INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

Intl. Trans. in Op. Res. 9 (2002) 97–112

Generalized loss models and queueing-loss networks

Hisashi Kobayashi[a] and Brian L. Mark[b]*

[a]*Department of Electrical Engineering, School of Engineering and Applied Science, Princeton University, Princeton, NJ 08544, USA*

[b]*Department of Electrical and Computer Engineering, School of Information Technology and Engineering, George Mason University, Fairfax, VA 22030, USA*

**Corresponding author. E-mail: bmark@gmu.edu*

Received 7 June 1999; received in revised form 17 April 2001; accepted 15 June 2001

Abstract

The classical Erlang and Engset loss models have been used extensively in the traffic engineering of traditional telephone exchanges. More recently, these models have been generalized to the so-called loss networks, which provide models for resource-sharing in multi-service telecommunication networks. In this paper, we introduce a new generalized class of models, queueing-loss networks, which captures both queueing and loss aspects of a system. The queueing-loss network model is a natural extension of queueing networks and loss networks that have the product-form solution. We discuss applications of the model and analyze a particular example of a simple queueing-loss network.

Keywords: loss models, loss networks, queueing models, queueing networks, product-form

1. Introduction

Recently, there has been an increasing interest in generalizations of the loss models originally studied by Erlang and Engset in the context of telephone exchanges (see e.g., Syski, 1986). Loss networks provide models for studying the blocking behavior of connection-oriented services in circuit-switched networks, ATM (Asynchronous Transfer Mode) networks, optical networks and wireless networks. As we discuss in this paper, loss networks have much in common with the traditional queueing network models. The earliest work on queueing networks with product form goes back to J. R. Jackson's original paper (1963). Theory for queueing network models has advanced considerably over the past several decades (Baskett et al., 1975; Kelly, 1979; Reiser and Kobayashi, 1975) and has been widely applied to the performance analysis of computing systems and packet-switched networks (Kobayashi, 1978).

This paper introduces a new class of models, *queueing-loss networks*, which are natural generalizations of queueing networks and loss networks. We give a brief development of loss networks by systematically generalizing the classical loss models using notions from the theory of queueing

networks. This development culminates in the introduction of queueing-loss networks and a discussion of their properties. Next, applications of the model are discussed and a particular example of a simple queueing-network model is analyzed. Finally, the paper concludes with a discussion of directions for further research on queueing-loss network models.

2. Generalized loss models

We use the general term 'station' to denote an entity which provides service to arriving calls or customers. A station consists of a number of servers or lines and possibly a waiting room or buffer. A loss station is one that has a finite number of servers and no waiting room. An arriving call either begins service immediately or is rejected due to the lack of a sufficient number of available servers. By contrast, a queueing station, as considered in this paper, has a sufficiently large waiting room such that no call is rejected.

The original loss model studied by Erlang is equivalent to an $M/M/S(0)$ queue[1] (see Fig. 1); i.e., a loss station with S servers where arriving calls form a Poisson process with rate λ and each call occupies a server for an exponentially distributed holding time with mean $1/\mu$. The stationary distribution of the number of *busy* servers is given by

$$P(n) = \frac{1}{G(S)}\frac{a^n}{n!}, \qquad 0 \leqslant n \leqslant S, \tag{1}$$

where $a = \lambda/\mu$ is the *offered load* and $G(S)$ is a normalization constant given by

$$G(S) = \sum_{n=0}^{S} \frac{a^n}{n!}. \tag{2}$$

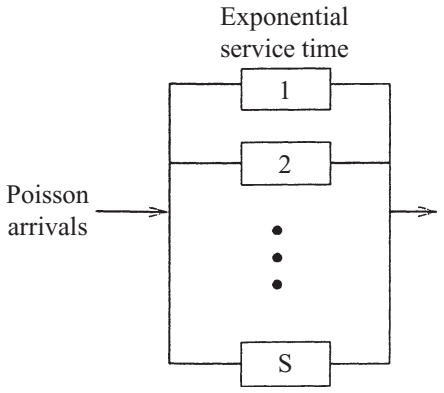

Fig. 1. Erlang loss model

[1] Often the notation $M/M/S/S$ is used in the queueing theory literature, where the second S signifies the maximum number of calls that can be accommodated in the system.

As $S \to \infty$, $G(S) \to e^a$, hence $P(n) \to a^n e^{-a}/n!$, which is the stationary distribution of an infinite-server (IS) station or $M/M/\infty$ queue (see e.g., Kobayashi, 1978). Therefore, the distribution (1) is a truncated Poisson distribution. The probability that all servers are found busy in the steady state is given by the celebrated Erlang loss formula:

$$B(S) \stackrel{\text{def}}{=} P(S) = \frac{a^S}{S!}\left[\sum_{i=0}^{S}\frac{a^i}{i!}\right]^{-1}. \tag{3}$$

The Erlang loss formula can be expressed in terms of the normalization constant as follows:

$$B(S) = 1 - \frac{G(S-1)}{G(S)}. \tag{4}$$

The above probability $B(S)$ is often referred to as the time congestion, since this represents the proportion of time that all the servers are busy. The call congestion or call loss probability $L(S)$ is defined as the probability that a newly-arriving call finds all servers occupied, and hence is lost or blocked, i.e., leaves the system without being served. When the arrival process is Poisson, as in the Erlang loss model, the call congestion and the time congestion can be seen to be equivalent, via the so-called PASTA (Poisson Arrivals See Time Averages) property (Wolff, 1989).

If we replace the Poisson arrival (i.e., an infinite source model) in the Erlang loss model by a finite number N of sources ($N > S$), then we obtain what is often termed the Engset loss model (see Fig. 2), which we denote as an $M(N)/M/S(0)$ queue.[2] Each source generates a call with an exponentially distributed inter-generation time with mean $1/\nu$ and then places the call at the loss station, where it either acquires a server for an exponentially distributed holding time or is blocked. Both completed and lost calls alike return to the sources and a new cycle begins. For this model, $n(t)$, the number of calls in progress at time t, will have, in the steady state, the following truncated binomial distribution:

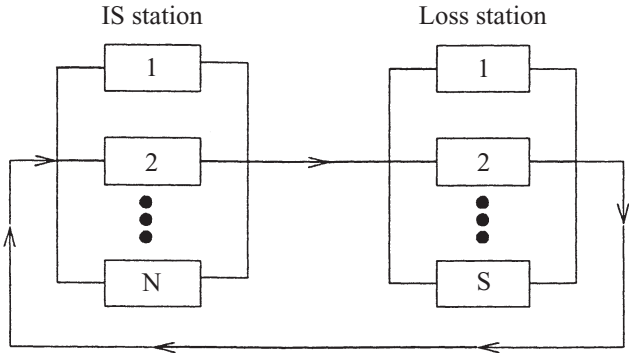

Fig. 2. Engset loss model

[2] In the literature it is often referred to as $M/M/S/N/S$, where the last two symbols represent, respectively, the number of sources and the number of customers that can be accommodated in this service station.

$$P(n,\ N) = \frac{1}{G(S,\ N)} \binom{N}{n} b^n, \qquad 0 \leqslant n \leqslant S, \tag{5}$$

where $b = \nu/\mu$ and the normalization constant $G(S,\ N)$ is given by

$$G(S,\ N) = \sum_{n=0}^{S} \binom{N}{n} b^n. \tag{6}$$

The time congestion $B(S,\ N)$ is given by

$$B(S,\ N) \stackrel{\text{def}}{=} P(S,\ N) = 1 - \frac{G(S-1,\ N)}{G(S,\ N)}. \tag{7}$$

Because the arrival process is not Poisson in the finite source model, the call congestion $L(S,\ N)$ is no longer the same as $B(S,\ N)$, but we find the following simple relation:

$$L(S,\ N) = 1 - \frac{G(S-1,\ N-1)}{G(S,\ N-1)} = B(S,\ N-1). \tag{8}$$

More generally, the distribution of number of calls in service seen by an arriving call is the time average distribution that would be observed if the number of sources were reduced by one. This is analogous to the result that holds in an $M(N)/M/1$ queue or a machine servicing model (Kobayashi, 1978).

We now define a generalized Erlang loss model as follows:

1. *Multi-class sources*. We introduce a set, $\mathscr{C}$, of call classes. The arrival pattern of class c calls is a Poisson process with rate λ_c. We denote by $n_c(t)$ the number of class c calls in progress at time t.
2. *Simultaneous acquisition of multiple servers*. A class c call requires to hold A_c servers simultaneously. If the total number of servers or lines is S, then the following constraint must be met:

$$\sum_{c \in \mathscr{C}} A_c n_c(t) \leqslant S. \tag{9}$$

3. *Generally distributed holding time*. The call holding time distribution is a general distribution $G_c(t)$ with mean $1/\mu_c$:

$$\int_0^\infty (1 - G_c(t))dt = \frac{1}{\mu_c}. \tag{10}$$

Let the state process of this generalized loss station be denoted by $\mathbf{n}(t) = (n_c(t) : c \in \mathscr{C})$. Let $P(\mathbf{n})$ denote the equilibrium state distribution when there are S servers. The set of feasible states is

$$\mathscr{F}(S) = \left\{ \mathbf{n} \geqslant \mathbf{0} : \sum_{c \in \mathscr{C}} A_c n_c \leqslant S \right\}. \tag{11}$$

The departure process from the station includes both calls that have successfully completed service and those which are rejected. The generalized Erlang station shares many of the properties associated with stations in queueing networks.

A queueing station is said to be quasi-reversible if its state process $\mathbf{n}(t)$ is a stationary Markov process with the property that the state at an arbitrary time t_0 is independent of:

(i) the arrival times of class c calls, $c \in \mathscr{C}$, after time t_0; and
(ii) the departure times of class c calls, $c \in \mathscr{C}$, prior to time t_0.

The property of quasi-reversibility was introduced by Kelly (1979) to characterize a wide class of queueing stations which, together with certain rules governing call routing, gives rise to product-form queueing networks. We extend this definition to loss stations by assuming the convention that the departure process includes both calls that successfully complete service and those which are blocked and do not receive service. A closely related property is reversibility. A stochastic process $n(t)$ is reversible if it is statistically identical with its time-reversed process $n_R(t) = n(\tau - t)$ for any τ. For a stationary Markov process, reversibility holds if and only if its stationary distribution satisfies the detailed balance equations (Baskett et al., 1975; Kelly, 1979).

The following important theorem is proved in Kobayashi and Mark (1994):

Theorem 2.1. *The generalized Erlang station is quasi-reversible and its state-process* $\mathbf{n}(t)$ *is a reversible Markov process with stationary distribution given by*

$$P(\mathbf{n}|S) = \frac{1}{G(S)} \prod_{c \in \mathscr{C}} \frac{a_c^{n_c}}{n_c!}, \quad \mathbf{n} \in \mathscr{F}(S) \tag{12}$$

where $a_c = \lambda_c/\mu_c$ *and* $G(S)$ *is the normalization constant defined by*

$$G(S) = \sum_{\mathbf{n} \in \mathscr{F}(S)} \prod_{c \in \mathscr{C}} \frac{a_c^{n_c}}{n_c!}. \tag{13}$$

This result can be easily extended to the loss station model in which the calls are generated from a finite number of sources of multiple classes. We define a generalized Engset loss station as follows:

1. *Multi-class sources*. Let N_c be the number of sources for class c calls, $c \in \mathscr{C}$, and let $\mathbf{N}$ be the vector $\{N_c, c \in \mathscr{C}\}$. We denote by $n_c(t)$ the number of class c calls in progress at time t. Then, clearly

   $$n_c(t) \leqslant N_c, \quad c \in \mathscr{C}. \tag{14}$$

   The inter-generation time at a class c source is characterized by a general distribution $F_c(t)$ with mean $1/\nu_c$:

   $$\int_0^\infty (1 - F_c(t))dt = \frac{1}{\nu_c}. \tag{15}$$

2. *Simultaneous acquisition of multiple servers*. As in the generalized Erlang loss model.
3. *Generally distributed holding time*. As in the generalized Erlang loss model.

The set of feasible states is now given by

$$\mathscr{F}(S, \mathbf{N}) = \left\{ \mathbf{n} \geqslant \mathbf{0} : \sum_{c \in \mathscr{C}} A_c n_c \leqslant S; \quad n_c \leqslant N_c, c \in \mathscr{C} \right\} \tag{16}$$

The following theorem (Kobayashi and Mark, 1994) is a generalization of a result first reported by Cohen (1957).

Theorem 2.2. *For the generalized Engset loss system,* $\mathbf{n}(t)$ *is a reversible Markov process with stationary distribution:*

$$P(\mathbf{n}|S, \mathbf{N}) = \frac{1}{G(S, \mathbf{N})} \prod_{c \in \mathscr{C}} \binom{N_c}{n_c} b_c^{n_c}, \quad \mathbf{n} \in \mathscr{T}(S, \mathbf{N}) \tag{17}$$

where $b_c = \nu_c/\mu_c$, *and the normalization constant* $G(S, \mathbf{N})$ *is given by*

$$G(S, \mathbf{N}) = \sum_{\mathbf{n} \in \mathscr{T}(S,\mathbf{N})} \prod_{c \in \mathscr{C}} \binom{N_c}{n_c} b_c^{n_c}. \tag{18}$$

3. Loss networks

One can further extend the above generalized loss station (GLS) models by introducing multiple server types. In the generalized Erlang and Engset models, we extend the second property as follows:

2′. *Simultaneous acquisition of multiple servers of different types*. Let $\mathscr{L}$ denote a set of server types. There are S_l servers of type $l \in \mathscr{L}$. A class c call requires to hold A_{lc} servers of type l simultaneously. For each server type $l \in \mathscr{L}$, the following constraint must be met:

$$\sum_{c \in \mathscr{C}} A_{lc} n_c(t) \leqslant S_l. \tag{19}$$

The results of Theorems 3.1 and 3.2 (below) can be generalized straightforwardly to accommodate the concept of server types. Fig. 3 shows a generalized loss station in which each call can simultaneously acquire multiple servers from among several server types.

We define the properties of a loss network as follows:

1. Let $\mathscr{L}$ denote the set of links in the loss network. A link $\ell \in \mathscr{L}$ contains S_ℓ channels.
2. A call class $c \in \mathscr{C}$ is defined as a pair (r, τ), where r is the route or path of the call in the loss network and τ is the type of the call. The sets of routes and call types in the loss network are denoted by $\mathscr{R}$ and $\mathscr{T}$, respectively. Thus, $\mathscr{C} = \mathscr{R} \times \mathscr{T}$.[3]
3. A class c call seeks to simultaneously acquire $A_{\ell c}$ channels of link ℓ for each link $\ell \in \mathscr{L}$.
4. The holding time of a class c call has a general distribution $G_c(t)$ with mean $1/\mu_c$.

The loss network can be seen to be equivalent to a generalized loss station (GLS) with multiple server types, where each link in the loss network corresponds to a server type in the GLS. The loss network provides a general model for a circuit-switched network that carries multi-rate traffic (i.e., different values of $A_{\ell c}$ for different c) among different types τ of calls (Kelly, 1991; Kobayashi and Mark, 1997). The model is equally applicable to bidirectional flows. All that is required is to assign different class parameters to traffic in the reverse directions. The reverse traffic for a given pair of nodes may have

[3] In the loss station models discussed in the preceding section, the class $\mathscr{C}$ and the type $\mathscr{T}$ are equivalent. In the loss network, for a given source-destination pair, different types of call may take different routes.

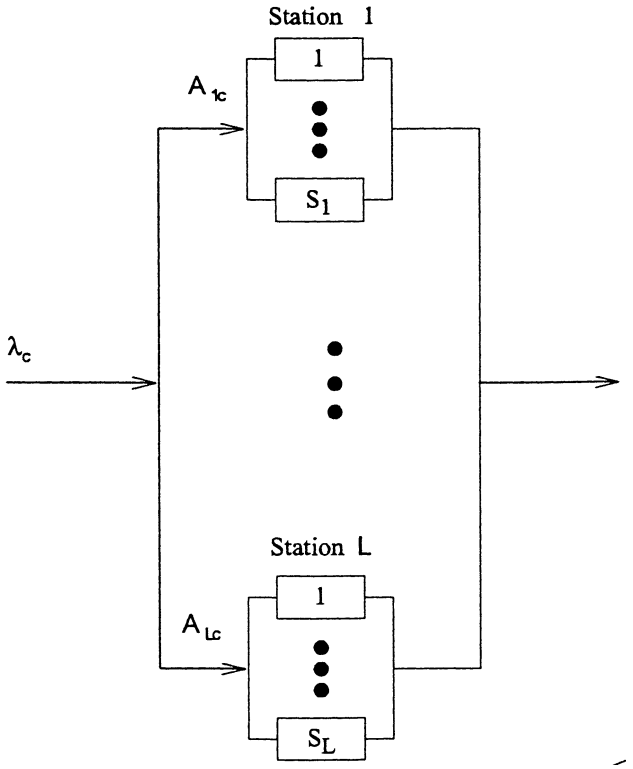

Fig. 3. Generalized loss station

different bandwidth requirements (i.e., values of $A_{\ell c}$ different from those for the forward direction). Similarly, the route used for the traffic in the reverse direction need not be the reverse of the forward route.

By specifying the arrival pattern of a loss network as a multi-class Poisson process as in property 1 of the generalized Erlang model in Section 2, we obtain an open loss network (OLN) (see Fig. 4). The OLN is equivalent to a generalized Erlang loss station with simultaneous acquisition among multiple server types. If we replace the multi-class Poisson process of the OLN by a multi-class finite source model as in property 1 of the generalized Engset model, we obtain a closed loss network (CLN). The CLN is equivalent to a generalized Engset loss station with multiple server types.

In the open loss network, the Poisson stream of class c arrivals is analogous to an open sub-chain in a queueing network (Baskett et al., 1975; Reiser and Kobayashi, 1975). Hence, in the OLN, each class c is said to be open. Similarly, dual to the concept of a closed sub-chain in a queueing network, we can define a *closed* class c in a loss network by replacing the Poisson stream of class c call arrivals by a finite source model of population N_c. The closed loss network is then a loss network wherein all the classes are closed. In a mixed loss network (MLN), as shown in Fig. 5, the set of call classes may be subdivided into the subset, $\mathscr{C}_O$, of open classes and the subset, $\mathscr{C}_C$, of closed classes, i.e., $\mathscr{C} = \mathscr{C}_O \cup \mathscr{C}_C$. The MLN further generalizes the generalized Erlang and Engset stations of the

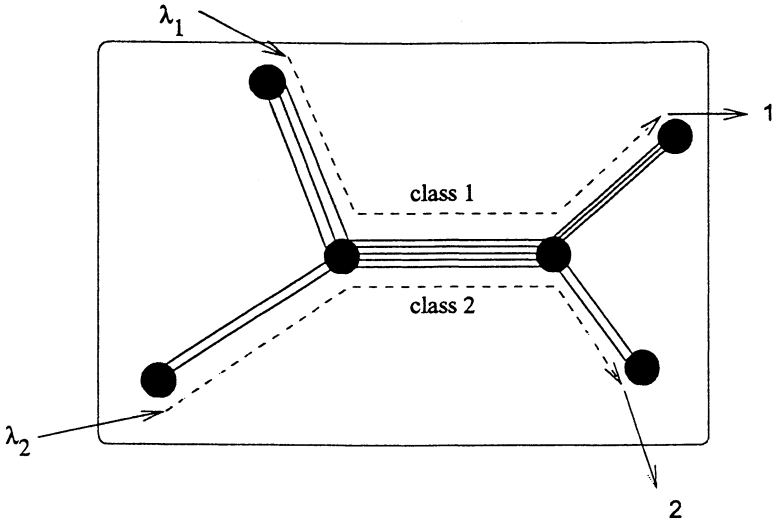

Fig. 4. Open loss network

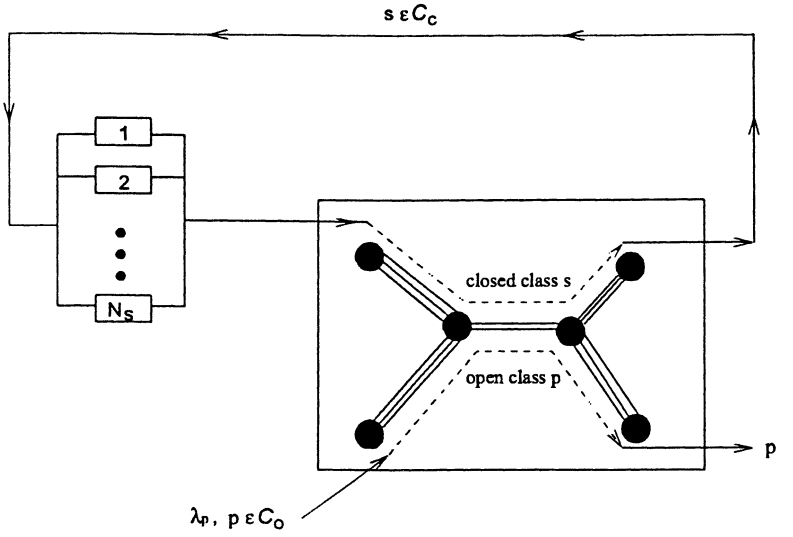

Fig. 5. Mixed loss network

previous section. Open, closed, and mixed loss networks are analogous to open, closed and mixed queueing networks (Baskett et al., 1975), respectively.

Denote the state process of a mixed loss network by $\mathbf{n}(t) = [\mathbf{n}_O(t), \mathbf{n}_C(t)]$, with $\mathbf{n}_O(t) = (n_p(t) : p \in \mathscr{C}_O)$ and $\mathbf{n}_C(t) = (n_s(t) : s \in \mathscr{C}_C)$. We have the following result for the MLN:

Theorem 3.1. *The state process of the mixed loss network is a reversible Markov process with equilibrium distribution given by*

$$P(\mathbf{n}|\mathbf{S},\mathbf{N}) = \frac{1}{G(\mathbf{S},\mathbf{N})} P_O(\mathbf{n}_O) P_C(\mathbf{n}_C|\mathbf{N}), \qquad \mathbf{n} \in \mathscr{F}(\mathbf{S},\mathbf{N}) \tag{20}$$

where

$$P_O(\mathbf{n}_O) = \prod_{p\in\mathscr{C}_O} \frac{a_p^{n_p}}{n_p!}, \qquad P_C(\mathbf{n}_C|\mathbf{N}) = \prod_{s\in\mathscr{C}_C} \binom{N_s}{n_s} b_s^{n_s} \tag{21}$$

with $a_p = \lambda_p/\mu_p$ $(p \in \mathscr{C}_O)$, $b_s = \nu_s/\mu_s$, $(s \in \mathscr{C}_C)$, *and*

$$\mathscr{F}(\mathbf{S},\mathbf{N}) = \left\{ \mathbf{n} \geqslant \mathbf{0};\ \sum_{c\in\mathscr{C}} A_{\ell c} n_c \leqslant S_\ell,\ \ell \in \mathscr{L};\ n_s \leqslant N_s,\ s \in \mathscr{C}_C \right\} \tag{22}$$

and

$$G(\mathbf{S},\mathbf{N}) = \sum_{\mathbf{n}\in\mathscr{F}(\mathbf{S},\mathbf{N})} P_O(\mathbf{n}_O) P_C(\mathbf{n}_C). \tag{23}$$

From the stationary distribution of the mixed loss network obtained above, we can express the time congestion and call congestion in terms of the normalization constant $G(\mathbf{S},\mathbf{N})$ as follows:

1. For calls belonging to an open class $p \in \mathscr{C}_O$:

$$B_p(\mathbf{S},\mathbf{N}) = 1 - \frac{G(\mathbf{S}-\mathbf{A}_p,\mathbf{N})}{G(\mathbf{S},\mathbf{N})} \tag{24}$$

$$L_p(\mathbf{S},\mathbf{N}) = B_p(\mathbf{S},\mathbf{N}), \tag{25}$$

   where $\mathbf{A}_p$ is the p-th column of the matrix $\mathbf{A} = [A_{\ell c}]$. The last equation is due to the PASTA property referred to earlier.
2. For calls belonging to a closed class $s \in \mathscr{C}_C$:

$$B_s(\mathbf{S},\mathbf{N}) = 1 - \frac{G(\mathbf{S}-\mathbf{A}_s,\mathbf{N})}{G(\mathbf{S},\mathbf{N})} \tag{26}$$

$$L_s(\mathbf{S},\mathbf{N}) = B_s(\mathbf{S},\mathbf{N}-\mathbf{1}_s), \tag{27}$$

   where $\mathbf{1}_s$ denotes the unit $|\mathscr{C}_C|$-vector whose s-th component is unity.

The above formulas for the time and call congestion are generalizations of the formulas obtained for the Erlang and Engset models. For numerical methods (exact, approximate, and asymptotic) to compute the normalization constants $G(\mathbf{S},\mathbf{N})$ for different values of $\mathbf{S}$ and $\mathbf{N}$, the reader is referred to Kobayashi and Mark (1997) and references cited therein.

4. Queueing-loss networks

Thus far, we have arrived at the mixed loss network by generalizing the classical Erlang and Engset loss models. We now carry the generalization further by introducing the concept of a queueing-loss network (QLN). A queueing-loss network (see Fig. 6) consists of a set of queueing sub-networks $\{\mathscr{Q}_j;\ j \in \mathscr{J}\}$ and a set of loss sub-networks $\{\mathscr{L}_k;\ k \in \mathscr{K}\}$. Calls are routed within each queueing sub-network and loss sub-network component as well as between queueing and loss network components. The call routing behavior can be governed by a Markov chain of arbitrary order (Kobayashi, 1978).

Each queueing sub-network, $\mathscr{Q}_j$, consists of a network of quasi-reversible queueing stations. Hence, if $\mathbf{n}_{\mathscr{Q}_j}$ denotes the population vector in the queueing sub-network $\mathscr{Q}_j$, its stationary state distribution $P_{\mathscr{Q}_j}(\mathbf{n}_{\mathscr{Q}_j})$ has the product form. Furthermore, the queueing network itself is quasi-reversible (Kelly, 1979). In general, each loss sub-network, $\mathscr{L}_k$, can be a mixed loss network (MLN). The loss network component of the MLN can be replaced by an equivalent generalized loss station (GLS) with simultaneous server acquisition. Each closed class in the MLN, representing a finite source population, can be decomposed as an infinite server (IS) station placed in tandem with the GLS (see Fig. 5). Let $\mathbf{n}_{\mathscr{L}_k}$ denote the population vector for the MLN. From Theorem 3.1, the state distribution $P_{\mathscr{L}_k}(\mathbf{n}_{\mathscr{L}_k})$ has the product form. In the decomposed representation of the MLN, each IS component is quasi-reversible, and by Theorem 2.1, the GLS component is also quasi-reversible.

Hence, the queueing-loss network can be decomposed into a set of quasi-reversible components. The

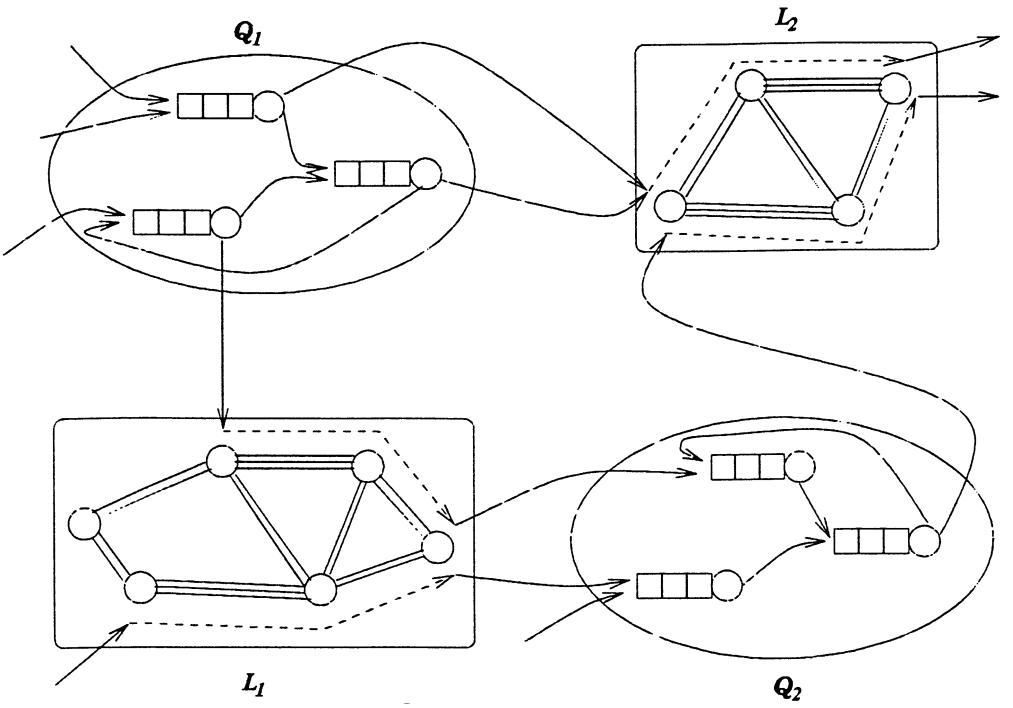

Fig. 6. Queueing-loss network

routing of calls between these components can be characterized by a Markov chain of arbitrary order. By combining these observations we now see that the QLN is a generalized queueing network. The only difference between this queueing network and those studied previously (Baskett et al., 1975; Kelly, 1979; Reiser and Kobayashi, 1975) is that it contains GLSs as its components. We have already established the fact that a GLS is a generalized version of an IS station and is a quasi-reversible station. Therefore, we can conclude that the QLN has a product-form solution. We state this general result for queueing-loss networks in the following theorem:

Theorem 4.1. *Consider a queueing-loss network* (*QLN*) *that contains a set of queueing sub-networks* $\{\mathscr{Q}_j;\ j \in \mathscr{J}\}$ *and a set of loss sub-networks* $\{\mathscr{L}_k;\ k \in \mathscr{K}\}$. *Let* $\mathbf{n}_{\mathscr{Q}_j}$ *and* $\mathbf{n}_{\mathscr{L}_k}$ *represent the population vectors in these sub-networks. The joint stationary distribution of the state process* $\mathbf{n}(t)$ *of the QLN takes the form:*

$$P(\mathbf{n}) = \frac{1}{G(\mathbf{S},\ \mathbf{N})} \prod_{j \in \mathscr{J}} P_{\mathscr{Q}_j}(\mathbf{n}_{\mathscr{Q}_j}) \prod_{k \in \mathscr{K}} P_{\mathscr{L}_k}(\mathbf{n}_{\mathscr{L}_k}), \tag{28}$$

where $P_{\mathscr{Q}_j}(\cdot)$ *and* $P_{\mathscr{L}_k}(\cdot)$ *themselves have product forms and are proportional to the marginal distributions of the sub-networks* $\mathscr{Q}_j$ *and* $\mathscr{L}_k$, $j \in \mathscr{J}$, $k \in \mathscr{K}$. *The normalization constant* $G(\mathbf{S},\ \mathbf{N})$ *and the feasible state set* $\mathscr{T}(\mathbf{S},\ \mathbf{N})$ *are defined over the capacity vector* $\mathbf{S}$ *of loss stations and the finite source vector* $\mathbf{N}$ *in the network. These vectors correspond to Cartesian products of the corresponding vectors of the queueing and loss sub-networks.*

5. Example of a queueing-loss network

A useful application of the queueing-loss network model may be found, for example, in a circuit-switched network in which call connection requests are served by either a centralized facility or distributed centers. Arriving calls may have to queue for the call-connection service if many such requests are already placed on the call-connection server. The call-connection server performs the function of admission control; i.e., it decides whether a new call can be accepted or not, based on the bandwidth resources requested by the call and the available resources of the network. In this application, the call-connection server is modeled by a queueing station, whereas the circuit-switched network itself is modeled by a loss network. The overall system is thus modeled by a queueing-loss network.

Consider the simple example of a QLN shown in Fig. 7. We label the three stations as stations 0, 1, and 2, respectively. Station 0 is an IS station, representing a finite source of population N. The inter-generation time of calls from each source is given by a general distribution. Station 1 is a single server queue representing, for example, a call-connection server. Station 2 is a loss station with S servers, and the call holding time can have a general distribution. We assume that $N > S$; otherwise, a call loss would not occur at this station. For the application discussed above, station 2 could be replaced by a more general loss network, representing, for example, a circuit-switched network.

As the results in the previous sections suggest, we can allow multiple classes of sources at station 0 and multiple types of servers at station 2. If station 1's queue discipline is FCFS (first-come, first-served) or any type of work-conserving queue discipline, then the service times at this station must be

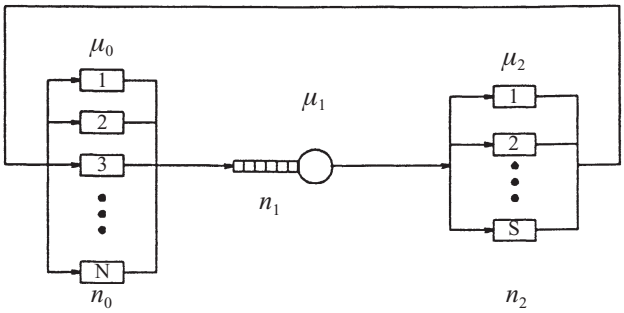

Fig. 7. Example of a queueing-loss network

drawn from the exponential service time that is common to all classes of customers (see e.g, Kobayashi, 1978). If the queue discipline of station 1 is either LCFS-PR (last-come, first-served with preemptive resume) or processor sharing (PS), then we can allow multiple classes for the service time and the distribution functions can be general, as long as their means are finite. We should also note that the processing rate of each station can be queue-dependent, i.e., the completion rate $\mu_i(n_i)$ of each server at station i can be an arbitrary function of its local queue size n_i, $i = 0, 1, 2$. For the IS station and the loss station, we allow the dependency $\mu_c(n_{ic})$ for different classes $c \in \mathscr{C}$, $i = 0, 2$. The same generality applies to a queueing station as well, if it adopts either LCFS-PR or PS.

For the sake of illustrative simplicity, we assume only one class of sources and a single type of server at the loss station, i.e., station 2. Further, we assume that the service rates are queue independent. Thus, the inter-generation time of each source at station 0 has mean $1/\mu_0$, the service time at station 1 is exponentially distributed with mean $1/\mu_1$, and the service time at station 2 has mean $1/\mu_2$. Using Theorem 4.1, we can write the stationary distribution of the queueing-loss network as:

$$P(n_0, n_1, n_2) \propto \frac{1}{n_0!}\left(\frac{\lambda}{\mu_0}\right)^{n_0}\left(\frac{\lambda}{\mu_1}\right)^{n_1}\frac{1}{n_2!}\left(\frac{\lambda}{\mu_2}\right)^{n_2} \propto \frac{1}{n_0!}\left(\frac{1}{\mu_0}\right)^{n_0}\left(\frac{1}{\mu_1}\right)^{n_1}\frac{1}{n_2!}\left(\frac{1}{\mu_2}\right)^{n_2}, \tag{29}$$

for (n_0, n_1, n_2) in the feasible set

$$\mathscr{F}(S, N) = \{(n_0, n_1, n_2) : n_0 + n_1 + n_2 = N;\ n_0, n_1 \geqslant 0;\ 0 \leqslant n_2 \leqslant S\}.$$

Here, λ is the rate of traffic through the closed route in the QLN, but this unknown parameter can be absorbed into the normalization constant. Hence, we can write

$$P(n_0, n_1, n_2) = \frac{1}{G(S, N)}\frac{1}{n_0!}\left(\frac{1}{\mu_0}\right)^{n_0}\left(\frac{1}{\mu_1}\right)^{n_1}\frac{1}{n_2!}\left(\frac{\mu_0}{\mu_2}\right)^{n_2}, \tag{30}$$

where

$$G(S,\,N) = \sum_{(n_0,n_1,n_2)\in\mathcal{F}(S,N)} \frac{1}{n_0!}\left(\frac{1}{\mu_0}\right)^{n_0}\left(\frac{1}{\mu_1}\right)^{n_1}\frac{1}{n_2!}\left(\frac{1}{\mu_2}\right)^{n_2} \tag{31}$$

$$= \frac{1}{\mu_1^N}\sum_{n_2=0}^{S}\sum_{n_0=0}^{N-n_2}\frac{1}{n_0!}\left(\frac{\mu_1}{\mu_0}\right)^{n_0+n_2}\frac{1}{n_2!}\left(\frac{\mu_0}{\mu_2}\right)^{n_2}. \tag{32}$$

The time congestion and call congestion at station 2 are then given by

$$B(S,\,N) = 1 - \frac{G(S-1,\,N)}{G(S,\,N)}, \tag{33}$$

$$L(S,\,N) = B(S,\,N-1) = 1 - \frac{G(S-1,\,N-1)}{G(S,\,N-1)}. \tag{34}$$

Suppose that we wish to find ρ_1, the utilization of station 1. By extending results known for closed queueing networks (see e.g., Kobayashi, 1978), we can write

$$\rho_1 = 1 - \frac{G^{(-1)}(S,\,N)}{G(S,\,N)}, \tag{35}$$

where $G^{(-1)}(S,\,N)$ represents the value of the normalization constant when station 1 is deleted from the system. This corresponds to the situation which would arise if we let $\mu_1 \to \infty$ in the above queueing-loss system. In the limit as $\mu_1 \to \infty$, only the terms corresponding to $n_1 = 0$ remain in (31) and we obtain the following expression for $G^{(-1)}(S,\,N)$:

$$G^{(-1)}(S,\,N) = \frac{1}{N!}\left(\frac{1}{\mu_0}\right)^N G_L(S,\,N), \tag{36}$$

where

$$G_L(S,\,N) = \sum_{n=0}^{S}\binom{N}{n}\left(\frac{\mu_0}{\mu_2}\right)^n \tag{37}$$

is the normalization constant of the Engset loss station resulting from deleting station 1 from the QLN. If station 1 has a constant rate, as in the present case, we can use the following alternative formula (see e.g., Kobayashi, 1978, p. 172):

$$\rho_1 = \frac{1}{\mu_1}\frac{G(S,\,N-1)}{G(S,\,N)}. \tag{38}$$

It is not difficult to confirm that the above two formulas for the server utilization ρ_1 are indeed equivalent.

To study the effect of the queueing station 1 on the system capacity of loss station 2, one can express the marginal distribution of station 2 as a function of μ_1 as follows:

$$P(n_2,\,\mu_1)=\frac{\binom{N}{n_2}\left(\frac{\mu_0}{\mu_2}\right)^{n_2}+N!\left(\frac{\mu_0}{\mu_1}\right)^{N}\sum_{n=0}^{N-n_2-1}\frac{1}{n!}\left(\frac{\mu_1}{\mu_0}\right)^{n+n_2}}{\sum_{m=0}^{S}\binom{N}{m}\left(\frac{\mu_0}{\mu_2}\right)^{m}+N!\left(\frac{\mu_0}{\mu_1}\right)^{N}\sum_{m=0}^{S}\sum_{n=0}^{N-m-1}\frac{1}{n!}\left(\frac{\mu_1}{\mu_0}\right)^{n+m}}. \tag{39}$$

Clearly, as $\mu_1 \to \infty$, $P(n_2,\,\mu_1)$ approaches the marginal distribution of an Engset loss station. One can further show that the time congestion at station 2, given by $P(S,\,\mu_1)$, is a monotonically increasing function of μ_1. The behavior of $P(S,\,\mu_1)$ as a function of μ_1 quantifies the trade-off between time congestion (or similarly, call congestion) in station 2 and queueing delay in station 1. As $\mu_1 \to \infty$, the queueing delay decreases to zero while the time congestion at station 2 increases to that of an Engset loss station. Thus, the effect of the queueing station is to alleviate call blocking in the loss station at the expense of introducing queueing delay. Alternatively, for a given call-loss probability, the QLN is able to handle a larger population N than the Engset station, provided that the queueing delay introduced in the QLN can be tolerated.

We now point out an equivalence between the QLN of Fig. 7 and the open loss network (OLN) defined in Section 3. Using the fact that the variables in (29) must satisfy $n_0 + n_1 + n_2 = N$, we can write the stationary distribution of $(n_0,\,n_2)$ as:

$$P(n_0,\,n_2) \propto \frac{1}{n_0!}\left(\frac{\mu_1}{\mu_0}\right)^{n_0}\frac{1}{n_2!}\left(\frac{\mu_1}{\mu_2}\right)^{n_2}. \tag{40}$$

Hence,

$$P(n_0,\,n_2)=\frac{1}{\tilde{G}(S,\,N)}\frac{a_0^{n_0}}{n_0!}\frac{a_2^{n_2}}{n_2!}, \tag{41}$$

where $a_0 = \mu_1/\mu_0$, $a_2 = \mu_1/\mu_2$ and

$$\tilde{G}(S,\,N)=\sum_{0\le n_0+n_2\le N,\,n_2\le S}\frac{a_0^{n_0}}{n_0!}\frac{a_2^{n_2}}{n_2!}. \tag{42}$$

We observe that (41) is the stationary state distribution of an open loss network with two links, l_1 and l_2, having link capacities N and S, respectively. There are two traffic classes, c_0 and c_2. Calls of class c_0 arrive according to a Poisson process of rate a_0 and use a route containing just link l_1. Calls of class c_2 arrive according to a Poisson process of rate a_2 and use a route containing links l_1 and l_2. The matrix $A = [A_{lc} : l \in \{l_1,\,l_2\},\, c \in \{c_0,\,c_2\}]$, which indicates the resource requirements of the OLN, is given by

$$A=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}. \tag{43}$$

Thus, with respect to the stationary probability state distribution of the two loss stations in Fig. 7, the simple QLN is equivalent to an OLN. Furthermore, certain methods for computing the normalization constant $\tilde{G}(S,\,N)$ for the OLN may be used to solve for the probability state distribution of the QLN.

6. Conclusion

In this paper we introduced the queueing-loss network model as a generalization of classical loss and queueing models. By further generalizing results on loss networks (Kobayashi and Mark, 1994, 1997), we showed that the product-form solution applies to this extended class of stochastic models. Queueing-loss networks allow multiple classes of calls, multiple types of servers, general call service times, and general call inter-generation time distributions. A key observation in making this generalization was that an entire open loss network (OLN) or sub-network could be replaced by a single generalized loss station (GLS).

The queueing-loss network can be used to model systems which involve both queueing and loss behaviors. For example, arriving calls to a circuit-switched network may first have to wait at a queueing station prior to being subjected to admission control. For a small three-stage closed queueing-loss network, we have shown how various performance measures can be calculated. For general queueing-loss networks, performance measures such as time congestion, call congestion at loss stations, and utilization at queueing stations can be expressed in terms of the normalization constant $G(\mathbf{S}, \mathbf{N})$. For a large network with large values in $\mathbf{S}$ (the vector of number of servers at various loss stations) and/or large values in $\mathbf{N}$ (the vector of the number of sources in closed classes), a direct evaluation of the normalization constants $G(\mathbf{S}, \mathbf{N})$ becomes computationally intensive. A large body of literature exists that addresses the computational aspect of approximating the normalization constant for loss networks (see e.g., Kobayashi and Mark, 1997 and references cited therein).

Future work could investigate the behaviors of more complicated queueing-loss network models and, in particular, the interaction between the queueing and loss aspects. A limitation of the queueing-loss network models discussed in this paper is that the subsequent behavior of calls, after being blocked at a loss station, is identical with that of calls which have successfully received service; i.e., they proceed to follow the same path. In practice, blocked calls are often tagged as such and subsequently receive different treatment from their successfully served counterparts. Although the product-form solution will no longer hold in this case, the practical implications of such a model make its thorough investigation an interesting open problem.

Acknowledgements

This research has been supported, in part, by the National Science Foundation, the New Jersey Commission on Science and Technology, and the Ogasawara Foundation for the Promotion of Science and Technology.

References

Baskett, F., Chandy, K.M., Muntz R.R. and Palacios, F.C., 1975. Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, *Journal of ACM* 22, 2, 248–260.

Cohen, J.W., 1957. The Generalized Engset Formulae. *Philips Telecommunication Review* 18(4), 158–170.

Jackson, J.R. 1963. Jobshop-like Queueing Systems, *Management Science* 10(1), 131–142.

Kelly, F.P., 1979. *Reversibility and Stochastic Networks*. Wiley, Chichester.

Kelly, F.P. 1991. Loss Networks. *Annals of Applied Probability* 1(3), 319–378.

Kobayashi, H. 1978. *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*. Addison-Wesley.

Kobayashi, H. and Mark, B.L., 1994. On Queueing Networks and Loss Networks, Proc. of the 28th Annual Conference on Information Sciences and Systems. Princeton, New Jersey, pp. 794–799.

Kobayashi, H. and Mark, B.L., 1997. Product-Form Loss Networks. In Dshalalow, J. (Ed.) *Frontiers in Queueing: Models and Applications in Engineering and Science*, CRC Press, pp. 147–195.

Reiser, M. and Kobayashi, H., 1975. Queueing Networks with Multiple Closed Chains: Theory and Computational Algorithms. *IBM Journal of Research and Development* 19, 283–294.

Syski, R., 1986. *Introduction to Congestion Theory in Telephone Systems*, 2nd edn, Elsevier Science, Amsterdam.

Wolff, R., 1989. *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs.