QUEUEING NETWORKS WITH FINITE CAPACITIES

Raif O. Onvural, IBM, Research Triangle Park, NC 27709

1. INTRODUCTION

Queueing theory has been developed to predict the performance characteristics of computer systems, production systems, communications systems, and flexible manufacturing systems. A queueing system consists of a number of nodes connected according to the topology of the system being modeled. Each node consists of a service facility and a queue for customers to wait when the service facility can not provide service to these customers.

Queueing networks may be classified as open and closed. In an open model, customers arrive at the network from outside, receive service at one or more nodes, and eventually leave the network. In closed queueing networks, there is a fixed population of customers circulating in the network at all times. Both types of networks have been frequently used in the literature as models of complex service systems.

Queueing networks have been studied in the literature under a multiplicity of assumptions. Most performance metrics of interest can be obtained from the joint steady state queue length distribution of the model. Almost all queueing networks of practical importance can, in theory, be solved numerically. However, in practice, obtaining the solution may not always be possible. In particular, the number of equations can easily reach millions, even for simple models, making it infeasible to solve the linear system of equations. Other two methods developed to analyze such networks are simulation and heuristics. Two major problems with simulations are determining how close the simulation results are to the exact values and how long to run a simulation to obtain accurate estimates. Heuristics have been developed to obtain the performance metrics of interest approximately when it is expensive to use the other methods. The main problem with heuristics is to bound the error in the solution. Perhaps, one of the most pioneering results in queueing networks is that under certain assumptions on the system parameters, it has been shown that the joint steady state queue length distribution of a class of queueing networks is the product of some particular functions of nodes. Product form networks are relatively easier to study and various algorithms have been developed in the literature to obtain the exact values of their performance metrics of interest.

Almost all queueing networks with product form queue length distributions require infinite queues, that is, it is assumed that there is always a space in the queue for arriving customers. In real systems, the storage space is always finite. Hence, a more realistic model of such systems requires modeling finite node capacities. An important feature of queueing networks with finite queues is that the flow of customers through a node may be momentarily stopped when another node in the network reaches its capacity. That is, a phenomenon called *blocking* occurs. In particular, consider a simplistic view of a computer communications system. The individual queues represent the finite space which is available for intermediate storage and servers correspond to communication channels. A message may not be transmitted until the destination node has space available to store the message, thus, sometimes causing the blocking of communication to that node. Similarly, in production systems, intermediate storage areas have finite capacities. A unit completing its service at a station may be forced to occupy the machine until there is a space available in the next station. While the unit blocks the machine, it may not be possible for the machine to process other units waiting in its queue.

In addition to the problem of blocking, *deadlocks* may occur in queueing networks with finite queues.A set of nodes is in a deadlock state when every node in the set is waiting for a space to become available at another node in the set. In this case, corresponding servers are blocked, and they can never get unblocked because the required space required will never be available.

Queueing networks with blocking are difficult to solve: in general, their steady state queue length distributions could not be shown to have product form solutions. Hence, most of the techniques that are employed to analyze these networks are in the form of approximations, simulation, and numerical techniques. In recent years, there has been a growing interest in the development of computational methods for the analysis of both open and closed queueing networks with blocking. A comprehensive survey of the literature on open queuing networks with blocking was compiled by Perros [1989] and on closed queueing networks with blocking by Onvural [1989] in addition to the two workshops and two special issues in this area (cf. Perros and Altiok [1988], Akyildiz and Perros [1989], and Onvural and Akyildiz [1992a] and [1992b]).

1.1. Blocking Mechanisms

The effect of a full node on its downstream nodes depends on the type of the system being modeled. To model different characteristics of various real life systems with finite resources, various blocking mechanisms that define distinct models of blocking have been reported in the literature. In particular, each blocking mechanism defines when a node is blocked, what happens during the blocking period, and how a node becomes unblocked. The most commonly used blocking mechanisms are classified as follows:

1.1.1. Blocked After Service (BAS)

A customer upon completion of its service at node i attempts to enter destination node j. If node j at that moment is full, the customer is forced to occupy server i until it enters destination node j, and node i is blocked. Node i remains blocked for this period of time, and server i can not serve any other customer which might be waiting in its queue. In queueing networks with arbitrary topologies, it is possible that a number of nodes may be blocked by the same node simultaneously. This necessitates imposing an ordering on the blocked nodes to determine which node will be unblocked first when a departure occurs from the blocking node. This problem has not been elaborated on in the literature. We are only aware of the *First-Blocked-First-Unblocked* rule (FBFU), which states that the node which was blocked first will be unblocked first. It is possible that deadlocks might occur in queueing networks under BAS blocking. It was assumed in the literature that deadlocks in networks under BAS blocking can be detected immediately and resolved by instantaneously exchanging blocked customers. Lemma 1 below gives the necessary and sufficient condition for a network under BAS blocking to be deadlock free, where a *cycle* in a network is defined as a directed path that starts and ends at the same node.

Lemma 1: A closed queueing network with K customers under BAS blocking is deadlock free if and only if for each cycle, C, in the network $K < \sum_{j \in C} B_j$*, where* B_j *is the capacity of node j.*

Simply stated, the total number of customers in the network must be smaller than the sum of node capacities in each cycle.

The BAS blocking (also referred to as type 1 blocking, manufacturing blocking, classical blocking, and transfer blocking) has been used to model systems such as production systems and disk I/O subsystems. In particular, consider a simplistic view of a production system consisting of a sequence stations. The availability of a storage space in the next station has no effect on the operation of a machine until it completes its service. Upon service completion, if there is no space available in the destination station then there are two possibilities: the unit which completed its service at node i is 1) moved back from the i-th machine to the storage area of machine i (i.e. its queue) allowing a unit to enter service, or 2) allowed to occupy the machine, blocking the operation for other units waiting in the storage space (i.e. BAS blocking). In most cases, it may not be possible to move the unit from the machine back to the storage area due to the physical constraints of the unit, the system, or both. Hence, the operation of the machine is blocked until there is a space available at the destination station. An exception to this is considered in Mitra and Mitrani [1988] where a blocked customer is moved back to its queue in the context of open networks to model the Japanese Kenban scheme, used for cell coordination in production lines.

1.1.2. Blocked Before Service (BBS)

A customer at node i declares its destination node j before it starts receiving its service. If node j is full, the i-th node becomes blocked. When a departure occurs from the destination node j, node i is unblocked and its server starts serving the customer. If the destination node j becomes full during the service of a customer at node i, then the service is interrupted and node i is blocked. The service is resumed from the interruption point as soon as a space becomes available at the destination node. Depending upon whether the customer is allowed to occupy the service area when the server is blocked, the following sub-categories are distinguished.

***BBS-SNO** (Server is Not Occupied):* Service facility of a blocked node can not be used to hold a customer.

***BBS-SO** (Server is Occupied):* Service facility of a blocked node is used to hold a customer.

In BBS blocking mechanism, a full node j blocks all nodes i that are connected to it (i.e. p_{ij}>0). When a departure occurs from node j, all blocked nodes become unblocked simultaneously and start serving their customers. Hence, there is no need to impose any ordering on the blocked nodes, unlike BAS blocking.

The BBS blocking mechanism (also called type 2 blocking, immediate blocking, service blocking, communications blocking) is motivated by considering servers which only move customers between stations and do no other work on them. In this case, the lack of downstream space must force the server to shut down.

The distinction between BBS-SO and BBS-SNO blocking mechanisms is meaningful when modeling different types of systems. For example, in communication networks, a server correspond to a communication channel. If there is no space in the downstream node then messages can not be transmitted. Furthermore, the channel itself can not be used to store messages due to physical constraints of the channel, i.e. BBS-SNO blocking. On the other hand BBS-SO blocking results if the service facility can be used to hold the blocked customer which, in this case, would be an approximate modeling of the system.

BBS-SO blocking has been used to model manufacturing systems, terminal concentrators, mass storage systems, disk to tape back up systems, window flow control mechanisms, and communication systems. Modeling these systems with BBS-SO blocking assumes that when its destination buffer is full the device is forced to stop its operation and the service facility can be used to hold a customer. A disk to tape back up model illustrated in figure 1 comprised of three servers, and two finite buffers between servers. The first server is the disk and channel which transfers blocks of data from from the disk to the main memory. The second server, the Central Processing Unit (CPU), transfer data from the main memory to the tape drive. The last server represents the tape drive. One of the performance objectives of interest is the tape back up rate (i.e. the throughput of the system). Blocking occurs due to finite spaces available for intermediate storage.

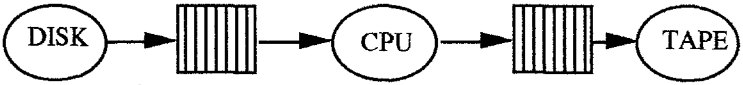


Figure 1: Disk to Tape Backup System

A simple terminal concentrator consists of a number of terminals, a concentrator, and a channel to transfer data to the main memory. The system configuration is the same as the disk to tape backup system illustrated above in figure 1 with the concentrator, the channel and the CPU replacing the disk, the CPU and the tape respectively. The two buffers in this terminal concentrator system have finite capacities which cause blocking of respective nodes. We note that the above examples are only sub-systems of larger configurations of computer systems, used only to illustrate the possibility of blocking due to finite storage capacities between the devices of such systems.

Queueing networks under BBS blocking are not always well defined for arbitrary topologies with an arbitrary number of customers in the network. This is because deadlocks in this blocking mechanism can not be resolved without violating its rules. As an example, let us assume that node i is blocked by node j and node j is blocked by node i. Then, the services at both nodes are suspended. Furthermore, the service can not start unless the blocking mechanism is temporarily switched to, for example, BAS blocking. In view of this, this blocking mechanism can only be used in deadlock free networks. Similar to lemma 1, it can be shown that a closed queueing network with BBS blocking is deadlock free if and only if for each cycle C in the network, i) $K< \sum_{j\in C}\{B_j-1\}$ in BBS-SNO blocking, and ii) $K< \sum_{j\in C} B_j$ in BBS-SO blocking. That is, a closed network under BBS-SO blocking is deadlock free if the number of customers in the network is less than the sum of node capacities in each cycle in the network, while in BBS-SNO, a network is deadlock free if the number of customers in the network is less than the sum of queue capacities (node capacity minus one for the server facility) in each cycle.

1.1.3. Repetitive Service (RS)

A customer upon service completion at node i attempts to join destination node j. If node j at that moment is full, the customer receives another service at node i. This is repeated until the customer completes a service at node i at a moment that the destination node is not full. Within this category of blocking mechanisms, the following two sub-categories are distinguished:

RS-FD (Fixed Destination): Once the customer's destination is determined it can not be altered.

RS-RD (Random Destination): A destination node is chosen at each service completion independently of the destination node chosen the previous time.

Similar to BBS-SO blocking, a network under RS-FD blocking is deadlock free if and only if $K< \sum_{j\in C} B_j$ for each cycle in the network, i.e. the number of customers in the network is less than the capacity of each cycle in the network. On the other hand, it can be shown that a network under RS-RD blocking is deadlock free if there is a path from every node to every other node in the network and, in case of closed networks, if there is at least one free space in the network, i.e. not for each cycle. This is because the existence of a free space in the network guarantees that all blocked customers will eventually depart, unblocking their servers.

The RS blocking (also called rejection blocking and type 3 blocking) arise in modeling telecommunication systems and it is mostly associated with reversible queueing networks. In particular, let us consider a packet switching network with fixed routing. The number of packets in the network is controlled by a window flow mechanism. A node transmits a packet to a destination node and waits for an acknowledgement. If the destination node does not accept the packet due to the fact that there is a lack of space, it will not send an acknowledgement. In this case, the packet may be retransmitted (RS-FD blocking) until it is accepted by the destination node (i.e.until an acknowledgement is received by the sender). Similarly, consider a manufacturing system consisting of a network of automated work stations (WS) linked by a computer controlled material handling device (MHD) to transport work-pieces that are to be processed from one station to another as illustrated in figure 2.

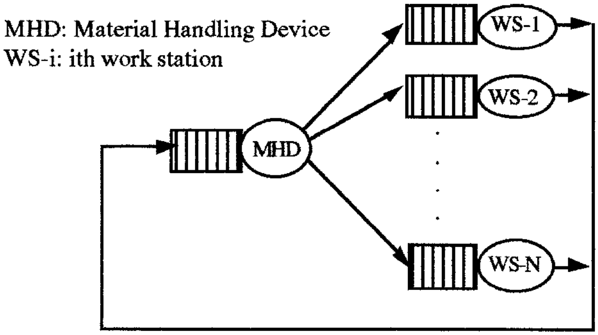


Figure 2: A Queueing Model of a Flexible Manufacturing System

In these systems, if a work-piece finds the next station full, then it has to wait for the next turn of the MHD. At the next turn, there are two possibilities: i) the work-piece can only be processed by one station, therefore, the next attempt can only be made to the previously chosen station (i.e. RS-FD blocking), or ii) the unit may be processed by all stations, hence, the next station is chosen independent of the previous choice(s) (i.e. RS-RD blocking). If the service time of the MHD is assumed to be exponentially distributed, the RS-RD blocking is equivalent to the following: The work-piece attempts to enter station 1, if station 1 is full, then it tries station 2, and so on, until a space is found in one of the stations.

1.2. Equivalencies of Blocking Mechanisms

Comparisons between these distinct types of blocking mechanisms have been carried out to obtain an equivalence between different blocking mechanisms applied to the same network. Two blocking mechanisms are said to be equivalent if the network under consideration has the same rate matrix under the both types of blocking mechanisms. We note that, all of the equivalences obtained in the literature assume that the service times are exponentially distributed. Furthermore, these equivalences are most often true only for specific topologies: cyclic networks and the central server model shown in figure 3 (Onvural (1987) and Balsamo, et al. (1986)).

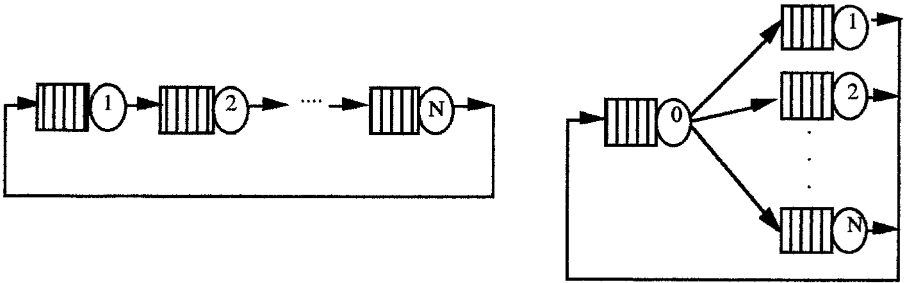


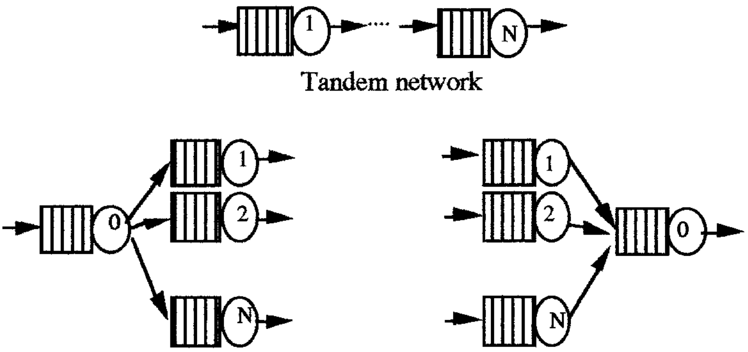
Figure 3: Cyclic Network and the Central Server Model

In the following two tables, the notation A ≡ B is used to denote that blocking mechanism A is equivalent to blocking mechanism B.

Topology	*Equivalencies*
Arbitrary	*BBS-SO ≡ RS-FD* *BBS-SO ≡ BBS-SNO for* $K \leq min\{B_i+B_j;\ i,j=1,\ldots,N\ s.t.\ p_{ij}>0\}-1$
Cyclic	*RS-FD ≡ RS-RD* *BBS-SNO with* B_i *≡ BAS with* B_i *-1,* $i=1,\ldots,N$
Central Server	*BBS-SO ≡ BBS-SNO if* $B_1=\infty$ *BBS-SO ≡ BBS-SNO if* $B_i=\infty$, $i=2,\ldots,N$ *RS-FD ≡ RS-RD if* $B_i=\infty$, $i=2,\ldots,N$
Two node cyclic	*BBS-SNO ≡ BBS-SO ≡ RS-FD ≡ RS-RD* *BBS-SNO with* B_1 *and* B_2 *≡ BAS with* B_1 *-1 and* B_2 *-1*

Table 1: Equivalencies of blocking mechanisms in closed networks

Similarly, equivalencies between different blocking mechanisms in open networks with finite capacity queues are given in table 2 for tandem, split, and merge configurations illustrated respectively in figure 4.



Tandem network

Split configuration

Merge configuration

Figure 4:Tandem, Split and Merge Configurations

Topology	*Equivalencies*
Arbitrary	*BBS-SO ≡ RS-FD*
Tandem	*BBS-SO ≡ RS-FD ≡ RS-RD* *BAS with B_i ≡ BBS-SNO with B_i +1, i=1,...,N*
Split	*BBS-SO ≡ BBS-SNO ≡ RS-FD if B_0=∞*
Merge	*RS-RD ≡ RS-FD* *BBS-SNO ≡ BBS-SO ≡ RS-RD ≡ RS-FD if B_i = ∞, i=1,...,N*
Two node tandem	*BBS-SNO ≡ BBS-SO ≡ RS-FD ≡ RS-RD* *BBS-SNO with B_1 = ∞ and B_2 ≡ BAS with B_1 = ∞ and B_2 -1*

Table 2: Equivalencies of blocking mechanisms in open networks

2. EXACT RESULTS

In special cases, networks with finite capacities are shown tò have product form solutions. Furthermore, various exact results are obtained on the behavior of various performance measures of interest. These results are reviewed next.

2.1. Closed Networks

In this section, we present various exact results obtained in the literature for closed queueing networks with blocking.

2.1.1. Reversible Networks

A stochastic process X(t) defined on the state space S is ***reversible*** if $\{X(t_1),...,X(t_n)\}$ has the same distribution as $\{X(t_0-t_1),...,X(t_0-t_n)\}$ for all $t_0, t_1,...,t_n$ ϵT. All results in this sub-section can be explained as a consequence of the following lemma given in Kelly [1979].

Lemma 2: *Consider a reversible Markov process with state space S and equilibrium distribution $\pi(j)$. If S is truncated to a sub-space A, then the Markov process is still reversible in equilibrium and it has distribution $\pi(j)/\sum_{k\in A}\pi(k)$, k $\in$A.*

A Markov process T(t) is called a ***truncated process*** of X(t) if it is irreducible and the state space of T(t) is a subset of the state space of X(t), i.e. the states of T(t) is a subset of X(t), and the transitions between the states of T(t) are the same as they are in X(t).

The ***routing in a network is reversible*** if the following condition holds for each subchain:

$$e_{ir}p_{ir;js}=e_{js}p_{js;ir} \qquad \text{for all i,j,r,s}$$

Two examples of reversible networks are the two-node cyclic networks and the central server model with BCMP type nodes.

Lemma 3: *A two-node cyclic network with node capacities B_i, i=1,2, BCMP type nodes and multiple classes of jobs has a product form equilibrium distribution under BAS, BBS-SO, BBS-SNO, RS-RD, and RS-FD blocking.*

In the cases of BBS-SO, RS-RD, or RS-FD blocking, we have the following lemma.

Lemma 4: *A closed queueing network under RS-RD blocking with node capacities B_i, multiple classes of jobs, BCMP type nodes, and state space A is a truncated process of the same network with infinite buffer capacities in which no more than B_i*

jobs are allowed at node i. Furthermore, if the underlined Markov process of the network with infinite buffer capacities is reversible then the blocking network has a product form equilibrium distribution.

A closed queueing network under BAS blocking has a product form equilibrium distribution if the number of jobs in the network is equal to the minimum buffer capacity plus one. If this condition is met, then there can be at most one node in the network blocked at any time, and, the blocked node has exactly one job, i.e. the server's operation is not blocked. In this case, the server of a blocked node behaves like an additional buffer capacity to the blocking node.

Lemma 5: *Consider a multi-class closed queueing network under BAS blocking. If* $\sum_{i=1}^{R} K_r = min\{B_i,\ i=1,...,N\}+1$ *then the network has a product form equilibrium distribution.*

2.1.2. Self Dual Networks

Closed queueing networks under BBS-SO blocking were first studied by Gordon and Newell [1967] in the context of cyclic networks. The service time at each node is assumed to be exponentially distributed. First, we will discuss the concept of holes as introduced by Gordon and Newell. Since the capacity of node j is B_j, let us imagine that this node consists of B_j cells. If there are i_j customers at node j, then i_j of these cells are occupied and B_j- i_j cells are empty. We may say that these empty cells are occupied by holes. Then, the total number of holes in the network is equal to $\sum_{j=1}^{N} B_j$-K. As the customers move sequentially through the cyclic network, the holes execute a counter sequential motion since each movement of a customer from the j-th node to the (j+1)st node corresponds to the movement of a hole in the opposite direction (i.e. from the j+1st node to the j-th node). It is then shown that these two systems are *duals*. That is, if a customer (hole) at node j is blocked in one system then node j+1 has no holes (customers) in its dual. Let (B_i,μ_i) be the capacity and the service rate of node i and $\{\ (B_1,\mu_1), (B_2,\mu_2),..., (B_N,\mu_N)\ \}$ be a cyclic network with K customers. Then, its dual is $\{\ (B_1,\mu_N, (B_N,\mu_{N-1}),..., (B_2,\mu_1)\ \}$ with $\sum_{j=1}^{N} B_j$-K customers. Let, $p(\underline{n})$ and $p^D(\underline{n})$ be the steady state queue length probabilities of a cyclic network and its dual, respectively, where $\underline{n}=(i_1,i_2,...,i_N)$ is the state of the network with i_j being the number of customers at node j. Then, for all feasible states, we have:

$$p(i_1,i_2,...,i_N) = p^D(B_1-i_1,\ B_N-i_N,...,B_2-i_2)$$

We note that, if the number of customers in the network is such that no node can be empty, then the dual network is a non-blocking network (i.e. the number of holes is less than or equal to the minimum node capacity) and it has a product form queue length distribution. But then, from the concept of duality, the original network has a product form queue length distribution. Hence, we have the following lemma:

Lemma 6: *Consider a cyclic network under BBS-SO blocking. The service time at each node is assumed to be exponentially distributed. The network has a product form* based on the following three conjectures:

Conjecture 1: The throughput of a closed queueing network with finite node capacities is less than or equal to the throughput of the same network with infinite node capacities, i.e. $\lambda(K)\leq\beta(K)$; $K=1,\ldots,M$

Conjecture 2: Probability that a node is empty does not increase as the number of customers in the network increases, i.e. $p_i^K(0)\geq p_i^{K+1}(0)$, $K=1,\ldots,M-1$, *where* $p_i^J(0)$ *is the probability that node i is empty when there are J customers in the network.*

Conjecture 3: Probability that a node is blocked does not decrease as the number of customers in the network increases, i.e. $p_i^{K+1}(b)\geq p_i^K(b)$, $K=1,\ldots,M-1$, *where* $p_i^J(b)$ *is the probability that node i is blocked when there are J customers in the network.*

Lemma 11: *Consider a closed exponential queueing network under BAS blocking, and let* $M=\sum_{i=1}^{N} B_i$, $n=\min\{B_i,\ i=1,\ldots,N\}$, *and* $\lambda^*=\{\lambda(K),\ K=1,\ldots,M\}$. *For a moment, assume that the network has infinite queue capacities and let* $\beta(K)$ *be its throughput when there are K customers in it. Then:* $\beta(n+1)\leq\lambda^*\leq\beta(M-\min\{B_i,\ i=1,\ldots,N\}+1)$.
Now, let K^* *be such that* $\lambda^*=\lambda(K^*)$. *Then:*

$$\max\left(\min\ \{B_j \text{ such that } p_{ij}\neq 0\ j=1\ldots N\}\ \ i=1\ldots N\right)\leq K^*\leq M-\min\{B_i,\ i=1,\ldots,N\}+1$$

Next, we present some equivalencies between closed queueing networks with respect to the buffer capacities and the number of jobs in the network. Two networks are said to be equivalent if they have the same steady state queue length distribution.

Under certain restrictions, the steady state queue length distribution of a closed queueing network under BAS or BBS-SO blocking is identical for a range of values of K. Consider a closed queueing network with exponential servers. If there is a node m with $B_m>\sum_{i=1}^{N} B_i-B_m$, i.e. the buffer capacity of node m is greater than the remaining capacity of the network, then the following lemma presents a sufficient condition for the network to have the queue length distribution for a range of values of K.

Lemma 12: *Consider an exponential closed queueing network with node capacities* B_i. *Let* $B_m=\max\{B_i,\ i=1,\ldots,N\}$. *If* $B_m>\sum_{i=1}^{N} B_i-B_m$ *then the network has the same steady state queue length distribution for all* $K\in S$, *where*

i) $S=\{L\colon \sum_{i=1}^{N} B_i-B_m+1\leq L\leq B_m\}$ *in BAS blocking, and, ii)* $S=\{L\colon \sum_{i=1}^{N} B_i-B_m\leq L\leq B_m\}$ *in BBS-SO blocking.*

Furthermore, a closed network under BAS blocking with $K=B_m+1$ *jobs has the same* *queue length distribution if*

$$K \geq \sum_{i=1}^{N} B_i - \min\{B_j, j=1,\ldots,N\}.$$

2.1.3. Other Results

Let $\lambda_i(K)$ and $\lambda(K)$ be respectively the throughput of node i and the throughput of the network when there are K customers in the network. By definition, $\lambda_i(K)=\{1-P_i^K(0)-P_i^K(b)\}\mu_i$, where μ_i is the service rate, $P_i^K(0)$ and $P_i^K(b)$ are the probabilities that node i is empty and blocked respectively given that there are K customers in the network. Shanthikumar and Yao [1989] considered exponential cyclic queueing networks under BBS-SO blocking and they identified the conditions under which the performance measures are monotone in service rate, node capacity, and population size. Let $\mu_i(k)$ be the load dependent service rate at station i. Furthermore, let $\underline{B}=(B_1,\ldots,B_N)$ and $\underline{\mu}=(\mu_1(k),\ldots,\mu_N(k))$ be the vector of node capacities and service rates respectively. The main properties obtained by Shanthikumar and Yao are given as follows:

Lemma 7: *Consider a cyclic network under BBS-SO blocking with two sets of service rates,* $\mu_i^1(k)$, $\mu_i^2(n)$. *If* $\mu_i^1(k) \geq \mu_i^2(n)$, $k \geq n$, $i=1,\ldots,N$, *then*

$$Throughput(\underline{\mu}^1, \underline{B}, K) \geq Throughput(\underline{\mu}^2, \underline{B}, K)$$

Lemma 8: *Consider a cyclic network under BBS-SO blocking with two sets of buffer capacities,* $\underline{B}^1$, $\underline{B}^2$ *and assume that* $\mu_i(k)$ *is increasing in k for each i,* $i=1,\ldots,N$. *If* $\underline{B}^1 \geq \underline{B}^2$ *then*

$$Throughput(\underline{\mu}, \underline{B}^1, K) \geq Throughput(\underline{\mu}, \underline{B}^2, K)$$

Lemma 9: *Let* $B^*=\max\{B_i, i=1,\ldots,N\}$. *Then, for* $0<K<B^*$,

$$Throughput(\underline{\mu}, \underline{B}, K+1) \geq Throughput(\underline{\mu}, \underline{B}, K)$$

Lemma 9 states that throughput is non-decreasing with respect to the number of customers as long as the number of customers is less than the maximum node capacity in the network. Similarly, lemmas 7 and 8 illustrate the monotonicity of the throughput with respect to service rates and node capacities, respectively. In particular, lemma 7 states that the throughput of the network does not decrease if the service rate of a node (or a group of nodes) increases. Similarly, lemma 8 states that the throughput of the network is non-decreasing as the buffer capacity of a node (or a group of nodes) increases.

The following lemma is a consequence of self-duality in cyclic networks conjectured by Persone and Grillo [1987] and Onvural [1987].

Lemma 10: *An exponential cyclic network under BBS-SO blocking has the same throughput with K and* $\sum_{j=1}^{N} B_j - K$ *customers in it.*

Lemma 11 below provides bounds on the maximum throughput and the number of customers, K*, that produces the maximum throughput (Onvural [1987]), which is *blocking another node are aggregated into one state.*

We note that a closed network with exactly one node with an infinite capacity is a special case that satisfies this condition.

Now, consider a closed queueing network under BAS blocking with K jobs such that the node with the maximum buffer capacity can not be empty. Then, the following lemma illustrates that if the buffer capacity of this node and the number of jobs in the network are increased by the same amount, then both networks have the same steady state queue length distribution.

Lemma 13: Consider an exponential closed queueing network under BAS blocking (CQN-1) and let node m has the maximum buffer capacity, i.e. $B_m=max\{B_j;$ $j=1,..,N\}$*. Also, consider another network with the same parameters as the CQN-1 except that the buffer capacity of node m is increased to* B^*_m *(CQN-2). Furthermore, assume that the number of jobs in CQN-1 be such that no node can be empty, i.e.* $K\geq\sum_{i=1}^{N} B_i-B_m+1$*. Then, CQN-1 with K jobs have the same steady state queue length distribution as CQN-2 with* $K^*=K+B^*_m-B_m$ *jobs.*

The concept of duality as introduced by Gordon and Newell does not provide a product form solution for closed queueing networks under BAS blocking, as these networks could not be shown to have self-dual configurations. However, the following corollary, which is a special case of lemma 12, illustrates that two networks under BAS blocking has the same steady state queue length distribution if they have the same dual network.

Corollary 1: Consider a CQN-1 under BAS blocking with K_1 *jobs and buffer capacities* B_i*. Also consider a CQN-2 identical to CQN-1 but with* K_2 *jobs and buffer capacities* C_i*. If* K_1 *and* K_2 *are such that no node can be empty, then the two networks have the same steady state queue length distribution if they have the same number of holes, i.e.* $\sum_{i=1}^{N} B_i-K_1=\sum_{i=1}^{N} C_i-K_2$.

Let us now consider a deadlock free open queueing networks with multiple arrival streams and exponentially distributed service times, and, let A be the set of nodes in which arrivals occur. For each node ieA, the interarrival times are assumed to be distributed exponentially with rate l_i. To find its equivalent closed queueing network, we will add a node (node 0) to this network with parameters $B^*_0=\infty$, $\mu^*_0=\lambda=\sum_{i\varepsilon A}\lambda_i$, and $p^*_{0i}=\lambda_i/\lambda$, iεA.

The total number of customers, K, in the CQN-B is set to be equal to $\sum_{i=1}^{N} B_i$. Departures in the open network are routed back to node 0 in the closed network. Furthermore, let node 0 in the closed network be subject to RS-RD blocking independent of the blocking mechanism used in the open network. Hence, we might have two different blocking mechanisms in the equivalent CQN-B. Then, we have:

Lemma 14: Consider a deadlock free open network under type i blocking, i=BAS, BBS, RS, and consider a closed network constructed as above. Node 0 is subject to RS-RD blocking while all other nodes are subject to type i blocking, same as in the open network. Let $K=\sum_{i=1}^{N} B_i$. *Then, the two networks have the same rate matrix.*

For other types of equivalencies between open and closed networks, i.e. with single arrival streams and multiple classes of customers, an interested reader may refer to Onvural and Perros [1988].

The throughput of cyclic networks under BAS blocking with number of customers being equal to the capacity of the network can be calculated efficiently using the following lemma.

Lemma 15: Consider an exponential cyclic network under BAS blocking with node capacities B_i *and K customers. If K=M then the throughput of the network is equal to* $1/E[max(X_1,X_2,\ldots,X_N)]$ *where* X_i *is the service time at node i. Furthermore, assuming* X_i*'s are distributed exponentially with rate* μ_i *,we have:*

$$E[max(X_1,X_2,\ldots,X_N)] = \int_0^\infty (1-\prod_{i=1}^{N}(1-e^{-\mu_i t}))\,dt.$$

For presentation purposes, let us consider a cyclic network with N=3, K=3, and $B_i=1$, i=1,2,3. Let X_i be the service time at node i and without loss of generality assume that $X_1 \leq X_2 \leq X_3$. Furthermore, assume that at t=0 all servers are busy working. Then, at $t=X_3$, all three servers will become blocked and a deadlock will occur. If we assume that deadlocks are detected immediately and resolved by instantaneously exchanging the blocked customers then at $t=X_3$, customer at node 1 will go to node 2, customer at node 2 will go to node 3, and customer at node 3 will go to node 1. At this point in time, all servers will start a new service. The points at which all three servers start a new service are the renewal points and the throughput of the cyclic network is 1/(expected time between arrivals) by definition.

2.1.4. Symmetric Networks

In this section, we discuss the concept of indistinguishable nodes as introduced by Onvural [1987] and Persone and Grillo [1987] in symmetric cyclic networks. When applicable, this notion allows the solution of the rate matrix of such networks on a reduced state space. It can be used as an efficient method to validate approximations as well as to study systems with symmetric parameters.

Consider an exponential cyclic network under BAS blocking with parameters $B_i=B$ and $\mu_i=\mu$, i=1,...,N. The algorithm presented next utilizes an aggregate state space obtained from the original state space after it is reduced by a factor of N. Consider this cyclic network under BAS blocking with B=2, K=4, and N=3. The state space of this network has the following structure with all transition rates being equal to μ.

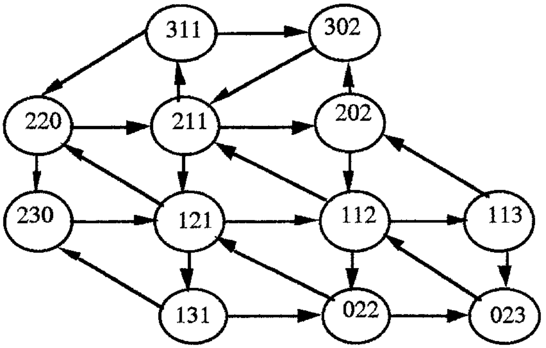


Figure 5: Transition rate diagram of a symmetric cyclic network under BAS blocking

with (n_1,n_2,n_3) denoting the state of the network and n_i=B+1 (=3) denoting that node i is blocking preceding node. Let us define the following classes, where a state is a member of a class if that state has the same steady state probability as all the other states in the same class.

$$S_1=\{ (2,2,0),(0,2,2),(2,0,2) \}$$
$$S_2=\{ (2,3,0),(0,2,3),(3,0,2) \}$$
$$S_3=\{ (2,1,1),(1,2,1),(1,1,2) \}$$
$$S_4=\{ (3,1,1),(1,3,1),(1,1,3) \}$$

Then, we have the following state space structure for these equivalence classes with all transition rates being equal to μ.

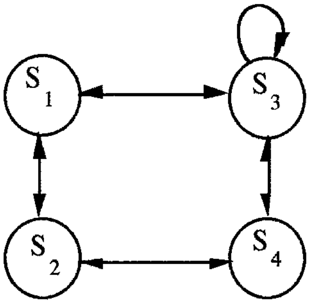


Figure 6: The transition rate diagram of the aggregated network

Then it can be shown that the respective steady state queue length distributions have the following relationship: $P(S_i)= \sum_{(i_1,i_2,i_3)\varepsilon S_i} P(i_1,i_2,i_3)$, i=1,...,4. Algorithm 1 summarizes the procedure to solve symmetric networks efficiently.

Algorithm 1:

S1. Generate the equivalence classes S_i, and set up the rate matrix.

S2. Solve the system numerically to obtain $P(S_i)$.

S3. Calculate the normalizing constant G_K for the original network as follows:

$$G_K = \sum_{i=1}^{S} R_i P(S_i)$$

where S is the number of equivalence classes and R_i is the number of states in the equivalence class i.

S4. $P(i_1,i_2,...i_N) = G_K^{-1} P(S_i)$.

Finally, we note that although the concept of indistinguishable nodes is discussed in cyclic networks under BAS blocking, it is also applicable to other blocking mechanisms defined in section 2 in cyclic exponential cyclic networks, and the central server model.

2.2. Open Networks

Unlike closed networks, open networks with blocking possess exact solutions in only a few special cases. Consider a tandem network with constant service times under BAS blocking. Arrivals occur only at the first node which has an infinite capacity. Interarrival times are assumed to be arbitrary. Then, we have (cf. Avi-Itzhak [1965]):

Lemma 16: i) The time spent in the system is independent of the order of the nodes and the capacity of nodes with finite capacities, and ii) the total time a customer spends in the network is the same as the same customer would spent waiting in a single server queue with a constant service time equal to the largest service time at the network, assuming that the service process in the equivalent single queue is the same as it is at the network.

Other than this result for tandem networks, the only other exact results in the literature, to the best of our knowledge, are reported in case of two node networks. A survey of two node open networks is given in Perros [1988]. In the following, the arrivals are assumed to occur at node 1 in a Poisson manner (with rate λ) which has an infinite capacity. The service times at each node is distributed exponentially with corresponding rates μ_1 and μ_2.

Konheim and Reiser [1976] obtained closed form expression for two node open networks under BBS blocking mechanisms assuming that a customer departing from the second server may be fed back to the first node.

Two limiting cases were investigated in Foster and Perros [1980], Konheim and Reiser [1976] and Hatcher [1969]. In the first case, it is shown that as μ_1-->∞ , the two node system is reduced to an M/M/1 queue with arrival rate l and service rate μ_2. In the second case, it is assumed that the first node is saturated, an assumption often considered in production systems. In order for the first node to be saturated, its service rate should satisfy the condition $\mu_1 \leq \mu_0$, where μ_0 is the critical service rate at which the first queue becomes unstable. μ_0 is obtained numerically. Once μ_0 is known, the second queue is an M/M/1/B+1 queue with arrival rate μ_0 and service rate μ_2, where B is the capacity of node 2.

Asare [1978] considered two node networks under processor sharing discipline, where customers cycle between two nodes infinitely quickly receiving an infitestimal amount of service at each server and showed that the queue length distribution of this system has a closed form solution.

3. APPROXIMATIONS

As discussed in section 2, queueing networks with blocking do not, in general, have product or closed form solutions that can be used to solve such networks efficiently. Furthermore, numerical techniques are often restricted to small configurations, limiting their applicability. Hence, approximations are often used to investigate the performance characteristics of queueing networks with blocking.

3.1. Closed Queueing Networks

Akyildiz [1988a,b] developed approximation algorithms for the throughput of closed queueing networks with exponential and general service times. He approximates the throughput assuming that the throughput of a blocking network is approximately the same as an equivalent non-blocking network with product form queue length distribution. The equivalent network with infinite queue capacities has the same parameters as the blocking network except K. The number of customers in the non-blocking network is chosen such that the number of states of the blocking network is as close to the number of states of the non-blocking network as possible. The only assumption in the algorithm is that the network under consideration should be deadlock free.

Onvural and Perros [1989b] developed an approximation algorithm to calculate the throughput of large closed exponential queueing networks with finite queues. The algorithm approximately determines the number of customers such that the throughput of the network is maximum and fits a curve that passes through a number of known points to estimate the unknown throughput values as the number of customers in the network varies.

Perros, Nilsson, and Liu [1989] developed a numerical procedure for the approximate analysis of closed queueing networks in which some of the queues have finite capacities.

Suri and Diehl [1986] introduced the concept of *variable buffer size* and used it together with the *flow equivalent approximations* to approximate the throughput of cyclic networks with at least one node with an infinite capacity. The service time at each node is assumed to be exponentially distributed with rate μ_i, i=1,...,N.

Yao and Buzacott [1985] reported an approximation algorithm for analyzing closed queueing networks under RD-RS blocking. They considered networks of queues with each queue being served by multiple servers. Service times are assumed to follow arbitrary Coxian distributions. The topology of the network is such that if each service distribution is approximated by an exponential distribution with the same mean as the Coxian server, then the resulting exponential network is reversible and has a product form queue length distribution. The approximation is based on the notion of exponentialization.

Kouvatsos and Xenios [1989] used the principle of maximum entropy to find an approximate product form queue length distribution for closed queueing networks under RS-RD blocking. The algorithm requires the solution of non-linear equations using the principle of maximum entropy. An interested reader may refer to Kouvatsos [1983] for a detailed description of the maximum entropy principle. The procedure is based on decomposing the network into individual nodes and analyzing them in isolation with each node being studied as a GE/GE/1/B/FCFS queue, i.e. with generalized exponential arrival and service distributions. The service distribution at each queue is revised to accommodate the delays a customer might undergo due to blocking.

Dallery and Frein [1989] developed an approximation algorithm for the analysis of cyclic networks under BAS blocking in which there is at least one node with an infinite capacity. The approach is similar to the ones developed in the literature for open queueing networks with blocking. The algorithm decomposes the network into individual nodes with revised capacities, revised service rates, and revised arrival rates. The service and the interarrival times at each node in isolation are assumed to be exponentially distributed. The corresponding rates are determined iteratively assuming at each step that the network throughput is known. The algorithm was proven to be convergent. It produces both the throughput of the network and the mean queue lengths. Frein and Dallery [1989] extended this algorithm to cyclic networks under BBS-SO blocking.

We now discuss how a closed queueing network with blocking can be decomposed into individual nodes so that the marginal queue length probabilities obtained with these nodes analyzed in isolation are exact, i.e. the same as they are obtained from the joint steady state queue length distribution. Consider a cyclic network under BBS-SO blocking with N nodes, buffer capacities B_i, exponentially distributed service times, and K customers in it. Let $\underline{n}=\{n_1,...,n_N\}$ denote the state of this network, where n_i is the number of customers at node i and $P(\underline{n})$ be the steady state queue length distribution of the network. We have $0 \le n_i \le B_i$ and $\sum_{i=1}^{N} n_i=K$, i=1,...,N. Let us now partition the state space of the network into disjoint sets, $S_i(j)$, with respect to the number of customers, j, at node i. That is:

$$S_i(j) = \{\underline{n} \mid n_i=j\}, \quad j=0,...,B_i$$

Then, the behavior of node i in isolation can be readily obtained from the global balance equations by summing them over the sets $S_i(j)$, $j=0,...,B_i$.

For presentation purposes, consider a four node cyclic network with parameters $B_i=2$, i=1,2,3,4 and K=5. The global balance equations for the set $S_1(0)$ are given as follows:

$$\mu_4 \, P(0,1,2,2) = \mu_1 \, P(1,0,2,2) + \mu_2 \, P(0,2,1,2)$$
$$(\mu_2+\mu_4) \, P(0,2,1,2)= \mu_1 \, P(1,1,1,2) + \mu_3 \, P(0,2,2,1)$$
$$(\mu_3+\mu_4) \, P(0,2,2,1)= \mu_1 \, P(1,1,2,1)$$

Summing these equations side by side and canceling common terms, we have:

$$\mu_4 \, P_1(0) = \mu_1 \, \{ P_1(1) - P(1_1,2_2) \},$$

where $P\{k_i,n_j\}$ is the joint probability of having k and n customers at nodes i and j, respectively, and, $P_i(k)$ is the marginal probability of having k customers at node i. Similarly, for the other two sets $S_1(k)$, k=1,2 we have:

$$\mu_4 \, \{ P_1(1) - P\{1_1,0_4\} \} + \mu_1 \, \{ P_1(1) - P\{1_1,2_2\} \} = \mu_4 \, P_1(0) + \mu_1 \, \{ P_1(2) - P\{2_1,2_2\}\}$$
$$\mu_1 \, \{ P_1(2) - P\{2_1,2_2\}\} = \mu_4 \, \{ P_1(1) - P\{1_1,0_4\} \}$$

These equations can equivalently be written as:

$$\mu_4 \, P_1(0) = \mu_1 \, \{ 1 - P(1_1,2_2) \, / \, P_1(1) \} \, P_1(1)$$
$$\mu_4 \, \{ 1 - P(1_1,0_4) \, / \, P_1(1)\} \, P_1(1) = \mu_1 \, \{1 - P(2_1,2_2) \, / \, P_1(2)\} \, P_1(2)$$

In this example, we observe that node 1 in isolation behaves like an $M/M/1/B_i$ queue with state dependent arrival and service rates, which is generalized in the following lemma.

Lemma 17: *Consider a deadlock free cyclic network under BBS-SO blocking with parameters B_i, K, μ_i, and p_{ij}; i,j=1,...,N. The service times are distributed exponentially. Then, the state dependent arrival, $\lambda_i(j)$, and service rates, $\mu_i(j)$, of a node* *in isolation is given in terms of the joint steady state probabilities of the original network as follows:*

$$\lambda_i(j) = \mu_{i-1}\{1-P(0_k,j_i) / P_i(j) \}, \qquad j=0,\dots,B_i-1;\ i=1,\dots,N$$

$$\mu_i(j) = \mu_i \{1 - P(j_i,B_k)) / P_i(j) \}, \qquad j=1,\dots,B_i$$

Furthermore, the marginal queue length probabilities of node i is obtained from the following set of equations:

$$\lambda_i(j)\, P_i(j) = \mu_i(j+1)\, P_i(j+1), j=0,\dots,B_i-1;$$

$$\text{and } \sum_{j=0}^{B_i} P_i(j)=1.$$

This result is rather surprising as a node of a closed queueing network under BBS-SO blocking in isolation behaves like an M/M/1/B queue. In particular, although the departure rate from a node of a blocking network is not Poisson, the above rate equations are the same rate equations of a $\lambda_i(j)/\mu_i(j)/1/B_i$ queue with state dependent Poisson arrivals and exponential service time distributions.

Although it is more involved than above construction, similar approach can be used to obtain the behavior of nodes in isolation in networks under BAS blocking. The decomposition approach for this type of blocking is discussed in the context of open networks.

Approximations developed for closed queueing networks in the literature are either based on some empirical observations or applications of approximations developed for open networks with blocking. Unlike open networks, the use of decomposition technique to approximately analyze closed networks is not a trivial task. In particular, in the simplest case, given the state of a node in isolation, i.e. the number of customers, the server may not be subject to blocking in closed networks depending on the total number of customers in the network whereas knowing the state of a node in isolation does not provide much information on the number of customers in other nodes of an open network. Hence, the state dependent arrival and service rates resulted by the exact decomposition appears to depend more strongly in closed networks on the number of customers at the node in isolation than they are in open networks, complicating the process.

3.2. Open Queueing Networks

Most approximations developed in the literature to analyze open queueing networks with blocking are based on decomposing the network into individual nodes and analyzing each node in isolation. In order to analyze a node in isolation, it is necessary to determine its arrival and service processes as well as its capacity.

In particular, consider a network under BAS blocking and let C be the capacity of a node with original capacity B in isolation. Furthermore, let the arrivals finding C customers at the node in isolation are assumed to be lost. In this case, if node i is blocked by node j, then the blocked customer is in fact can be viewed as a part of node j, i.e. joined queue j. This is because, when a departure occurs from node j then one of the blocked customers immediately joins node j. Hence, in general, C is equal to B plus the number of upstream servers connected directly to the node. However, if the node in isolation is analyzed with an arrival process which is subject to blocking then C=B. In this case, upon its attempt to enter the node, if the arriving customer finds B customers already in the queue then the arrival process is suspended. When a space becomes available, the blocked arriving customer joins the queue unblocking the arrival process.

In other types of blocking mechanisms, the blocked customer can not be viewed as part of its destination node as the blocked customer would be currently receiving service at its original node at a time a space becomes available at its destination. Hence, we have C=B in BBS and RS blocking mechanisms.

For presentation purposes, let us consider a tandem network under BAS blocking with N nodes, buffer capacities B_i, and exponentially distributed service times with corresponding rates μ_i. Arrivals are assumed to occur in a Poisson manner at the first node with rate l. The exact behavior of a node in isolation can be obtained from the global balance equations, similar to the way it was done in the context of closed networks. Instead, we discuss intuitively how the arrival and service processes look like in isolation. Since node N is not subject to blocking, its service process in isolation is the same as it is in the original network. Node N-1, on the other hand, is subject to blocking and its service process in isolation should include the delay that a blocked customer goes through. In particular, if upon service completion at node N-1, there is no space available at node N then the customer is blocked and the service at node N-1 is suspended. The blocking delay in this case is the remaining service time at node N. However, due to the memoryless property of exponential distributions, this time is exponentially distributed with rate μ_N. Hence, the service process at node N-1 when analyzed in isolation is two phase Coxian, as illustrated in figure 7.

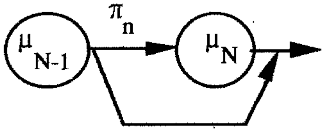


Figure 7: Service process at node N-1 in isolation.

The branching probability π_n depends on the state of the system and defined in terms of the joint queue length distribution of nodes N-1 and N-2. In particular, from the exact decomposition, we have π_n equal to the conditional probability that node N is full given that there are n customers at node N-1. Similarly, the blocking delay a customer goes through at node N-2 is given as follows: Upon service completion at node N-2, there are two possibilities. There is a space at node N-1 and the customer at node N-2 joins node N-1, or node N-1 is full at that time blocking node N-2. If blocked, there are two more possibilities. In the first case, node N-1 may be blocked by node N. Then, the blocking delay is the remaining service time at node N which is distributed exponentially. On the other hand, if node N-1 is busy serving then the blocked customer first has to wait for service completion at node N-1. Upon service completion at node N-1, customer at node N-1 may join node N, unblocking node N-2, or may get blocked by node N. Hence, the service process at node N-2 in isolation is phase type with three stages corresponding to servers N-2, N-1, and N, as illustrated in figure 8.

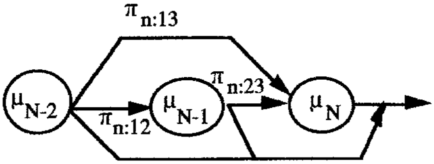


Figure 8: Service process at node N-2 in isolation.

The branching probabilities are now expressed in terms of the joint queue length distributions of nodes N-2, N-1, and N. For example, $\pi_{n:12}$ is the conditional probability that node N-1 is full and server N-1 is busy serving given that there are n customers at node N-2; $\pi_{n:23}$ is the conditional probability that node N is full given that there are n customers at node N-2 and node N-2 is blocked by node N-1; and $\pi_{n:13}$ is the probability that node N-1 is full and blocked by node N given that there are n customers at node N-2.

In general, the service process at node i includes all downstream servers with branching probabilities being positive for every j and k such that k>j.

The arrival process at a node in isolation "behaves like" a state dependent Poisson process. In particular the arrival rate to node i when it is in state (n,p) (i.e. there are n customers in it and service is in phase p) is equal to the product of the service rate at node i-1 and the conditional probability that node i-1 is not empty given that node i is in state (n,p). We note that the state (n,p) corresponds to, in general, nodes i to i+p-1 are blocked and node p is busy serving. Accordingly, the exact values of arrival rates at node i are given in terms of the joint queue length distributions of nodes i-1 to N.

An interested reader may refer to Perros [1989] for a survey of approximation algorithms reported in the literature for open networks with blocking. Tandem networks with blocking have been studied under a multiplicity of assumptions in the literature. Perhaps, one of the earliest result is developed by Hillier and Boiling [1967]. Assuming that the throughput of the network is known, the service rate at node i in isolation is fixed so that the throughput of the node is equal to the network throughput, whereas, the arrival rate is assumed to be equal to the throughput of the preceding node obtained at previous iteration. Further assuming that the corresponding distributions are distributed exponentially, each queue is analyzed as an $M/M/1/B_i+1$ queue. The algorithm iterates between nodes until a convergence on the network throughput is achieved.

Perros and Altiok [1986] developed an approximation in which each queue is analyzed as an $M/PH/1/B_i+1$ queue. The service process in isolation is obtained exactly the same way as they are in the exact decomposition except branching probabilities which are approximated by applying Little's relation to downstream queues. The arrival process to the queue is assumed to be Poisson and its rate is determined iteratively assuming the network throughput is known. We note that if the first queue is infinite then the network throughput is known. Otherwise, the network throughput has to be approximated iteratively. Over the years, this algorithm is extended to split and merge as well as arbitrary configurations, Coxian server and arrival processes, and tandem networks with Coxian arrivals, Coxian servers and multiple classes of customers.

Gershwin [1987, 1981] and Gun and Makowski [1989] studied tandem queues under BAS blocking where the arrival process is subject to blocking (as opposed to arrivals being lost in above cases). We note that in this case, the buffer capacity of a node in isolation is the same as it is in the original network. Caseau and Pujolle [1979] analyzed tandem networks under RS blocking using single node decomposition. Kouvatsos and Xenios [1989] used the maximum entropy principle to solve each node in isolation to approximately obtain the performance metrics of queueing networks under RS-RD blocking. Brandwajn and Jow [1988] used two node decomposition to analyze tandem networks under BAS and BBS blocking. The network is decomposed into N-1 subsystem with each subsystem i consisting of nodes i and i+1. To solve each subsystem, it is necessary to approximate the interaction between nodes i and i-1 for the arrival process and nodes i+1 and i+2 for the effective service time at node i+1, which are obtained from the analysis of subsystems i-1 and i+1. The algorithm, in general, captures more information than that can be obtained in a single decomposition and produces more accurate results. However, the solution of a subsystem numerically is more time consuming than it is to solve single nodes.

4. BUFFER ALLOCATION

The performance of a system highly depends on its topology, routing in the network, and the capacities of its queues. Although a single optimization model may be formulated, for practical purposes, the problem is generally decomposed into three interrelated optimization problems (Smith and Daskalaki [1988]): optimal topology problem, optimal routing problem, and optimal resource allocation problem.

In the buffer allocation problem, it is assumed that the topology of the system and the routing in the network are given. Then, the buffer capacities at service stations are determined such that the network throughput is close to its maximum value. This problem has a long history of research and development. However, most of the approximations reported in the literature considered tandem topologies of queueing networks in which service stations are connected in series (c.f. Buzacott and Shanthikumar [1992], Yamashita and Suzuki [1987] and Jafari and Shanthikumar [1989], Soyster. et. al. [1979], Sheskin [1976]). This model does not take the interactions between various tandem lines in the system into consideration. Smith and Daskalaki [1988] developed a heuristic to address the buffer allocation problem for tandem, merge.and split topologies of automated assembly lines. More recently, Yamashita and Onvural [1993]] proposed two approximation algorithms are developed to allocate the buffer capacities at each node such that the network throughput is close to its optimum value.

5. CONCLUSIONS

In this paper, we give a tutorial of queueing networks with blocking. Except for a few special cases, these networks could not be shown to have product form solutions. Although the steady state queue length distributions of these networks can, in theory, be calculated by solving the global balance equations together with the normalization equation numerically, this procedure can, in practice, be restrictive due to the time complexity of the procedure and the large storage required to store the rate matrices, particularly for large networks. Since exact values of their steady state queue length distributions are, in general, not attainable, good approximation algorithms are required to analyze queueing networks with finite queues.

REFERENCES

The list of references listed here by no means complete. They are the ones referred explicitly in the text and those of surveys, conference proceedings, and special issues for the interested reader to refer to. In particular, a bibliography of related papers on queueing networks with blocking can be found in Perros (1989).

Akyildiz, I.F. 1987, Exact Product Form Solutions for Queueing Networks with Blocking, IEEE Tran. Computers, 1, 121-126

Akyildiz, I.F. 1988b, On the Exact and Approximate Throughput Analysis of Closed Queueing Networks with Blocking, IEEE Trans. Software Engineering, SE-14-1, 62-71

Akyildiz, I.F. and Von Brand, H. 1989a, Exact Solutions for Open, Closed and Mixed Queueing Networks with Rejection Blocking, Theor. Comp. Sci. J., 64, 203-219

Akyildiz, I.F. and Perros, H.G. 1989, Special Issue of Performance Evaluation on Queuing Networks with Finite Capacity Queues, 10-3

Altiok, T. and Perros, H.G. 1987, Approximate Analysis of Arbitrary Configurations of Queueing Networks with Blocking, Annals of OR, 481-509

Asare, B. 1978, Queue networks with blocking, Ph.D. Thesis, Trinity College Dublin, Ireland, 1978.

Avi-Itzhak, B. 1965, A Sequence of Service Stations with Arbitrary Input and Regular Service Times, Management Science, 11, 565-571

Balsamo, S., Clo, M.C., and Donatiello, L. 1992, Cycle Time Distributions of Cyclic Networks with Blocking, Proc. Second Int. Workshop on Queueing Networks with Blocking, Onvural and Akyildiz (Eds.), North Holland

Balsamo, S. and Donatiello, L. 1988, Two-Stage Cyclic Network with Blocking: Cycle Time Distribution and Equivalence Properties, Proc. Modeling Tech. and Tools for Computer Perf. Eval., Potier and Puigjaner (Eds), 513-528

Balsamo, S., Persone V. De Nitto, and Iazeolla, G. 1986, Some Equivalencies of Blocking Mechanisms in Queueing Networks with Finite Capacity", Manuscript, Dipartimento di Informatica, Universite di Pisa, Italy

Balsamo, S. and Iazeolla, G. 1983, Some Equivalence Properties for Queueing Networks with and without Blocking", Performance'83, Agrawala and Tripathi (Eds), 351-360, North Holland Publishing Company, Amsterdam

Bocharov, P.P. and Albores, F.K. 1980, On Two-Stage Exponential Queueing System with Internal Losses or Blocking, Problems of Control and Information Theory, 9, 365-379

Boxma, O. and Konheim, A. 1981, Approximate Analysis of Exponential Queueing Systems with Blocking, Acta Informatica, 15, 19-66

Brandwajn, A. and Jow, Y.-L.L. 1988, An Approximation Method for Tandem Queues with Blocking, Operations Research, 36, 73-83

Caseau, P. and Pujolle, G. 1979, Throughput Capacity of a Sequence of Transfer Lines with Blocking Due to Finite Waiting Room", IEEE Trans. Software Engineering, 5, 631-642

Dallery, Y. and Frein, Y. 1989, A Decomposition Method for the Approximate Analysis of Closed Queueing Networks with Blocking, Proc. First International Workshop on Queueing Networks with Blocking, Perros and Altiok (Eds), 193-215, North Holland

Daskalaki, S. and MacGregor Smith, J. 1989, The Static Routing Problem in Open Finite Queueing Networks, Proc. First International Workshop on Queueing Networks with Blocking, Perros and Altiok (Eds), 193-215, North Holland

Foster, F.G. and Perros, H.G. 1980, On the Blocking Process in Queue Networks, Eur. J. Oper. Res., 5, 276-283

Frein, Y. and Dallery, Y. 1989, Analysis of Cyclic Queueing Networks with Finite Buffers and Blocking Before Service, Performance Evaluation, 197-210

Gershwin, S.B. 1987, An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking, Operations Research, 35, 291-305

Gershwin, S.B. 1987, An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking, Operations Research, 35, 291-305

Gordon, W.J. and Newell, G.F. 1967a, Cyclic Queueing Systems with Restricted Queues, Oper. Res., 15, 266-278

Gun, L. and Makowski, A.M. 1988, Matrix Geometric Solution for Finite Queues with Phase Type Distributions, Performance'87, Courtois and Latouche (Eds.), 269-282, North Holland

Hatcher, J.M. 1969, The Effect of Internal Storage on the Production Rate of a Series of Stages Having Exponential Service Times, AIIE Trans., 1, 150-156

Hillier, F.S. and Boling, R.W. 1967, Finite Queues in Series with Exponential or Erlang Service Times-A Numerical Approach, Operations Research, 15, 286-303

Hillier, F.S. and So, K.C. 1989, The Assignment of Extra Servers to Stations in Tandem Queueing Systems with Small or No Buffers, Performance Evaluation, 10-3, 219-232

Hordijk, A. and Van Dijk, N. 1981, Networks of Queues with Blocking, Performance'81, Klystra (Ed.), 51-65, Elsevier Science Publishers B.V. (North Holland)

Jafari, M.A. and Shanthikumar, J.G. 1984, Allocation of Buffer Storages Along a Multi-Stage Automatic Transfer Line, IE Rept. 84-003, Arizona University

Jafari, M.A. and Shanthikumar, J.G. 1985, Determination of Optimal Buffer Storage Capacities and Optimal Allocation in Multi-Stage Automatic Transfer Lines, IE&OR Rept. 85-011, Syracuse University

Kelly, F.P. 1984, Blocking, Reordering, and the Throughput of a Series of Buffers, Stochastic Processes and Their Applications, 17, 327-336

Kelly, K.P. 1979, Reversibility and Stochastic Networks, John Wiley and Sons Ltd., Chichester, England

Konheim, A.G. and Reiser, M. 1976, A Queueing Model with Finite Waiting Room and Blocking, J. ACM, 23, 328-341

Kouvatsos, D.D. 1983, Maximum Entropy Methods for General Queueing Networks, Modeling Tech. and Tools for Perf. Analysis, Potier (Ed.), 589-608, North Holland, Amsterdam

Kouvatsos, D.D. and Xenios, N.P. 1989, MEM for Arbitrary Queueing Networks with Multiple General Servers and Repetitive Service Blocking, Performance Evaluation, 10-3, 169-196

Labetoulle, J. and Pujolle, G. 1980, Isolation Method in a Network of Queues, IEEE Transactions on Software Engineering, SE-6, 373-381

Langaris, C. and Conolly, C. 1984, On the Waiting Time of a Two-Stage Queueing System with Blocking, J. Appl. Prob., 21, 628-638

Lavenberg, S.S. 1978, Stability and Maximum Departure Rate of Certain Open Queueing Networks Having Finite Capacity Constraints, RAIRO Informatique/Computer Science, 12, 353-370

MacGregor Smith, J. and Daskalaki, S. 1988, Buffer Space Allocation in Automated Assembly Lines, Operations Research, 36, 343-358

Mitra, D. and Mitrani, I. 1988, Analysis of a Novel Discipline for Cell Coordination in Production Lines, AT&T Bell Labs Res. Rep.

Onvural, R.O. 1987, Closed Queueing Networks with Finite Buffers, Ph.D. thesis, CSE/OR, North Carolina State University

Onvural, R.O. 1989a, On the Exact Decomposition of Exponential Closed Queueing Networks with Blocking, First International Workshop on Queueing Networks with Blocking, Perros and Altiok (Eds), 73-83, North Holland

Onvural, R.O. 1989b, A Note on the Product Form Solutions of Multi-Class Closed Queueing Networks with Blocking, Performance Evaluation, 10-3, 247-254

Onvural, R.O. 1990, A Survey of Closed Queueing Networks with Blocking, ACM Computing Surveys, 22-2, 83-122

Onvural, R.O. and Akyildiz, I.F. 1993, Proc. Second International Workshop on Queueing Networks with Finite Capacity, North Holland, 1993

Onvural, R.O. 1993, Queueing Networks with Finite Capacity (Ed.), Performance Evaluation, April 1993

Onvural, R.O. and Perros, H.G. 1986, On Equivalencies of Blocking Mechanisms in Queueing Networks with Blocking, Oper. Res. Letters, 5-6, 293-298

Onvural, R.O. and Perros, H.G. 1988, Equivalencies Between Open and Closed Queueing Networks with Finite Buffers, Performance Evaluation, 9, 263-269

Onvural, R.O. and Perros, H.G. 1989a, Some Equivalencies on Closed Exponential Queueing Networks with Blocking", Performance Evaluation, 9, 111-118

Onvural, R.O. and Perros, H.G. 1989b, Throughput Analysis in Cyclic Queueing Networks with Blocking, IEEE Trans. Software Engineering, SE 15-6, 800-808

Perros, H.G. 1984, Queueing Networks with Blocking: A Bibliography, ACM Sigmetrics, Performance Evaluation Review, 12-2, 8-12

Perros, H.G. 1989, A Bibliography of Papers on Queueing Networks with Finite Capacity Queues, Performance Evaluation, 10-3, 255-260

Perros, H.G. 1988, Two Node Open Networks with Finite Capacity Queues, CS Tech. Rep., North Carolina State University

Perros, H.G. and Altiok, T. 1986, Approximate Analysis of Open Networks of Queues with Blocking: Tandem Configurations, IEEE Trans. Software Engineering, SE-12, 450-461

Perros, H.G. 1989, Approximation Algorithms for Open Queueing Networks with Blocking, Chapter in Stochastic Analysis of Computer and Communications Systems, Takagi (Ed.), Elsevier SciencePublishers B.V. (north Holland)

Perros, H.G., Nilsson, A., and Liu, Y.C. 1989, Approximate Analysis of Product Form Type Queueing Networks with Blocking and Deadlock, Performance Evaluation, to appear

Persone, De Nitto, V., and Grillo, D. 1987, Managing Blocking in Finite Capacity Symmetrical Ring Networks, 3rd Conference on Data and Communication Systems and Their Performance, Rio de Jenerio, Brasil

Shanthikumar, G.J. and Yao, D.D. 1989, Monotonicity Properties in Cyclic Queueing Networks with Finite Buffers, First International Workshop on Queueing Networks with Blocking, Perros and Altiok (Eds), 325-344, North Holland

Sheskin, T. J. 1976, Allocation of Interstage Storage Along an Automated Production Line, AIIE Trans., 8, 146-152

Soyster, A.L., Schmidt, J.W., and Rohrer, M.W. 1979, Allocation of Buffer Capacities for a Class of Fixed Cycle Production Lines, AIIE Trans., 11, 140-146

Suri, R. and Diehl, G.W. 1986, A Variable Buffer Size Model and Its Use in Analytical Closed Queueing Networks with Blocking, Management Science, 32-2, 206-225

Takahashi, Y., Miyahara, H. and Hasegawa, T. 1980, An Approximation Method for Open Restricted Queueing Networks, Operations Research, 28, 594-602

van Dijk, N.M. 1989, A Simple Throughput Bound for Large Closed Queueing Networks with Finite Capacities, Performance Evaluation, 10-3, 153-168

Van Dijk, N.M. and Tijms, H.C. 1986, Insensitivity in Two Node Blocking Models with Applications, Teletraffic Analysis and Computer Performance Evaluation, Boxma, Cohen and Tijms (Eds), 329-340, Elsevier Science Publishers B.V. (North Holland), Amsterdam, The Netherlands

Yamashita, H. and Onvural, R.O. 1993, Buffer Allocation in Queueing Networks with Arbitrary Topologies, to appear in Annals of OR on Queueing Networks, van Dijk (Ed.)

Yamashita, H. and Suzuki, S. 1987, An Approximate Solution Method for Optimal Buffer Allocation in Serial n-stage Automatic Production Lines, Trans. Japan. Soc. Mech. Eng., 53-C, 807-814

Yao, D.D. and Buzacott, J.A. 1985, Queueing Models for Flexible Machining Stations Part II: The Method of Coxian Phases, Eur. J. Operations Research, 19, 241-252