topology networks. The first two methods assume BAS blocking and exponential service time and evaluate the network throughput. The third method assumes RS-RD blocking, generalized exponential service time and evaluates the queue length distribution and average performance indices.

Matching State Space

This method approximates the throughput of a network with BAS blocking and exponential service times. The basic idea 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.

Let M denote the Markov chain associated to the network with M nodes and blocking. Let K(N) denote the state space cardinality of M when there are N customer in the network. The algorithm defines a new network with infinite capacity queues and N' customers with underlying Markov chain M'. Let $K'(N')=\binom{M+N'-1}{M-1}$ be the state space cardinality of M' when there are N' customer in the network. The algorithm determines N' to approximate equation K(N)_K'(N'), that is to minimize the difference function $|K(N)-K'(N')|$. Since one can observe that $K(N) \le K'(N)$ _N then $N' \le N$. Finally, the network without blocking is analysed. The algorithm, whose details are given in (Akyildiz 1988a), has the following structure:

1. Computation of K(N) by a convolution algorithm.
2. Determine N' to minimize $|K(N)-K'(N')|$, $1 \le N' \le N$, by linear search in [1,N].
3. Computation of the throughput of the network without blocking by a convolution algorithm.

The algorithm implementation is simple and the time computational complexity is of $O(M^3+MN^2)$ operations.

Approximate MVA

Network with BAS blocking and exponential service times are analyzed by a modification of the MVA algorithm originally defined for product-form networks with unlimited queue capacities. The MVA algorithm is based in the Little theorem and the arrival theorem Raiser and Lavenberg 1980). Let $R_i(n)$, $L_i(n)$ and $X_i(n)$ denote respectively the average response time, mean queue length and throughput of node i when there are n customers in the network and let e_i be the mean number of customers at node i. For load independent service center MVA is based on the following recursive scheme:

$$R_i(n)= (1/\mu_i)\,[1+L_i(n-1)] \qquad 1\le i\le M \qquad (3.7)$$

$$X_i(n)= n\, e_i/[\Sigma_{1\le j\le M}\, e_j R_j(n)] \qquad 1\le i\le M \qquad (3.8)$$

$$L_i(n)= X_i(n)R_i(n) \qquad 1\le i\le M \qquad (3.9)$$

for n=1,…, N, with $L_i(0)=0$, $1\le i\le M$. This algorithm is based on the arrival theorem defined for product-form networks that does not apply to networks with blocking (Balsamo 1994).

The approximation algorithm modifies formula (3.7) 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. The approximate algorithm modifies relation (3.7) for the full node i and for the blocked node j as follows:

$$R_i(n)= (1/\mu_i)\, L_i(n-1) \qquad (3.10)$$

$$R_j(n)= (1/\mu_j)\,[1+L_j(n-1)]+BT_i\,(e_j p_{ji}/e_i) \qquad (3.11)$$

where $BT_i= (1/\mu_i)$. For node i only the customers already in the node contribute to the average response time, while in the blocked node j the time increase of a blocking time due to node i. Then the modified MVA algorithm works as follows (see (Akyildiz 1988b) for details):

1. Initialization.
2. For each population n=1,…N

   repeat

   computation of the MVA equations by formulas (3.7)-(3.9), where (3.7) is substituted by (3.10) or (3.11) for full and blocked nodes, respectively

   until $L_i(n) \le B_i$ for each node i.

The algorithm can be simply implemented and the time computational complexity is of $O(M^3+kMN)$ operations where k is the number of iterations of the internal cycle at step 2.

Maximum Entropy Algorithm

This method 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 an algorithm defined for open networks. It has successively be extended to multiclass exponential networks (Kouvatsos and Denasis 1993, Kouvatsos and Awan 1995).

Let $a_i=\max\{0, N-B+B_i\}$ denote the minimum number of customer in node i, $1\le i\le M$. The algorithm approximates the joint queue length distribution $\pi(\mathbf{S})$ for each network state $\mathbf{S}=(n_1,\dots, n_M)$ by maximizing the entropy function $H(\pi)=-\Sigma_{\mathbf{S}}\, \pi(\mathbf{S})\log(\pi(\mathbf{S}))$ subject to the following constraints:

(I) (normalization) $\Sigma_{\mathbf{S}}\, \pi(\mathbf{S})=1$

(II) (u_i is the probability that node i has more than a_i customers) $\Sigma_{n_i>a_i}\pi_i(n_i)=u_i$

(III) (L_i is the mean queue length) $\Sigma_{a_i\le n_i\le B_i} h_i(n_i)\pi_i(n_i)=L_i$

(IV) (Σ_i is probability that node i is full) $\Sigma_{a_i\le n_i\le B_i} f_i(n_i)\pi_i(n_i)=\Phi_i$

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 approximation $\pi(\mathbf{S})$ that has the following product form expression:

$$\pi(\mathbf{n})=(1/Z)\prod_{i=1}^{M} x_i(n_i)\, y_i^{h_i(n_i)} z_i^{f_i} \qquad (3.12)$$

where Z is a normalizing constant, $x_i(n_i)=1$ if $n_i=a_i$ and $x_i(n_i)=x_i$ if $a_i\le n_i\le B_i$, and x_i, y_i and z_i are the Lagrangian coefficients corresponding to constraints (II)-(IV). The network cannot be decomposed into single nodes and coefficients x_i, y_i and z_i do not have a closed form expression. The algorithm approximates the closed network with a *pseudo* open without exogenous departures and arrivals. This open network is analysed by the approximation based on the same principle by adding the constraint on the average queue lengths $N =\Sigma_i L_i$ and with slight modifications to derive a solution for x_i, y_i and an approximation for z_i, Σi. Then coefficients z_i are iteratively evaluated.

The algorithm, whose details are given in (Kouvatsos and Xenios 1989), has two phases, as follows:

1. Analysis of the corresponding pseudo open network with the approximation for open networks slightly modified to derive coefficients x_i, y_i and a approximation for z_i, _i.
2. Iteratively evaluate coefficients z_i by applying a convolution algorithm that computes network throughputs.

The time computational complexity of the algorithm for step 1 depends on the algorithm for open networks and for step 2 is of $O(kM^2N^2)$ operations where k is the number of iterations.

4. EXPERIMENTAL RESULTS AND COMPARISON OF APPROXIMATE METHODS

In this section we present the experimental resultss and the comparison between the approxiamte algorithm for closed queueing networks with finite capcaty queues. In order to evaluate the approximation accuracy, we have performed an extensive numerical study, since no a priori error bound is known. In particular we have performed a parametric analysis by varying the model components, i.e. the number of service centers (M), the number of customers (N) for closed networks, the network topology, the service rates (μ_i, $1 \le i \le M$) and finite capacities (B_i, $1 \le i \le M$). To study whether the symmetry of the network parameters affects the approximation accuracy we have considered various classes of networks with identical and with different service center parameters, i.e., service rates and finite capacities. If the approximate algorithm is based on particular properties, they have been validated through a specific test. We have implemented the approximate methods that have been integrated in the software tool QNBA (Queueing Networks with Blocking Analyzer). QNBA has been developed for the specification and analysis of queueing networks with finite capacity queues and population constraints (Rainero 1997). The accuracy of the approximate methods has been validated by comparing numerical results with either simulation or exact results. Exact solution of the networks with blocking has been obtained by QNBA with Markov chain analysis or product-form expression, while simulation results have been obtained by the RESQ/IBM package (Kurose et al. 1982). The accuracy is evaluated by the relative error. We define REL as the percentage relative error of the approximate result (A) and the exact or simulation result (B) defined as follows: REL=[(A-B)/A]100%. A-REL denotes the average of percentage relative errors, where given a set $\mathbf{A}=\{A_i\}$ of the approximate results and the corresponding relative errors $\{REL_i\}$, we define A-REL=$\Sigma_i\, REL_i/|\mathbf{A}|$. Max-REL is the maximum of the percentage relative errors, that is M-REL=$\max_i REL_i$.

We have implemented and evaluated each algorithm on a class of networks by varying the parameters up to M=10 nodes and N=150 customers, with service rates ranging in [0.01,12] customers per unit of time and finite queue capacities ranging in [1, 40]. Both symmetrical and non-symmetrical networks have been considered, i.e. with identical or different service rates and/or queue capacities.
Table II summarises some observations of the performance comparison of the six approximate methods. Table III IV and V shown the numerical results, i.e. the average and maximum relative errors. From the experimental results we can derive the following observations.
For *Throughput Approximation* (TA) we have observed a very good accuracy with an average relative error always within 2% and a maximum relative error always within 5% both for BAS and BBS blocking. For both blocking types we observed that the error increases when the number of customers is close to B^-. The width of the interpolation values does not seem to affect the approximation accuracy. The approximation underestimates or overestimates the real throughput without any apparent regularity. The main drawback of this method is the cumbersome computational complexity required to evaluate the exact throughput at step 2. Hence, as observed by the authors, this method 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.
For *Network Decomposition* (ND) we have observed a good accuracy with an average relative error always within 3% and a maximum relative error almost always within 7%. The average relative error is affected by the number of nodes but it seems to be independent of the other network parameters (service rate and capacity unbalancing). Like the previous algorithm, the error increases when the number of customers is close to B^- and we observed both cases of underestimation and overestimation of the real throughput.

For *Variable Queue Capacity Decomposition* (VQCD) we have observed accurate results for small networks (i.e. with M=3,4), but an average relative error even greater than 7-10% as the number of nodes increases. We observed maximum relative errors up to 34%, although most of the relative errors were within 15%. By the algorithm definition it was expected that the accuracy worsens as the network dimension increases, because the final composite node C_1 represents M-1 aggregated nodes and each aggregation step introduces an approximation error. The accuracy is not affected by the number of customers, but we observed underestimation of the throughput for small populations and overestimation for large ones. The average relative error seems to be independent of the unbalancing between node service rates and/or queue capacities, as observed in some specific test with stressed parameter values.

Comparison of Throughput Approximation, Network Decomposition and Variable Queue Capacity Decomposition
By comparing the numerical results of the three approximation algorithms we can derive some observations on accuracy, efficiency and generality:

- Network Decomposition (ND) is more accurate than Variable Queue Capacity Decomposition (VQCD) for both the average and the maximum relative error. This difference increases with the number of network nodes.
- Throughput Approximation (TA) is more accurate than ND for both the average and the maximum relative error. TA accuracy is more stable that ND as the number of network nodes increases.
- ND is more efficient than TA, which is limited to small networks.
- the time computational complexities of ND, given by $O(kM^4(B^+)^3)$, and of VQCD, $O(MN^3)$, show a different dependence on network parameters. We observe that if $N<MB^+$ then VQCD approximation is better than ND, while the opposite is true otherwise. This confirms that VQCD approximation is less efficient than the ND for large N.
- VQCD and TA provide the throughput for all the network population from 1 to N.
- ND is based on a fixed-point iteration and can show some numerical instability, as observed for a network with M=20 service centers.
- ND and AT apply to a more general class than VQCD.

For *Matching State Space* (MSS) numerical evaluation does not show a good accuracy. Although we have observed an average relative error of 4.6%, the maximum relative error is more than 25% even for networks with few service centers. This can be explained by observing that the basic assumption of the approximation usually is not verified. The algorithm tries to approximate the state space cardinality independently of the model structure. Hence the algorithm provides the same approximation for all those networks with different parameters that have the same number of feasible states. We have observed a good accuracy only for networks with very small state spaces, while the error increases with the state space cardinality. There is a high variability in the observed errors and we cannot say whether the approximation under or overestimates the throughput. By comparing the results for various classes of topology, we have observed the smallest error for central server networks and the worst case for cyclic networks.
For *Approximate MVA* (A-MVA) Numerical experiments have shown small values of k. The algorithm does not show a good accuracy: the average relative error is over 10% for mean response time and about 8% for the throughput. The maximum relative error is more than 55% for mean response time and more than 38% for the throughput, even for networks with few service centers. Moreover for mean response time we have observed a high variability in the relative errors of for different service centers in the same network (e.g. from 1% to 20%), while the approximation behavior is more regular for the throughput. The observed approximation error is not affected by the number of network nodes, but it depends on the network topology. Like the previous algorithm we have observed better results for central server networks and the worst case for cyclic networks.

Comparison of Matching State Space and Approximate MVA

We can compare the throughput obtained by the two approximation algorithms Matching State Space and Approximate MVA since they apply to the same class of networks.

- Matching State Space (MSS) is more accurate than Approximate MVA (A-MVA) both in terms of average and maximum relative errors.
- The approximations are quite different, since their rationales are not related.
- Both approximations seem to be independent of network parameters (number of nodes, service rates and queue capacities), but dependent on the topology. Specifically they provide better results for central server networks and worse results for cyclic networks.
- MSS is more efficient than A-MVA.
- Both algorithms are stable.

For *Maximum Entropy Algorithm* (ME) numerical experiments we have observed a good accuracy for the average relative error of about 3.8% for the mean queue length and 3.9% for the throughput. However, we have observed relevant maximum relative errors up to 41% and 59% for throughput and mean queue length, respectively. The throughput accuracy is not affected by the topology and the symmetry of network parameters and. The observed accuracy depends on the coefficient of variation (c.v.) of the service distributions, i.e. the error increases as the c.v. increase. This is due to the approximation introduced by the algorithm for open networks applied at step 2 that modifies the c.v. when it is different from 1, corresponding to exponential case. Moreover, we observed that the approximate results do not necessarily satisfy the balance equations of the throughput and often find the exact solution when applied to product-form networks. We point out that this is the only method that provides an approximation of the queue length distribution. For this performance index we have observed an average relative error of about 10% a relevant maximum relative error, but a maximum absolute error of 0.21. Finally, we have observed some problems of convergence for cyclic networks with asymmetric parameters (service rates and queue capacities).

5. CONCLUSIONS

We have compared some approximate methods for queueing networks with finite capacity queues and blocking. We have discussed the conditions under which each method can be applied, the algorithm efficiency and accuracy through a parametric analysis based on a numerical study by varying several network parameters.

REFERENCES

(Akyildiz 988a) Akyildiz, I.F. 1988. "On the exact and approximate throughput analysis of closed queueing networks with blocking." *IEEE Trans. on Soft. Eng.*, no.14: 62-71.

(Akyildiz 988b) Akyildiz, I.F. 1988. "Mean value analysis of blocking queueing networks." *IEEE Trans. on Soft. Eng.,* no. 14: 418-129.

(Akyildiz and Perros 1989) Akyildiz, I.F., and H.G. Perros.1989. *Special Issue on Queueing Networks with Finite Capacity Queues, Performance Evaluation*, vol. 10, n. 3.

(Balsamo and De Nitto Personè 1991) Balsamo, S., and V. De Nitto Personè. 1991. "Closed queueing networks with finite capacities: blocking types, product-form solution and performance indices." *Performance Evaluation*, vol.12, n.4: 85-102.

(Balsamo and De Nitto Personè 1994) Balsamo, S., and V. De Nitto Personè. 1994. "A Survey of product-form queueing networks with blocking and their equivalences." *Annals of Operations Research*, no. 48: 31-61.

(Balsamo 1994) Balsamo, S., 1994. "Properties and analysis of queueing network models with finite capacities." in *Performance Evaluation of Computer and Communication Systems*, L. Donatiello and R. Nelson Eds., Lecture Notes in Computer Science, no. 729, Springer-Verlag.

(Balsamo and Clò 1995) Balsamo, S. and C. Clò. 1995. "A Convolution Algorithm for Product-Form Queueing Networks with Blocking." *Annals of Operations Research, Baltzer,,* no. 79: 97-117.

(Courtois and Semal 1986) Courtois, P.J. and P. Semal. 199("Computable bounds for conditional steady-state probabilities in larg Markov chains and queueing models." *IEEE Journal on SAC* , vol. 4 no. 6: 920-936.

(Dallery and Frein 1989) Dallery, Y. and Y. Frein. 1989. "/ decomposition method for the approximate analysis of close queueing networks with blocking." In *Proceedings of First In Workshop on Queueing Networks with Blocking*, (H.G. Perros and T Altiok Eds.) North Holland

(Dallery and Towsley 1991) Dallery, Y. and D.F. Towsley. 1991 "Symmetry property of the throughput in closed tandem queuein networks with finite buffers." *Op. Res. Letters,* no. 10: 541-547.

(Frein and Dallery 1989) Frein, Y. and Y. Dallery. 1989. "Analysis c cyclic queueing networks with finite buffers and blocking befor service." *Performance Evaluation*, no. 10: 197-210.

(Gordon and Newell 1976) Gordon, W.J. and G.F. Newell. 197("Cyclic queueing systems with restricted queues." *Oper. Res.,* no. 1: 286-302.

(Kouvatsos and Xenios 1989) Kouvatsos, D. and N.P. Xenios. 198("MEM for arbitrary queueing networks with multiple general server and repetitive-service blocking." *Performance Evaluation,* no. 10, 3.

(Kouvatsos and Denasis 1993) Kouvatsos, D. and S.G. Denazis. 199: "Entropy maximized queueing networks with blocking and multipl job classes." *Performance Evaluation*, no. 17 :189-205.

(Kouvatsos and Awan 1995) Kouvatsos, D. and I.U. Awan. 199: "Arbitrary closed queueing networks with blocking and multiple jo classes." In *Proceedings of Third International Workshop o Queueing Networks with Finite Capacity,* (Bradford, UK, 6-7 July).

(Kurore and al. 1982) Kurose, J.F.; E.A. MacNair; and C.H. Saue 1982. "The Research Queueing package." Technical Report RA 13(IBM, NY.

(Onvural 1990) Onvural, R.O. 1990. "Survey of Closed Queuein Networks with Blocking." *ACM Computing Surveys*, vol. 22, vo. 2 83-121.

(Onvural and Perros 1989a) Onvural, R.O. and H.G. Perros. 198("Some equivalencies on closed exponential queueing networks wit blocking." *Performance Evaluation.*, vol.9: 111-118.

(Onvural and Perros 1989b) Onvural, R.O. and H.G. Perros. 198("Throughput analysis in cyclic queueing networks with blocking. *IEEE Trans. on Software Eng.* SE-15: 800-808.

(Onvural 1993) Onvural, R.O. 1993. *Special Issue on Queuein Networks with Finite Capacity, Performance Evaluation,* vol. 17, nc 3.

(Perros 1994) Perros, H. 1994. *Queueing Networks with Blockin* Oxford.

(Raiser and Lavenberg 1980) Raiser, M. and .S.S. Lavenberg. 198("Mean Value Analysis of closed multi-chain queueing networks. *Journal of ACM*, no. 27: 217-224.

(Rainero 1997) Rainero, A. 1997. "Analisis approssimata di modelli rete di code con blocco." Laurea Thesis in Computer Scienc University of Udine (in italian).

(Suri and Diehl 1986) Suri, R. and G.W. Diehl. 1986. "A variable buff size model and its use in analytical closed queueing networks wit blocking." *Management Sci.* , no..32, 2: 206-225.

(Van Dijk 1991a) Van Dijk, N. 1991. "On 'stop = repeat' servicing fc non-exponential queueing networks with blocking." *J. Appl. Prob* vol. 28 :159-173.

(Van Dijk 1991b) Van Dijk, N. 1991. "'Stop = recirculate' fc exponential product form queueing networks with departur blocking." *Oper. Res. Lett..,* no. 10: 343-351.

(Yao and Buzacott 1987) Yao, D.D. and J.A. Buzacott. 1987. "Modelin a class of flexible manufacturing systems with reversible routing. *Oper. Res.,* no. 35: 87-93.

Method	Network Constraint		Blocking Type	Approximation key idea
	Topology	Node type		
Throughput Approximation	cyclic	G/M/1/B	BAS or BBS-SO	Exact model analysis for some network population. Interpolation of the throughput values by varying network population
Network Decomposition	cyclic	G/M/1/B	BBS-SO	Network decomposition into single nodes analyzed in isolation as M/M/1/B queues
Variable Queue Capacity Decomposition	cyclic a node with unlimited capacity	G/M/1/B	BBS-SO	Network aggregation of the set of finite capacity queue nodes in a single composite node having state dependent service rate and variable buffer size
Matching State Space	general	G/M/1/B	BAS	Analysis of the network with unlimited queue capacity and by choosing the network population to approximately match the same state space cardinality
Approximate MVA	general	G/M/1/B	BAS	Modification of the MVA algorithm to take into account blocking
Maximum Entropy Algorithm	general	G/GE/1/B	RS-RD	Approximate product-form for the queue length distribution

Table I - Approximate methods for closed queueing networks with blocking.

Method	Performance Indices	Accuracy	Efficiency
Throughput Approximation	X(N): network throughput as a function of the network population	Very good	Poor for networks with more than 5 nodes
Network Decomposition	X: network throughput	Good	Good
Variable Queue Capacity Decomposition	X(N): network throughput as a function of the network population	Good for networks with up to 4 nodes, inaccurate otherwise.	Fair
Matching State Space	X_i: node throughput	Fair	Good
Approximate MVA	L_i: mean queue length X_i: node throughput T_i: node mean response time	Fair for throughput, poor for other performance indices	Very good
Maximum Entropy Algorithm	π_i: queue length distribution L_i: mean queue length X_i: node throughput R_i: node mean response time	Fair for all the performance indices.	Fair

Table II - Performance comparison of approximate methods.

M	TA BBS		TA BAS		ND $B_1=\infty$		ND $B_1<\infty$		VQCD	
	Max REL	A REL	Max REL	A REL	Max REL	A REL	Max REL	A REL	Max REL	A REL
3	3.1	1.6	2.3	1.7	4.8	1.6	2.2	0.9	2.5	1.4
5	4.2	2.1	3.4	1.8	5	1.8	6.6	2.4	11.2	5.1
7	3.8	1.7	4.5	2.0	.0	1.9	4.1	2.2	25.4	7.2
8	2.9	1.6	3.1	1.4	3.4	2.5	4.8	2.6	18.0	8.7
10	4.6	2.0	4.0	1.9	5.1	5.7	6.1	3.8	34.2	10.9
total	4.6	1.92	4.5	1.87	8.7	2.7	6.6	2.4	34.2	7.1

(a) Queuing networks with cyclic topology

M	MSS		A-MVA X_i		A-MVA L_i		ME X_i		ME L_i	
	Max REL	A REL	Max REL	A REL	Max REL	A REL	Max REL	A REL	Max REL	A REL
3	2.8	1.1	18.2	4.9	29.2	8.5	16.6	3.6	24.2	4.0
5	25.6	5.3	37.3	9.0	55.9	14.5	28.4	4.7	32.6	3.3
7	14.4	5.2	9.1	2.8	28.4	9.7	15.8	3.2	32.2	3.2
8	26.9	6.0	16.0	5.3	9.1	6.0	41.4	3.7	58.9	3.9
10	17.1	5.3	38.8	11.2	44.1	12.3	21.3	4.0	48.2	3.7
total	26.9	4.6	38.8	7.8	55.9	10.1	41.4	3.92	58.9	3.78

(b) Queueing networks with general topology

Table III: Average and maximum relative error for network population: (a) cyclic topology, (b): general networks

service rates	queue capacities	TA BBS	TA BAS	ND $B_1=\infty$	ND $B_1<\infty$	VQ-CD	MSS	A-MVA X_i	A-MVA L_i	ME X_i	ME L_i
balanced	balanced	1.8	1.9	3.1	1.8	5.6	3.6	7.1	9.9	3.7	2.8
balanced	unbalanced	1.5	1.6	2.8	2.7	7.2	4.3	7.6	9.8	3.9	4.3
unbalanced	balanced	2.0	1.9	2.4	2.7	6.7	3.9	8.1	10.1	4.0	4.1
unbalanced	unbalanced	1.9	2.1	2.5	2.4	8.0	4.9	7.5	11.3	4.1	4.3

Table IV: Average relative error for parameters unbalancing

network topology	MSS	A-MVA X_i	A-MVA L_i	ME X_i	ME L_i
cyclic	7.1	11.6	13.9	3.9	3.3
central server	2.8	5.1	5.0	3.8	3.1
arbitrary	3.9	6.9	11.6	4.0	5.2

Table V: Average relative error for network topology