ELSEVIER

Performance Evaluation 29 (1997) 15–33

PERFORMANCE EVALUATION
An International Journal

Study of a class of partially ordered service strategies for a system of two discrete-time queues ★

Ioannis Stavrakakis [a,*], Sophia Tsakiridou [b]

[a] *Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA*
[b] *Department of Electrical and Computer Engineering, Queen's University, Kingston, Ont., Canada K7L 3N6*

Received 2 February 1993; revised 20 November 1995

Abstract

A discrete-time system of two queues served according to distinct policies is studied in this paper: one queue receives 1-limited service; the other is served according to some policy from the introduced class of (state dependent) policies $\mathcal{S}$. Class $\mathcal{S}$ contains traditional service policies as well as ad hoc policies. A partial performance ordering of policies in $\mathcal{S}$ is derived which can be useful in designing ad hoc policies with improved performance compared to that of traditional policies. Some examples of ad hoc policies in $\mathcal{S}$ are presented and numerical results are derived to illustrate the potential for efficiency provided by service policies in $\mathcal{S}$.

Keywords: Discrete-time queueing systems; M/G/1 type Markov chain; Matrix analytic technique; Service strategies

1. Introduction

Discrete-time queueing systems have widely been adopted for the study of packet communication networks, where packet processes are described in terms of discrete-time stochastic point processes. The service policies associated with such systems model the protocols that govern the availability of network resources to the packets, users or classes of users. The distributed nature of a networking system and/or the diversified service requirements of the supported users necessitate the development of sophisticated resource allocation protocols modeled in terms of queueing systems under various service policies. Examples are the exhaustive, gated and limited service policies which have been studied extensively, primarily in continuous time, in the context of polling [1,2], and vacation systems [3], and the consistent gated/limited policy [4] which has been applied for the study of the DQDB network [5,6].

★ Research supported in part by the National Science Foundation under grant NCR-9011962 and by the Advanced Research Project Agency (ARPA) under Grant F49620-93-1-0564 monitored by the Air Force Office of Scientific Research (AFOSR). This work was conducted while the authors were with the Department of Computer Science and Electrical Engineering of the University of Vermont, Burlington, VT 05405, USA.

* Corresponding author. Tel.: (617) 373-3053; fax: (617) 373-8970.

PII S0166-5316(96)00002-8

A discrete-time single server queueing system consisting of two queues is considered in this paper. One of the queues receives 1-limited service, the other is served according to a policy contained in a versatile class of state-dependent service policies $\mathcal{S}$. This class contains some of the well-known service policies as well as new ad hoc policies which may provide for improved performance compared to that under traditional policies.

The system studied in this paper may be considered as a variant of a two-queue asymmetric polling system with 1-limited service at one queue and service according to a policy in $\mathcal{S}$ at the other. When both queues receive 1-limited service the system operates under an alternating service policy, which has been analyzed in continuous time [7]. A continuous-time two-queue model with mixed exhaustive and limited services has been studied in [8]. A discrete-time model with gated and 1-limited services has been analyzed in [4].

The system considered in this paper may also be described in terms of a GI/D/1 infinite queue with server vacations. The vacation period distribution depends on the occupancy of the queue which is served during the vacation. The queueing model and analysis developed in this paper are also applicable to a system with vacation period distributions that depend on the state of a Markov process which evolves independently of the system [9]. This state-dependent vacation model is different from most of those presented in the past, for example [10,11], in the following aspect: the process whose state determines the occurrence and distribution of vacation periods is external to the queue under study. The analysis is based on matrix analytic techniques similar to those applied for the study of the vacation model in [12].

The detailed description of the queueing system and a unified representation of the class of service policies $\mathcal{S}$ are presented in Section 2. In Section 3, a queueing model for the system is formulated and analyzed. The joint probability distribution of the queue occupancies is derived by applying matrix analytic methods and Markov renewal theory arguments. Section 4 presents a partial performance ordering for policies in $\mathcal{S}$. Some numerical results which illustrate the effect of various service disciplines on the performance of the system are presented in Section 5. The work is summarized in Section 6.

2. Description of the queueing system

2.1. Introduction

Consider the discrete-time queueing system shown in Fig. 1. The packet service time is assumed to be constant and equal to the system time unit (slot). Unless otherwise stated, a superscript I (F) will indicate a quantity associated with queue Q^{I} (Q^{F}). The capacities of queues Q^{F} and Q^{I} are assumed to be $N < \infty$ and infinite, respectively. Let $\{q_j^{\mathrm{I}}\}_{j\geq 0}$ and $\{q_j^{\mathrm{F}}\}_{j\geq 0}$ denote the associated queue occupancy processes with state spaces $R^{\mathrm{I}} = \{0, 1, 2, \ldots\}$ and $R^{\mathrm{F}} = \{0, 1, \ldots, N\}$.

The system service discipline is determined by the general rules presented in Section 2.2 and the adopted service policy in $\mathcal{S}$ presented in Section 2.3. A brief introduction to the service discipline is presented first.

The server never idles as long as there is service to be performed. Upon switching to Q^{F}, the server serves a number of F-packets (packets in Q^{F}) which depends on the number of packets found in Q^{F} and the number of packets left in Q^{I}. If the number of F-packets served cannot exceed the number of F-packets already present in Q^{F} at the switching instant, then this state-policy (see Section 2.3) will be called gated. Otherwise, it will be called non-gated. If all state-policies of a service policy are gated (non-gated) the service policy will also be called gated (non-gated). Following the service to Q^{F} the server switches to Q^{I}

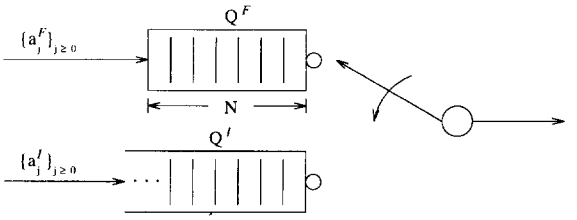


Fig. 1. The queueing system.

and provides 1-limited service to that queue. Although a service policy other than the 1-limited could be considered for Q^{I}, the analysis developed in this work is applicable only under 1-limited service to Q^{I}.

2.2. *The general service mechanism*

The system is work conserving; that is, the server is never idle when the system is non-empty. The server is assumed to be at Q^{F} when idle. Let a *decision time-instant* t^{d} be defined as the time instant at which a packet is forwarded to the head of Q^{I}; the collection of decision time-instants is denoted by $\mathcal{A}$.

(i) At t^{d}, $t^{\mathrm{d}} \in \mathcal{A}$, the server switches to Q^{F}; the amount of service provided to this queue is determined by the adopted policy in $\mathcal{S}$ (Section 2.3). Upon completion of this service to Q^{F}—which can possibly be equal to 0—the server switches to Q^{I} and serves one packet.

(ii) Upon completion of the 1-limited (packet) service to Q^{I}:
   (a) If Q^{I} is left non-empty, a packet is instantaneously forwarded to the head of Q^{I} and, thus, a decision time-instant t^{d} is reached.
   (b) If Q^{I} is left empty, Q^{F} is served uninterruptedly until the next decision time-instant t^{d}, to be determined by the first future packet arrival to Q^{I}.

Notice that before service is provided to a packet forwarded to the head of Q^{I}, an amount of service is provided to Q^{F}, as determined by a policy in $\mathcal{S}$. In general, the amount of service provided to Q^{F}—determined by the policy in $\mathcal{S}$—will depend on the state (queue occupancy) of Q^{F}. Thus, in general, the server behavior is 'regulated' by the state of Q^{F}. Since the server considers switching back to Q^{F} at any (discrete) time instant at which it is away from this queue, Q^{F} may be viewed as the critical queue. Although almost the entirety of the adopted service disciplines for Q^{F} would provide prioritized service to this queue, there is at least one such service discipline under which the priority is reversed.

2.3. *The class of service policies* $\mathcal{S}$

As will become clear, a number of traditional service disciplines, as well as ad hoc policies—potentially improving the performance achieved by traditional ones—can be represented by selecting a proper service discipline in $\mathcal{S}$.

Let i, $0 \leq i \leq N$, denote the state of $\{q_j^{\mathrm{F}}\}_{j\geq 0}$ at some decision time-instant t^{d}, $t^{\mathrm{d}} \in \mathcal{A}$. A *state-policy* associated with state i, $\mathcal{P}_i$, is defined to be equal to the *potential* amount of service (in packets) provided to Q^{F} by the server, immediately after t^{d}, $0 \leq \mathcal{P}_i \leq \infty$. This amount of service will be provided to Q^{F} unless the queue becomes empty. In view of rule (i), the definition of $\mathcal{P}_i$ and the work-conserving nature of the service discipline, the server will switch to Q^{I} at $t^{\mathrm{d}} + \tau(i)$, where $\tau(i) = \min\{\mathcal{P}_i, t^{\mathrm{F}}(i)\}$ and $t^{\mathrm{d}} + t^{\mathrm{F}}(i)$

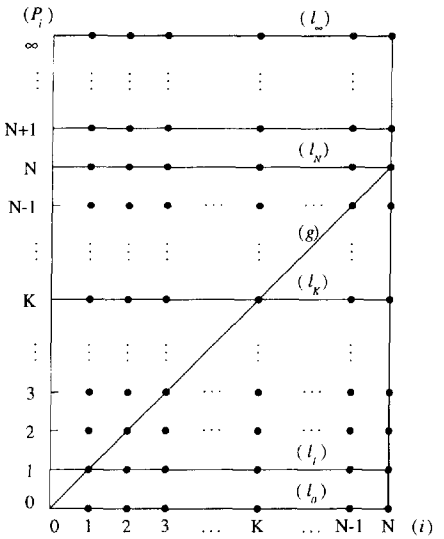


Fig. 2. The space of state-policies $\mathcal{SP}$ (each dot marks a state-policy point).

denotes the first time-instant following t^d at which Q^F becomes empty; $\tau(q^F_{t^d})$ will be referred to as the *service horizon* following t^d or determined by $q^F_{t^d}$.

A service policy in $\mathcal{S}$ will be defined in terms of the state-policies $\mathcal{P}_i$ associated with all states i, $0 \leq i \leq N$. Since the service policies considered here are work-conserving, it is easily established that any selection for $\mathcal{P}_0$ will induce identical server behavior as $\mathcal{P}_0 = 0$. $\mathcal{P}_0 = 0$ will be assumed for any policy in $\mathcal{S}$ and, thus, a policy in $\mathcal{S}$ will be defined in terms of $\mathcal{P}_i$, $1 \leq i \leq N$. A service policy in $\mathcal{S}$ will be uniquely represented by the N-dimensional vector $\bar{\mathcal{P}} = (\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_N)$, $\bar{\mathcal{P}} \in \mathcal{S} = \mathbb{Z}_0^+ \times \mathbb{Z}_0^+ \times \cdots \times \mathbb{Z}_0^+$, where $\mathbb{Z}_0^+$ denotes the set of non-negative integers. As will become clear below, this vector representation of a service policy in terms of the state-policies $\{\mathcal{P}_i\}_{i=1}^N$ not only provides for a unified description (and study) of many well known service policies, but also allows for the introduction of new ad hoc policies. The space of state-policies $\mathcal{SP}$ is shown in Fig. 2. A service policy $\bar{\mathcal{P}} \in \mathcal{S}$ may be graphically represented by the set of *state-policy points* $\{(i, \mathcal{P}): 1 \leq i \leq N\}$, or the lines connecting these points (*policy-lines*) on $\mathcal{SP}$.

The following definitions will provide for a classification of state-policies and service policies in $\mathcal{S}$.

Definition 1. A state-policy $\mathcal{P}_i$ will be called *gated* if only packets that are present in Q^F at some $t^d \in \mathcal{A}$ at which $q^F_{t^d} = i$ will be served over the service horizon following t^d; otherwise, it will be called *non-gated*. It will be called *unlimited (limited) gated* if all (not all) packets that are present in Q^F at some $t^d \in \mathcal{A}$ at which $q^F_{t^d} = i$ will be served over the service horizon following t^d. It will be called *unlimited non-gated* if exhaustive service is provided to Q^F; otherwise, it will be called *limited non-gated*.

Definition 2. A service policy $\bar{\mathcal{P}} = (\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_N) \in \mathcal{S}$ will be called (*limited* or *unlimited*) *gated* if all associated state-policies are (limited or unlimited) gated. It will be called (*limited* or *unlimited*) *non-gated* if all associated state-policies are (limited or unlimited) non-gated.

Based on Definition 1, it can be easily established that $\mathcal{P}_i$ is gated if $\mathcal{P}_i \leq i$. In this case, $\mathcal{P}_i$ packets will be served over the service horizon $\tau(i)$, and thus, $\Pr\{\tau(i) = \mathcal{P}_i\} = 1$. $\mathcal{P}_i$ is unlimited gated if $\mathcal{P}_i = i$ and limited gated if $\mathcal{P}_i < i$. Similarly, $\mathcal{P}_i$ is non-gated if $\mathcal{P}_i > i$. In this case, *at least* i packets will be served over the service horizon $\tau(i)$ and $\Pr\{\tau(i) = \mathcal{P}_i\} < 1$. $\mathcal{P}_i$ is unlimited non-gated if $\mathcal{P}_i = \infty$ and limited non-gated if $\mathcal{P}_i < \infty$.

If $\bar{\mathcal{P}}$ is gated, the amount of service provided to Q^{F} over any service horizon $\tau(i)$ is equal to $\mathcal{P}_i$, for all i, $1 \leq i \leq N$, thus; $\Pr\{\tau(i) = \mathcal{P}_i\} = 1$. The subclass of gated service policies $\mathcal{S}_{\mathrm{G}}$ is defined on the subspace $\mathcal{SP}_{\mathrm{G}} = \{(i, \mathcal{P}_i): 0 \leq \mathcal{P}_i \leq i,\ 0 \leq i \leq N\}$ of $\mathcal{SP}$.

If $\bar{\mathcal{P}}$ is non-gated, the amount of service provided to Q^{F} over a service horizon $\tau(i)$ is at least equal to $\mathcal{P}_i$ and depends on the evolution of $\{q_j^{\mathrm{F}}\}_{j \geq 0}$ over that service horizon; thus, $\Pr\{\tau(i) = \mathcal{P}_i\} < 1$. The subclass of non-gated service policies $\mathcal{S}_{\mathrm{NG}}$ is defined on the subspace $\mathcal{SP}_{\mathrm{NG}} = \{(i, \mathcal{P}_i): \mathcal{P}_i > i, 0 \leq i \leq N\}$ of $\mathcal{SP}$.

Based on the above definitions, the space of state-policies $\mathcal{SP}$ can be divided into the subspaces of gated, $\mathcal{SP}_{\mathrm{G}}$, and non-gated, $\mathcal{SP}_{\mathrm{NG}}$, state-policies; a policy-line in $\mathcal{SP}_{\mathrm{G}}$ ($\mathcal{SP}_{\mathrm{NG}}$) defines a gated (non-gated) service policy in $\mathcal{S}$. The boundary between the two subspaces identifies the diagonal policy-line (g) (Fig. 2) which represents the unlimited gated service policy. The following definition provides for another classification of service policies in $\mathcal{S}$.

Definition 3. A service policy $\bar{\mathcal{P}}$ is k-limited, if $\mathcal{P}_i = k$, $1 \leq i \leq N$, for some k, $0 \leq k \leq \infty$. Note that a k-limited service policy is not state-dependent. This is also indicated by the graphical representation of the k-limited service policies in $\mathcal{SP}$; each k-limited service policy, $0 \leq k \leq \infty$, is represented by a horizontal policy-line (l_k) (Fig. 2).

The state-policies $\mathcal{P}_i$ associated with a k-limited service policy $\bar{\mathcal{P}}$ may be gated ($i \geq k$) or non-gated ($i < k$). For the extreme values of k, $k = 0$ and $k = \infty$, the system service policy becomes HoL priority for queues Q^{I} and Q^{F}, respectively. For $k = 1$, the system operates under an alternating service policy; a single packet is served at each queue, in a alternating manner, while both queues are non-empty.

2.4. *An expansion of class* $\mathcal{S}$, $\mathcal{S}_{\mathrm{exp}}$

Based on the definition of a state-policy stated at the beginning of Section 2.3, the potential amount of service over any service horizon $\tau(i)$, for any policy $\bar{\mathcal{P}} \in \mathcal{S}$ described so far, is deterministic and equal to $\mathcal{P}_i$. A generalized definition of a state-policy is introduced below to describe probabilistically limited service policies [13] and expand the class of service policies in $\mathcal{S}$. A probabilistically limited state-policy associated with state i, $1 \leq i \leq N$, may be defined in terms of the probability vector $\bar{g}_i = [g_{i0}\ g_{i1}\ g_{i2} \ldots]^{\mathrm{T}}$, where g_{ij}, $1 \leq i \leq N$, $j \geq 0$, denotes the probability that the *potential* amount of service provided to Q^{F} over a service horizon $\tau(i)$ is equal to j. A service policy in the expanded class of service policies $\mathcal{S}_{\mathrm{exp}}$ may be represented by a matrix $\tilde{\mathcal{P}}$: $\tilde{\mathcal{P}} = [\bar{g}_1\ \bar{g}_2\ \bar{g}_3 \ldots \bar{g}_N]$.

Clearly, $\mathcal{S} \subset \mathcal{S}_{\mathrm{exp}}$. A service policy $\bar{\mathcal{P}} \in \mathcal{S}$ can be expressed in terms of a matrix $\tilde{\mathcal{P}}$ by assigning the proper distributions to every (deterministic) state-policy i:

$$g_{ij} = \begin{cases} 1 & \text{for } j = \mathcal{P}_i, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \leq i \leq N.$$

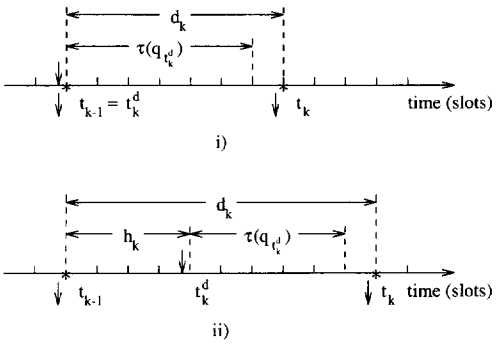


Fig. 3. Illustration of the definitions h_k and d_k: (i) $h_k = 0$, (ii) $h_k > 0$. An arrow pointing to (away from) the time axis indicates the movement of a packet to (away from) the head of Q^{I}.

A probabilistically limited service policy is gated if $\sum_{j=0}^{i} g_{ij} = 1$ and non-gated if $\sum_{j=i+1}^{\infty} g_{ij} = 1$, $1 \le i \le N$. Examples of probabilistically limited service policies are the Bernoulli gated:

$$g_{ij} = \begin{cases} 1 - p_i, & j = 0, \\ p_i, & j = 1, \\ 0, & \text{otherwise}, \end{cases} \qquad 1 \le i \le N,$$

and the binomial gated:

$$g_{ij} = \begin{cases} \binom{i}{j} p_i^j (1 - p_i)^{i-j}, & 0 \le j \le i, \\ 0, & \text{otherwise}, \end{cases} \qquad 1 \le i \le N,$$

where $0 \le p_i \le 1$, $1 \le i \le N$, and $\binom{i}{j}$ is the binomial coefficient.

For $p_1 = p_2 = \cdots = p_N = 1$ the Bernoulli gated policy becomes 1-limited and the binomial gated service policy becomes unlimited gated. For $p_1 = p_2 = \cdots = p_N = 0$ both policies become HoL priority at Q^{I}. In the rest of this paper, only service policies in $\mathcal{S}$ will be considered.

3. Modeling and analysis of the queueing system

The queueing behavior of the system under the adopted service discipline is analyzed in this section under i.i.d. and mutually independent arrival processes $\{a_j^{\mathrm{F}}\}_{j\ge 0}$ and $\{a_j^{\mathrm{I}}\}_{j\ge 0}$ to Q^{F} and Q^{I}, respectively. The joint probability distribution of the occupancy at queues Q^{I} and Q^{F} upon service completion of a packet at Q^{I} is derived first, based on a matrix analytic approach. The joint probability distribution of the queue occupancies at arbitrary time instants is then obtained by applying Markov renewal theory arguments. Consider the following time sequences defined at the end of slots:

$\{t_k\}_{k\ge 0}$: Defined to be the sequence of time instants at which the service of a packet is completed at Q^{I} ($t_0 = 0$).

$\{t_k^{\mathrm{d}}\}_{k\ge 1}$: Defined to be the sequence of decision time-instants, or time instants at which a packet is forwarded to the head of $Q^{\mathrm{I}}(\{t_k^{\mathrm{d}}\}_{k\ge 1} \equiv \mathcal{A})$.

According to the system service discipline rules described in Section 2.2, the server switches to Q^F at t_{k-1} for an amount of time which depends on $q^I_{t_{k-1}}$ and $q^F_{t^d_k}$:

(i) If $q^I_{t_{k-1}} > 0$, by rule (ii(a)), the kth packet is instantaneously forwarded to the head of Q^I, and thus, $t_{k-1} = t^d_k$ (Fig. 3(i)). In this case, the server will remain at Q^F for $\tau(q^F_{t^d_k})$ slots.

(ii) If $q^I_{t_{k-1}} = 0$, by rule (ii(b)), Q^I will remain empty for a time interval h_k, until the first future packet arrival to Q^I; at $t_{k-1} + h_k$ the decision time-instant t^d_k is reached (Fig. 3(ii)). In this case, the server will remain at Q^F for $h_k + \tau(q^F_{t^d_k})$ slots.

The service of the kth packet at Q^I will be completed at time instant $t_k = t^d_k + \tau(q^F_{t^d_k}) + 1$ and, therefore, the time interval d_k between the successive service completions of the $(k-1)$st and kth packets at Q^I is given by $d_k = t_{k-1} - t_k = h_k + \tau(q^F_{t^d_k}) + 1$ where $h_k = 0$, for $q^I_{t_{k-1}} > 0$ (Fig. 3). It will be shown in the following that the process $\{q^I_{t_k}, q^F_{t_k}\}_{k\geq 0}$ is a Markov chain embedded at service completion time instants $\{t_k\}_{k\geq 0}$. In order to proceed with the analysis the following random variables need to be defined:

a^I (a^F) : Denotes the number of packet arrivals to Q^I (Q^F) in an arbitrary slot. The probability distribution of a^I (a^F) is denoted by $p^I_a(n)$ $(p^F_a(n))$, $n \geq 0$.

$a^{I,l}$: Denotes the number of packet arrivals to Q^I over l slots. It is distributed according to $p^{I,l}_a(n)$, the l-fold convolution of $p^I_a(n)$, $n \geq 0$ $(p^{I,1}(n) \equiv p^I_a(n))$.

h : Denotes the generic random variable for h_k; it describes the time between the service completion of a packet that leaves Q^I empty and the first packet arrival to the empty queue. The distribution of random variable h is given by $p_h(n) = [p^I_a(0)]^{n-1}[1 - p^I_a(0)]$, $n \geq 1$.

Proposition 1. *The process $\{q^I_{t_k}, q^F_{t_k}\}_{k\geq 0}$ is a two-dimensional Markov chain embedded on packet service completion time instants, $\{t_k\}_{k\geq 0}$.*

Proof. It suffices to show that $q^I_{t_k}$ and $q^F_{t_k}$ can probabilistically be determined from $q^I_{t_{k-1}}$ and $q^F_{t_{k-1}}$. The evolution of the occupancy process $\{q^F_j\}_{j\geq 0}$ is determined based on the server availability to Q^F; transitions take place at the end of slots. If the server is at Q^F at the end of a slot, then process $\{q^F_j\}_{j\geq 0}$ makes a transition according to the probability matrix $\boldsymbol{P}_s$,

$$\boldsymbol{P}_s = \begin{bmatrix} p^F_a(0) & p^F_a(1) & p^F_a(2) & \cdots & p^F_a(N-1) & \sum_{n=N}^{\infty} p^F_a(n) \\ p^F_a(0) & p^F_a(1) & p^F_a(2) & \cdots & p^F_a(N-1) & \sum_{n=N}^{\infty} p^F_a(n) \\ 0 & p^F_a(0) & p^F_a(1) & \cdots & p^F_a(N-2) & \sum_{n=N-1}^{\infty} p^F_a(n) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p^F_a(0) & \sum_{n=1}^{\infty} p^F_a(n) \end{bmatrix}. \quad (1)$$

If the server is absent from Q^{F} at the end of a slot, that is, at time instants $\{t_k\}_{k\geq 0}$ transitions take place according to the probability matrix $\boldsymbol{P}_v$,

$$\boldsymbol{P}_v = \begin{bmatrix} p_a^{\mathrm{F}}(0) & p_a^{\mathrm{F}}(1) & p_a^{\mathrm{F}}(2) & \cdots & p_a^{\mathrm{F}}(N-1) & \sum_{n=N}^{\infty} p_a^{\mathrm{F}}(n) \\ 0 & p_a^{\mathrm{F}}(0) & p_a^{\mathrm{F}}(1) & \cdots & p_a^{\mathrm{F}}(N-2) & \sum_{n=N-1}^{\infty} p_a^{\mathrm{F}}(n) \\ 0 & 0 & p_a^{\mathrm{F}}(0) & \cdots & p_a^{\mathrm{F}}(N-3) & \sum_{n=N-2}^{\infty} p_a^{\mathrm{F}}(n) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{2}$$

Therefore, the l-step transition probability $\Pr\{q_{j+l}^{\mathrm{F}} = j \mid q_j^{\mathrm{F}} = i\}$ of the process over an interval $(j, j+l]$ of uninterrupted service to Q^{F} of length l (slots) is given by $\boldsymbol{P}_q^{\mathrm{F}}(l) = [\boldsymbol{P}_s]^l$, $l \geq 1$. Given $(q_{t_{k-1}}^{\mathrm{I}}, q_{t_{k-1}}^{\mathrm{F}})$, $q_{t_k^{\mathrm{d}}}^{\mathrm{F}}$ is determined by $q_{t_{k-1}}^{\mathrm{F}}$ and the transition matrix $\boldsymbol{P}_q^{\mathrm{F}}(h_k)$, where $h_k = 0$, if $q_{t_{k-1}}^{\mathrm{I}} > 0$ and distributed as random variable h otherwise; since $t_k = t_k^{\mathrm{d}} + \tau(q_{t_k^{\mathrm{d}}}^{\mathrm{F}}) + 1$, $q_{t_k}^{\mathrm{F}}$ is determined by $q_{t_k^{\mathrm{d}}}^{\mathrm{F}}$ and the transition matrix $\boldsymbol{P}_q^{\mathrm{F}}(\tau(q_{t_k^{\mathrm{d}}}^{\mathrm{F}}))\boldsymbol{P}_v$. Since $q_{t_k}^{\mathrm{I}} = q_{t_{k-1}}^{\mathrm{I}} + a^{\mathrm{I},d_k} - 1$ and $d_k = h_k + \tau(q_{t_k^{\mathrm{d}}}^{\mathrm{F}}) + 1$, $q_{t_k}^{\mathrm{I}}$ is determined by $(q_{t_{k-1}}^{\mathrm{I}}, q_{t_{k-1}}^{\mathrm{F}})$, in view of the above discussion. □

By identifying $q_{t_k}^{\mathrm{I}}$ and $q_{t_k}^{\mathrm{F}}$ as the level and phase of process $\{q_{t_k}^{\mathrm{I}}, q_{t_k}^{\mathrm{F}}\}_{k\geq 0}$, respectively, the transition probability matrix $\boldsymbol{P}_{(q^{\mathrm{I}},q^{\mathrm{F}})}$ of this embedded Markov chain can be written as a stochastic matrix of M/G/1 type:

$$\boldsymbol{P}_{(q^{\mathrm{I}},q^{\mathrm{F}})} = \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \boldsymbol{B}_3 & \boldsymbol{B}_4 & \cdots \\ \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{A}_3 & \boldsymbol{A}_4 & \cdots \\ \boldsymbol{0} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{A}_3 & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3}$$

where $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$, $n \geq 0$, are the $(N+1)$-dimensional matrices with elements $A_n(j_1, j_2)$ and $B_n(j_1, j_2)$, j_1, $j_2 \in R^{\mathrm{F}}$, given by

$$A_n(j_1, j_2) = \Pr\left\{q_{t_k}^{\mathrm{I}} = i_1 - 1 + n, q_{t_k}^{\mathrm{F}} = j_2 \mid q_{t_{k-1}}^{\mathrm{I}} = i_1, q_{t_{k-1}}^{\mathrm{F}} = j_1\right\}, \quad i_1 \geq 1, \tag{4}$$

$$B_n(j_1, j_2) = \Pr\left\{q_{t_k}^{\mathrm{I}} = n, q_{t_k}^{\mathrm{F}} = j_2 \mid q_{t_{k-1}}^{\mathrm{I}} = 0, q_{t_{k-1}}^{\mathrm{F}} = j_1\right\}. \tag{5}$$

Note that $B_n(j_1, j_2)$ is the probability that the process $\{q_{t_k}^{\mathrm{I}}, q_{t_k}^{\mathrm{F}}\}_{k\geq 0}$ makes a transition from state $(0, j_1)$ to state (n, j_2) over the time interval between two successive packet service completions at Q^{I}. $A_n(j_1, j_2)$ is the probability that over the time interval between two successive service completions at Q^{I}, the phase process $\{q_{t_k}^{\mathrm{F}}\}_{k\geq 0}$ makes a transition from state j_1 to state j_2, and there are n packet arrivals to Q^{I}, given that

Q^{I} is non-empty at the beginning of the transition. Let $A_n(j_1, j_2, l)$ ($B_n(j_1, j_2, l)$) be defined similarly to $A_n(j_1, j_2)$ ($B_n(j_1, j_2)$) but under the additional requirement that the duration of a transition d_k is equal to l slots, $l \geq 1$:

$$A_n(j_1, j_2, l) = \Pr\left\{q^{\mathrm{I}}_{t_k} = i_1 - 1 + n, q^{\mathrm{F}}_{t_k} = j_2, d_k = l \,|\, q^{\mathrm{I}}_{t_{k-1}} = i_1, q^{\mathrm{F}}_{t_{k-1}} = j_1\right\}, \quad i_1 \geq 1, \tag{6}$$

$$B_n(j_1, j_2, l) = \Pr\left\{q^{\mathrm{I}}_{t_k} = n, q^{\mathrm{F}}_{t_k} = j_2, d_k = l \,|\, q^{\mathrm{I}}_{t_{k-1}} = 0, q^{\mathrm{F}}_{t_{k-1}} = j_1\right\}. \tag{7}$$

Let $p^{\mathrm{F}}_{\tau}(i, j, n) = \Pr\{q^{\mathrm{F}}_{t^{\mathrm{d}}+n} = j, \tau(q^{\mathrm{F}}_{t^{\mathrm{d}}}) = n \,|\, q^{\mathrm{F}}_{t^{\mathrm{d}}} = i\}$, $i, j \in R^{\mathrm{F}}$, $0 \leq n \leq \mathcal{P}_i$, denote the transition probabilities of $\{q^{\mathrm{F}}_j\}_{j\geq 0}$ over a service horizon of length n slots. These are determined by the adopted service policy $\bar{\mathcal{P}}$ and the associated state-policies $\mathcal{P}_i$, $i \in R^{\mathrm{F}}$, and can be expressed in terms of the one-step transition probabilities of $\{q^{\mathrm{F}}_j\}_{j\geq 0}$ given by probability matrix $\boldsymbol{P}_s$ [14]. The stochastic matrix with elements $p^{\mathrm{F}}_{\tau}(i, j, n)$ is denoted by $\boldsymbol{P}^{\mathrm{F}}_{\tau}(n)$. It can be easily shown that

$$A_n(j_1, j_2, l) = p^{\mathrm{I},l}_a(n) \sum_{j=0}^{N} p^{\mathrm{F}}_{\tau}(j_1, j, l-1) p_v(j, j_2), \quad l \geq 1, \tag{8}$$

$$B_n(j_1, j_2, l) = \sum_{m=1}^{l-1} [p^{\mathrm{I}}_a(0)]^{m-1} \sum_{j=0}^{N} p^{\mathrm{F}}_q(j_1, j, m) \sum_{i=1}^{n+1} p^{\mathrm{I}}_a(i) p^{\mathrm{I},l-m}_a(n-i+1) \times \sum_{j'=0}^{N} p^{\mathrm{F}}_{\tau}(j, j', l-m-1) p_v(j', j_2), \quad l \geq 2, \tag{9}$$

$$B_n(j_1, j_2, 1) = 0.$$

From (8) and (9) matrices $\boldsymbol{A}_n(l)$ and $\boldsymbol{B}_n(l)$, $l \geq 1$, may be written as

$$\boldsymbol{A}_n(l) = p^{\mathrm{I},l}_a(n) \boldsymbol{P}^{\mathrm{F}}_{\tau}(l-1) \boldsymbol{P}_v, \tag{10}$$

$$\boldsymbol{B}_n(l) = \sum_{m=1}^{l-1} [p^{\mathrm{I}}_a(0)]^{m-1} \boldsymbol{P}^{\mathrm{F}}_q(m) \sum_{i=1}^{n+1} p^{\mathrm{I}}_a(i) \boldsymbol{A}_{n-i+1}(l-m), \qquad \boldsymbol{B}_n(1) = \boldsymbol{0}. \tag{11}$$

Application of the law of total probability yields

$$\boldsymbol{A}_n = \sum_{l=1}^{\infty} \boldsymbol{A}_n(l) = \sum_{l=1}^{\infty} p^{\mathrm{I},l}_a(n) \boldsymbol{P}^{\mathrm{F}}_{\tau}(l-1) \boldsymbol{P}_v, \tag{12}$$

$$\boldsymbol{B}_n = \sum_{l=1}^{\infty} \boldsymbol{B}_n(l) = [\boldsymbol{I} - p^{\mathrm{I}}_a(0) \boldsymbol{P}_s]^{-1} \boldsymbol{P}_s \sum_{i=1}^{n+1} p^{\mathrm{I}}_a(i) \boldsymbol{A}_{n-i+1}. \tag{13}$$

The stationary joint probabilities for the queue occupancies upon packet service completions at Q^{I}, $y(i, j)$, $i \in R^{\mathrm{I}}$, $j \in R^{\mathrm{F}}$, as determined by the transition probability matrix $\boldsymbol{P}_{(q^{\mathrm{I}}, q^{\mathrm{F}})}$, are derived by applying matrix analytic techniques [14]. The stationary joint probabilities at arbitrary time instants, $\psi(i, j)$, $i \in R^{\mathrm{I}}$, $j \in R^{\mathrm{F}}$ are then obtained in terms of $y(i, j)$ by invoking Markov renewal theory arguments [14].

4. Partial performance ordering of policies in $\mathcal{S}$

In view of the structure of class $\mathcal{S}$ it is possible to establish a partial performance ordering of policies in $\mathcal{S}$, as shown below.

Definition 4. Consider service policies $\bar{\Sigma} = \{\Sigma_1, \ldots, \Sigma_N\}$ and $\bar{\Omega} = \{\Omega, \ldots, \Omega_N\}$ in $\mathcal{S}$. Let $q_k^{M,\bar{\mathcal{P}}}$ denote the occupancy of Q^M at time slot k under policy $\bar{\mathcal{P}}$, $M \in \{\mathrm{I}, \mathrm{F}\}$, $\bar{\mathcal{P}} \in \{\bar{\Omega}, \bar{\Sigma}\}$. Policy $\bar{\Omega}$ is said to outperform policy $\bar{\Sigma}$ with respect to Q^M if and only if $q_k^{M,\bar{\Omega}} \leq q_k^{M,\bar{\Sigma}}$, $M \in \{\mathrm{I}, \mathrm{F}\}$, for all $k \geq 0$. That is,

$$\bar{\Sigma} \overset{M}{\prec} \bar{\Omega} \iff q_k^{M,\bar{\Omega}} \leq q_k^{M,\bar{\Sigma}}, \quad M \in \{\mathrm{I}, \mathrm{F}\}. \tag{14}$$

Notice that (14) implies that if $\bar{\Sigma} \overset{\mathrm{F}}{\prec} \bar{\Omega}$, then

$$L^{\mathrm{F},\bar{\Omega}} \leq L^{\mathrm{F},\bar{\mathcal{P}}} \quad \text{and} \quad D^{\mathrm{I},\bar{\Omega}} \geq D^{\mathrm{I},\bar{\mathcal{P}}}, \tag{15}$$

where $L^{\mathrm{F},\bar{\mathcal{P}}}$ denotes the packet loss probability at Q^{F} under policy $\bar{\mathcal{P}}$ and $D^{\mathrm{I},\bar{\mathcal{P}}}$ denotes the average packet delay at Q^{I} under policy $\bar{\mathcal{P}}$, $\bar{\mathcal{P}} \in \{\bar{\Sigma}, \bar{\Omega}\}$. The following proposition provides for a partial performance ordering of policies in $\mathcal{S}$; its proof may be found in Appendix A.

Proposition 2. *Consider two policies $\bar{\Sigma} = \{\Sigma_1, \ldots, \Sigma_N\}$ and $\bar{\Omega} = \{\Omega_1, \ldots, \Omega_N\}$ in $\mathcal{S}$. Then*

$$\Sigma_i \leq \min_{1 \leq k \leq N} \{\Omega_k\} \qquad \textit{for all } i,\ 1 \leq i \leq N, \tag{16}$$

implies that

$$\bar{\Sigma} \overset{\mathrm{F}}{\prec} \bar{\Omega} \quad \textit{and} \quad \bar{\Omega} \overset{\mathrm{I}}{\prec} \bar{\Sigma}. \tag{17}$$

Corollary 1. *Let $\bar{\mathcal{L}}_k = \{k, k, \ldots, k\}$ denote the k-limited service policy in $\mathcal{S}$, $0 \leq k \leq \infty$. Then Proposition 2 implies that*

$$\textit{if } k \leq j \quad \textit{then } \bar{\mathcal{L}}_k \overset{\mathrm{F}}{\prec} \bar{\mathcal{L}}_j \textit{ and } \bar{\mathcal{L}}_j \overset{\mathrm{I}}{\prec} \bar{\mathcal{L}}_k. \tag{18}$$

Under some conditions on the arrival process to Q^{F} a weaker than (16) sufficient condition for performance ordering of policies in $\mathcal{S}$ may be obtained as described in Proposition 3.

Proposition 3. *Consider two policies $\bar{\Sigma} = \{\Sigma_1, \ldots, \Sigma_N\}$ and $\bar{\Omega} = \{\Omega_1, \ldots, \Omega_N\}$ in $\mathcal{S}$. Let $a^{\mathrm{F}}(\Delta t)$ denote the number of packet arrivals to Q^{F} over an interval of length Δt (slots). If*

$$a^{\mathrm{F}}(\Delta t) \leq \Delta t \tag{19}$$

for all Δt, $\Delta t \in \{1, 2, \ldots\}$, and

$$\Sigma_i \leq \Omega_k \qquad \textit{for all } i \geq k \textit{ and } k,\ 1 \leq k \leq N, \tag{20}$$

then

$$\bar{\Sigma} \overset{\mathrm{F}}{\prec} \bar{\Omega} \quad \textit{and} \quad \bar{\Omega} \overset{\mathrm{I}}{\prec} \bar{\Sigma}. \tag{21}$$

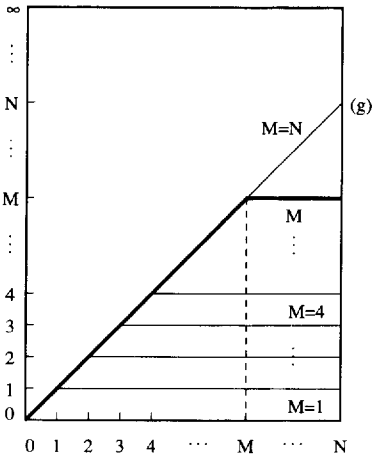


Fig. 4. Policy-line representation of the set of policies in Example 1.

The proof of Proposition 3 may be found in Appendix A.

When the arrrival process to Q^F is Bernoulli, condition (19) in Proposition 3 holds and thus, the following corollary is evident in view of Proposition 3.

Corollary 2. *Consider two policies* $\bar{\Sigma} = \{\Sigma_1, \ldots, \Sigma_N\}$ *and* $\bar{\Omega} = \{\Omega_1, \ldots, \Omega_N\}$ *in* $\mathcal{S}$. *Let the packet arrival process to* Q^F *be a Bernoulli process. If*

$$\Sigma_i \leq \Omega_k \qquad \text{for all } i \geq k,\ 1 \leq k \leq N, \tag{22}$$

then

$$\bar{\Sigma} \overset{F}{\prec} \bar{\Omega} \quad \text{and} \quad \bar{\Omega} \overset{I}{\prec} \bar{\Sigma}. \tag{23}$$

5. Examples, numerical results and discussion

The HoL priority policy for Q^F (upper boundary of the state-policy space $\mathcal{SP}$) minimizes packet delay/loss for this queue. Similarly, the HoL priority policy for Q^I (lower boundary of the state-policy space $\mathcal{SP}$) minimizes packet delay for this queue. A (state-dependent) service policy in $\mathcal{S}$ is expected to provide greater flexibility in meeting performance requirements associated with each queue. The vector representation of a service policy in $\mathcal{S}$ allows for the description of ad hoc, customized policies as illustrated in the following examples.

Example 1. The service policy $\bar{\mathcal{P}}$ defined by the following state-policies $\mathcal{P}_i$,

$$\mathcal{P}_i = \begin{cases} i, & 1 \leq i \leq M, \\ M, & M < i \leq N, \end{cases} \qquad 1 \leq M \leq N,$$

provides unlimited gated service to Q^F when the content (state) of Q^F is below some threshold M, $1 \leq M \leq N$, and M-limited service otherwise (Fig. 4). This policy provides increased service to Q^I compared

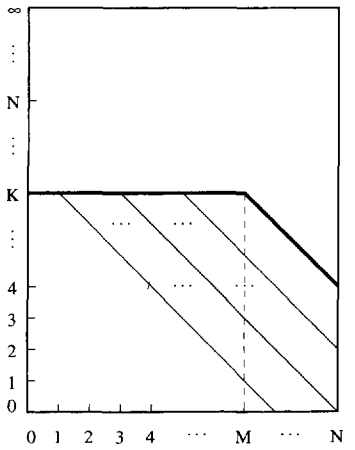


Fig. 5. Policy-line representation of the set of policies in Example 2.

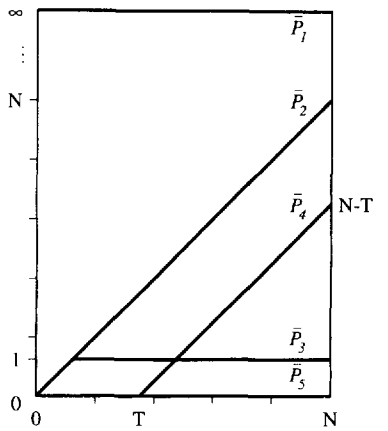


Fig. 6. Policy-line representation of the set of policies in Example 3.

to the unlimited gated policy when the content of Q^F exceeds the threshold M; it also provides increased service to Q^I compared to the M-limited policy for $1 \leq i < M$. The policy becomes 1-limited for $M = 1$ and unlimited gated for $M = N$.

Example 2. Consider the policies L_k^M which are k-limited up to a certain state M, $1 \leq M \leq N$, and then drop linearly with fixed slope, as shown in Fig. 5. If packet arrivals to Q^F satisfy the condition in Proposition 3, for instance, Bernoulli arrivals, then the following performance ordering holds true: $L_k^{M_1} \overset{F}{\prec} L_k^{M_2}$ and $L_k^{M_2} \overset{I}{\prec} L_k^{M_1}$ if $M_1 < M_2$.

Example 3. Consider the following policies (Fig. 6):

$\bar{\mathcal{P}}_1$: $\mathcal{P}_i = \infty$, $1 \leq i \leq N$ (unlimited non-gated or ∞-limited or HoL for Q^F);
$\bar{\mathcal{P}}_2$: $\mathcal{P}_i = i$, $1 \leq i \leq N$ (unlimited gated);
$\bar{\mathcal{P}}_3$: $\mathcal{P}_i = 1$, $1 \leq i \leq N$ (1-limited);

$\bar{\mathcal{P}}_4$: $\mathcal{P}_i = 0, 1 \le i \le T$ and $\mathcal{P}_i = i - T, T < i \le N$; according this service policy Q^{I} receives HoL priority service whenever the content of Q^{F} is less than a threshold T. If the content i of Q^{F} at a decision time-instant exceeds the value of the 'safety' threshold T, service is provided to $i - T$ packets at Q^{F}. Compared to the HoL priority service policy for Q^{I}, this 'mixed' policy provides increased service to the finite queue when it is critically congested ($i > T$) and is expected to decrease the loss probability.

$\bar{\mathcal{P}}_5$: $\mathcal{P}_i = 0, 1 \le i \le N$ (0-limited or HoL for Q^{I}).

In view of Proposition 1, the following performance ordering of the policies may be easily established:

$$\bar{\mathcal{P}}_5 \overset{\mathrm{F}}{\prec} \bar{\mathcal{P}}_3 \overset{\mathrm{F}}{\prec} \bar{\mathcal{P}}_2 \overset{\mathrm{F}}{\prec} \bar{\mathcal{P}}_1 \quad \text{and} \quad \bar{\mathcal{P}}_5 \overset{\mathrm{F}}{\prec} \bar{\mathcal{P}}_4 \overset{\mathrm{F}}{\prec} \bar{\mathcal{P}}_1,$$
$$\bar{\mathcal{P}}_1 \overset{\mathrm{I}}{\prec} \bar{\mathcal{P}}_2 \overset{\mathrm{I}}{\prec} \bar{\mathcal{P}}_3 \overset{\mathrm{I}}{\prec} \bar{\mathcal{P}}_5 \quad \text{and} \quad \bar{\mathcal{P}}_1 \overset{\mathrm{I}}{\prec} \bar{\mathcal{P}}_4 \overset{\mathrm{I}}{\prec} \bar{\mathcal{P}}_5.$$

Notice that $\bar{\mathcal{P}}_1 \overset{\mathrm{I}}{\prec} \bar{\mathcal{P}} \overset{\mathrm{I}}{\prec} \bar{\mathcal{P}}_5$ and $\bar{\mathcal{P}}_5 \overset{\mathrm{F}}{\prec} \bar{\mathcal{P}} \overset{\mathrm{F}}{\prec} \bar{\mathcal{P}}_1$ for all service policies $\bar{\mathcal{P}} \in \mathcal{S}$, which implies that $\bar{\mathcal{P}}_1$ and $\bar{\mathcal{P}}_5$ determine bounds on the achievable performance by any service policy in $\mathcal{S}$.

The performance of some service policies in $\mathcal{S}$ is evaluated in terms of the induced packet loss probability at the finite queue, L^{F}, and the average packet delays D^{I} and D^{F}. These can be easily determined from the joint probability distribution of the queue occupancies. Let λ^{F} (λ^{I}) denote the packet arrival rate to Q^{F} (Q^{I}). Then

$$L^{\mathrm{F}} = \frac{\lambda^{\mathrm{F}} + \lambda^{\mathrm{I}} - \rho'}{\lambda^{\mathrm{F}}},$$

where $\rho' = 1 - \psi(0, 0) = 1 - \Pr\{q^{\mathrm{I}} = 0, q^{\mathrm{F}} = 0\}$ is the system utilization. The average packet delays D^{I} and D^{F} are derived from the mean queue occupancies $E[q^{\mathrm{I}}]$ and $E[q^{\mathrm{F}}]$ by applying Little's theorem:

$$D^{\mathrm{I}} = \frac{E[q^{\mathrm{I}}]}{\lambda^{\mathrm{I}}}, \qquad D^{\mathrm{F}} = \frac{E[q^{\mathrm{F}}]}{\lambda^{\mathrm{F}}(1 - L^{\mathrm{F}})}.$$

Some numerical results for L^{F} and D^{I} induced by service policies $\bar{\mathcal{P}}_1 - \bar{\mathcal{P}}_5$ in Example 3 have been derived and are shown in Figs. 7 and 8 under Bernoulli and in Figs. 9 and 10 under truncated Poisson packet arrival processes to each of the queues. The finite queue capacity is $N = 29$. For $\bar{\mathcal{P}}_4$, two different threshold values are considered, $T = 10$ and $T = 15$. For each type of arrival processes the mean packet delay D^{I} and the packet loss probability L^{F} are plotted versus: (i) the arrival rate λ^{I} for a fixed arrival rate $\lambda^{\mathrm{F}} = 0.5$ (packets/slot) and a range of offered traffic load between 0.8 and 0.925 (Figs. 7 and 9) and (ii) the arrival rate λ^{F} for a fixed arrival rate $\lambda^{\mathrm{I}} = 0.5$ (packets/slot) and the same total load range as in (i) (Figs. 8 and 10).

The results shown in Figs. 7–10 are in accordance with the performance ordering established by Proposition 2. Furthermore, they illustrate the potential for improved performance provided by the ad hoc service policies in $\mathcal{S}$, such as $\bar{\mathcal{P}}_4$. For instance, if upper bounds on L^{F} equal to 10^{-12} (under Bernoulli arrivals) or 10^{-9} (under truncated Poisson arrivals) are necessary to meet quality of service requirements, then only service policies $\bar{\mathcal{P}}_1$, $\bar{\mathcal{P}}_2$ and $\bar{\mathcal{P}}_4$ ($T = 10$) are acceptable for all values of λ^{I} and λ^{F} considered in Figs. 7–10. Among these policies, $\bar{\mathcal{P}}_4$ ($T = 10$) induces significantly lower D^{I} compared to that under $\bar{\mathcal{P}}_1$ and $\bar{\mathcal{P}}_2$ and, thus, it is the most effective in the sense that it minimizes D^{I} while satisfying the constraint on L^{F}.

Results for finite capacity $N = 50$ and Poisson packet arrival processes are shown in Figs. 11 and 12. For policy $\bar{\mathcal{P}}_4$, the thresholds $T = 15$ and $T = 25$ have been considered. Results for $N = 50$ and Bernoulli

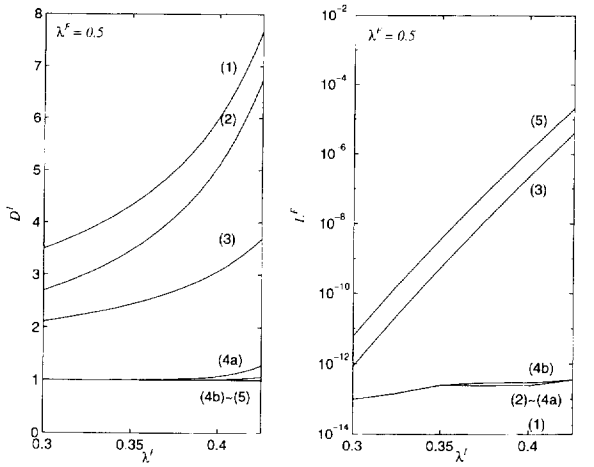


Fig. 7. Mean packet delay D^{I} and loss probability L^{F} under Bernoulli packet arrivals ($N = 29$, $\lambda^{\mathrm{F}} = 0.5$). Curve (i) corresponds to policy $\bar{\mathcal{P}}_i$; (4a) and (4b) correspond to $\bar{\mathcal{P}}_4$ with $T = 10$ and $T = 15$, respectively. $(i) \sim (j)$ indicates that the results under policies $\bar{\mathcal{P}}_i$ and $\bar{\mathcal{P}}_j$ are approximately the same.

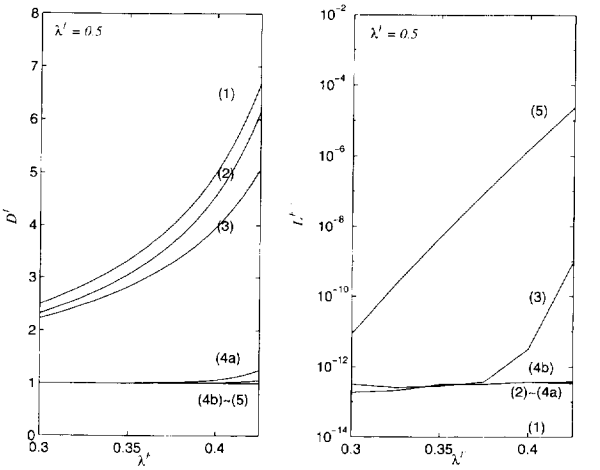


Fig. 8. Mean packet delay D^{I} and loss probability L^{F} under Bernoulli packet arrivals ($N = 29$, $\lambda^{\mathrm{I}} = 0.5$). Curve (i) corresponds to policy $\bar{\mathcal{P}}_i$; (4a) and (4b) correspond to $\bar{\mathcal{P}}_4$ with $T = 10$ and $T = 15$, respectively. $(i) \sim (j)$ indicates that the results under policies $\bar{\mathcal{P}}_i$ and $\bar{\mathcal{P}}_j$ are approximately the same.

arrival processes have not been presented since most of policies $\bar{\mathcal{P}}_1$–$\bar{\mathcal{P}}_5$—with the exception of $\bar{\mathcal{P}}_3$ and $\bar{\mathcal{P}}_4$ over certain ranges of traffic loads—induce very low loss probability at Q^{F}. In this case, the accuracy of the derived loss probabilities depends on the numerical precision of the machine.

The queueing system introduced and studied in this work could model the queueing behavior of a protocol which allocates a common resource to two traffic classes with different Quality-of-Service (QoS) requirements. One class may be viewed as a best effort class [15] and the second as having more stringent QoS requirements. A sufficiently large buffer (Q^{I}) can be assumed to be available to the best effort traffic—called also Available Bit Rate (ABR)—for which a mean value of the induced queueing intensity (such

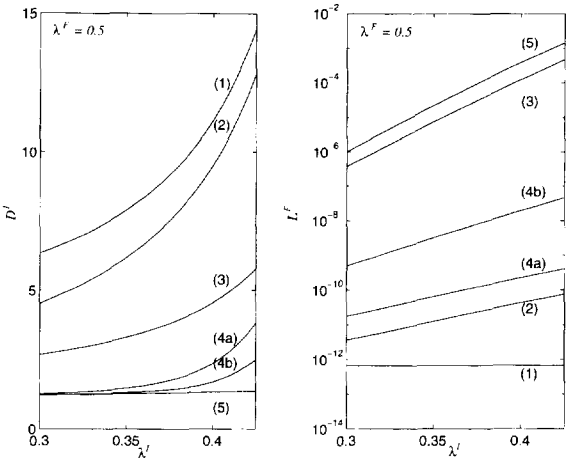


Fig. 9. Mean packet delay D^{I} and loss probability L^{F} under Poisson packet arrivals ($N = 29$, $\lambda^{\mathrm{F}} = 0.5$). Curve ($i$) corresponds to policy $\tilde{\mathcal{P}}_i$; (4a) and (4b) correspond to $\tilde{\mathcal{P}}_4$ with $T = 10$ and $T = 15$, respectively.

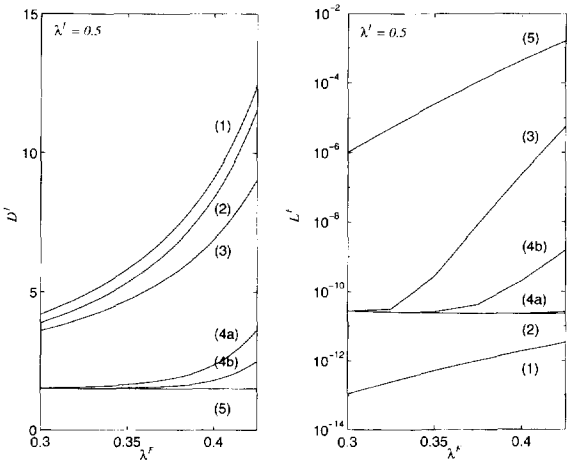


Fig. 10. Mean packet delay D^{I} and loss probability L^{F} under Poisson packet arrivals ($N = 29$, $\lambda^{\mathrm{I}} = 0.5$). Curve ($i$) corresponds to policy $\tilde{\mathcal{P}}_i$; (4a) and (4b) correspond to $\tilde{\mathcal{P}}_4$ with $T = 10$ and $T = 15$, respectively.

as delay) would be meaningful. For the traffic with more stringent QoS requirements a finite buffer (Q^{F}) could be considered to capture, for instance, a requirement associated with a maximum tolerable delay. For example, if a packet delay exceeding N (slots) is unacceptable, then it is evident that a buffer capacity larger than N is meaningless. Furthermore, the cell loss probability would serve as a lower bound on the probability that a packet is delayed beyond N slots. Under the HoL priority policy for Q^{F}, the packet loss probability and the probability that a packet is delayed by more than N slots will coincide, assuming that such packets do not receive service.

Finally, the partial ordering of the proposed policies can provide for an efficient search for a policy which delivers a desired performance. At first, the Head-of-Line (HoL) priority for one class would establish if there exists *any* policy in the class which would induce the target-performance (or desired Quality-of-

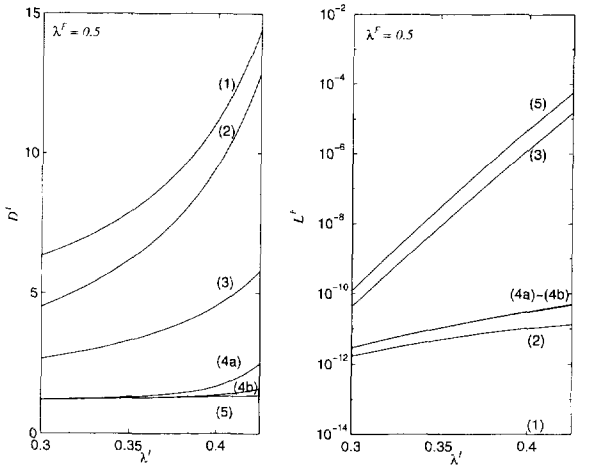


Fig. 11. Mean packet delay D^I and loss probability L^F under Poisson packet arrivals ($N = 50$, $\lambda^F = 0.5$). Curve (i) corresponds to policy $\bar{\mathcal{P}}_i$; (4a) and (4b) correspond to $\bar{\mathcal{P}}_4$ with $T = 15$ and $T = 25$, respectively.

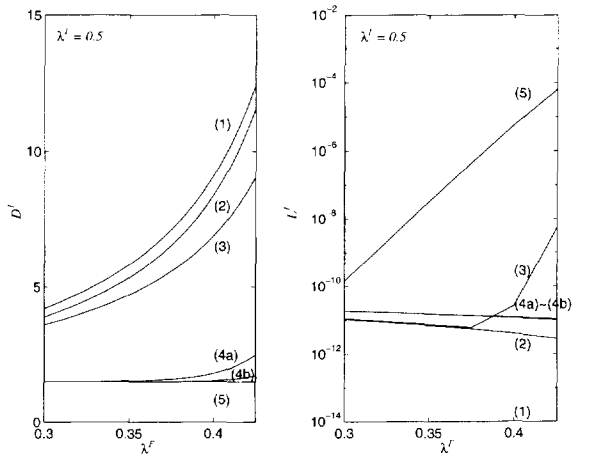


Fig. 12. Mean packet delay D^I and loss probability L^F under Poisson packet arrivals ($N = 50$, $\lambda^I = 0.5$). Curve (i) corresponds to policy $\bar{\mathcal{P}}_i$; (4a) and (4b) correspond to $\bar{\mathcal{P}}_4$ with $T = 15$ and $T = 25$, respectively.

Service) for this class. If it exists, then Proposition 2 could provide guidance in the search for an acceptable policy. It may not identify the best policy or even find an existing acceptable policy, but it is one—possibly effective—alternative to the exhaustive search. By considering different starting policies, the entire space could be searched in a non-exhaustive manner by moving 'toward' the better policies as indicated by Proposition 2.

6. Conclusions

The main contribution of this paper is the introduction and study of the versatile class of service policies $\mathcal{S}$ based on the concept of the state-policy. Through the construction of service policies in terms of policy-lines

on the space of state-policies $\mathcal{SP}$ (Fig. 2), the flexibility in designing policies in $\mathcal{S}$ is illustrated. Through the introduced simple performance ordering of service policies a direction for the design of service policies in $\mathcal{S}$ whose performance approach a desired level is presented. This indicates the policy design flexibility and potential for closely achieving a desired level of performance associated with the introduced class of service policies. The latter have been illustrated through some examples and numerical results.

Appendix A. Proof of Propositions 2 and 3

A.1. *Proof of Proposition 2*

It suffices to show that under any realization of the packet arrival process

$$q_k^{\mathrm{F},\bar{\Omega}} \leq q_k^{\mathrm{F},\bar{\Sigma}} \qquad (\text{or } q_k^{\mathrm{I},\bar{\Sigma}} \leq q_k^{\mathrm{I},\bar{\Omega}})$$

for all $k \geq 0$. Consider an initially empty system. Let t_1 mark the (end of the time) slot at which the first packet arrives to Q^{I}; clearly,

$$q_{t_1}^{\mathrm{F},\bar{\Omega}} = q_{t_1}^{\mathrm{F},\bar{\Sigma}},$$

since uninterrupted service to Q^{F} has been provided under both policies up to t_1. Condition (16) implies that the service horizons under policies $\bar{\Sigma}$ and $\bar{\Omega}$—denoted by $\tau^{\bar{\Sigma}}(q_{t_1}^{\mathrm{F},\bar{\Sigma}})$ and $\tau^{\bar{\Omega}}(q_{t_1}^{\mathrm{F},\bar{\Omega}})$, respectively—will satisfy

$$\tau^{\bar{\Sigma}}(q_{t_1}^{\mathrm{F},\bar{\Sigma}}) \leq \tau^{\bar{\Omega}}(q_{t_1}^{\mathrm{F},\bar{\Omega}}),$$

and thus,

$$q_t^{\mathrm{I},\bar{\Sigma}} \leq q_t^{\mathrm{I},\bar{\Omega}}, \quad t_1 \leq t \leq t_2^{\bar{\Omega}},$$

since a packet from Q^{I} is served under policy $\bar{\Sigma}$ no latter than under policy $\bar{\Omega}$; $t_2^{\bar{\Omega}} = t_1 + \tau^{\bar{\Omega}}(q_{t_1}^{\mathrm{F},\bar{\Omega}}) + 1$ marks the slot at which the server serves the first packet in Q^{I} under policy $\bar{\Omega}$. At $t_2^{\bar{\Omega}}$ the following cases may be distinguished.

(a) If $q_{t_2^{\bar{\Omega}}}^{\mathrm{I},\bar{\Omega}} = 0$, then it is easy to see that $q_{t_2^{\bar{\Omega}}}^{\mathrm{I},\bar{\Sigma}} = 0$, since the former implies that only one packet arrival to Q^{I} has occurred up to $t_2^{\bar{\Omega}}$. Let t_1' mark the arrival time of the next packet to Q^{I}. Then,

$$q_t^{\mathrm{I},\bar{\Sigma}} = q_t^{\mathrm{I},\bar{\Omega}} = 0, \quad t_2^{\bar{\Omega}} \leq t < t_1'.$$

Thus, it has been proven that

$$q_t^{\mathrm{I},\bar{\Sigma}} \leq q_t^{\mathrm{I},\bar{\Omega}}, \quad t_1 \leq t < t_1',$$

where $[t_1, t_1')$ denotes the time interval of an I-cycle, defined to be the interval between consecutive arrivals to Q^{I} which find the buffer empty under both policies $\bar{\Sigma}$ and $\bar{\Omega}$.

(b) If $q_{t_2^{\bar{\Omega}}}^{\mathrm{I},\bar{\Omega}} > 0$, then it is easy to see that the time t_0 at which the second packet is forwarded to the head of Q^{I} under policy $\bar{\Sigma}$ will satisfy

$$t_1 + \tau^{\bar{\Sigma}}(q_{t_1}^{\mathrm{F},\bar{\Sigma}}) + 1 \triangleq t_2^{\bar{\Sigma}} \leq t_0 \leq t_2^{\bar{\Omega}}.$$

The departure time instant of the second packet under policies $\bar{\Sigma}$ and $\bar{\Omega}$, will be

$$t_3^{\bar{\Sigma}} \triangleq t_0 + \tau^{\bar{\Sigma}}(q_{t_0}^{\mathrm{F},\bar{\Sigma}}) + 1 \quad \text{and} \quad t_3^{\bar{\Omega}} \triangleq t_2^{\bar{\Omega}} + \tau^{\bar{\Omega}}(q_{t_2^{\bar{\Omega}}}^{\mathrm{F},\bar{\Omega}}) + 1,$$

respectively. Since in view of (16),

$$\tau^{\bar{\Sigma}}(q_{t_0}^{\mathrm{F},\bar{\Sigma}}) \leq \tau^{\bar{\Omega}}(q_{t_2^{\bar{\Omega}}}^{\mathrm{F},\bar{\Omega}}), \tag{A.1}$$

it is implied that $t_3^{\bar{\Sigma}} \leq t_3^{\bar{\Omega}}$ and

$$q_t^{\mathrm{I},\bar{\Sigma}} \leq q_t^{\mathrm{I},\bar{\Omega}}, \quad t_0 \leq t \leq t_3^{\bar{\Omega}}. \tag{A.2}$$

By reiterating the argument presented above until Q^{I} becomes empty under both policies, it is established that (A.2) holds at any time instant of an I-cycle with an arbitrary number of packet arrivals to Q^{I}.

Parts (a) and (b) above complete the proof of Proposition 2.

A.2. Proof of Proposition 3

The proof can be established by following the proof of Proposition 2. The introductory part and part (a) of that proof is directly applicable. In order for part (b) to be valid, the validity of (A.1) needs to be established. While condition (16) in Proposition 2 guarantees that for any values of $q_{t_0}^{\mathrm{F},\bar{\Sigma}}$ and $q_{t_2^{\bar{\Omega}}}^{\mathrm{F},\bar{\Omega}}$ inequality (A.1) will hold true, condition (20) in Proposition 3 would also guarantee it provided that

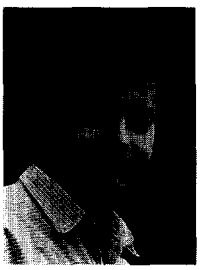
$$q_{t_0}^{\mathrm{F},\bar{\Sigma}} \geq q_{t_2^{\bar{\Omega}}}^{\mathrm{F},\bar{\Omega}}. \tag{A.3}$$

That is, provided that the finite queue occupancy when the second packet is forwarded to the head of Q^{I} under policy $\bar{\Sigma}$ be not less than under policy $\bar{\Omega}$. Since Q^{F} will have been reduced by $t_2^{\bar{\Omega}} - t_0$ at $t_2^{\bar{\Omega}}$ under policy $\bar{\Omega}$ compared to its value at t_0 under policy $\bar{\Sigma}$ and the potential new packet arrivals to Q^{F} over the interval $(t_0, t_2^{\bar{\Omega}})$ will be at most $t_2^{\bar{\Omega}} - t_0$ (condition (19)), it is evident that (A.3) will hold true. Reiterating the argument for the packet arrival to the head of Q^{I} that may follow, the proof of Proposition 3 can be established.

References

[1] H. Takagi, Queueing analysis of polling models, *ACM Comput. Surveys* **20**(1) (1988) 5–28.
[2] H. Levy and M. Sidi, Polling systems: Applications, modelling, and optimization, *IEEE Trans. Comm.* **38**(10) (1990) 1750–1760.

[3] H. Takagi, *Queueing analysis: A foundation of performance evaluation. Vacation and priority systems,* Vol. 1, Part 1, **42**(2) (1994) 615–624.
[4] I. Stavrakakis, Delay bounds on a queueing system with consistent priorities, *IEEE Trans. Comm.*, **42**(2) (1994) 615–624. *Proc. IEEE INFOCOM'92*, Florence, Italy, May 4–8 (1992) pp. 2151–2160.
[5] C. Bisdikian, A queueing model for a data station within the IEEE 802.6 MAN, *IEEE 17th Local Computer Networks Conf.*, Minneapolis, MN, September 13–16, 1992.
[6] R. Landry and I. Stavrakakis, A three-priority queueing policy with applications to communication networks, *Proc. INFOCOM'93*, San Francisco, CA, March 30–April 1, 1993, pp. 1067–1074.
[7] M. Eisenberg, Two queues with alternating service, *SIAM J. Appl. Math.* **36**(2) (1979) 287–303.
[8] T. Ozawa, Alternating service queues with mixed exhaustive and K-limited services, *Perform. Eval.* **11** (1990) 165–175.
[9] I. Stavrakakis and S. Tsakiridou, Occupancy distribution for a DQDB station based on a queueing system with Markov-structured service requirements, *Proc. INFOCOM'93*, San Francisco, CA, March 30–April 1, 1993, pp. 1083–1090.
[10] C.M. Harris and W.G. Marchal, State dependence in M/G/1 server-vacation models, *Oper. Res.* **36**(4) (1988) 560–565.
[11] J. Keilson and L.D. Servi, Blocking probability for M/G/1 vacation systems with occupancy level dependent schedules, *Oper. Res.* **37**(1) (1989) 134–140.
[12] D.M. Lucantoni, K.S. Meier-Hellstern and M.F. Neuts, A single-server queue with server vacations and a class of non-renewal arrival processes, *Adv. in Appl. Probab.* **22** (1990) 676–705.
[13] K.K. Leung, Cyclic-service systems with probabilistically-limited service, *IEEE J. Select. Areas Commun.* **9**(2) (1991) 185–193.
[14] S. Tsakiridou and I. Stavrakakis, Study of a class of partially ordered service strategies for a system of two discrete-time queues, Technical Report, TR-CDSP-96-37, Communications and Digital Signal Processing (CDSP) Center, ECE Department, Northeastern University, Boston, 1996.
[15] D. Towsley, Providing quality of service in packet switched networks, in: L. Donatiello and R. Nelson (Eds.) *Performance Evaluation of Computer and Communication Systems*, Lecture Notes in Computer Science, Vol. 729, Springer, Berlin (1993) pp. 560–586.
[16] M.F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and their Applications*, Marcel Dekker, New York (1989).



Ioannis Stavrakakis received the Diploma in Electrical Engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, 1983, and the Ph.D. degree in Electrical Engineering from the University of Virginia, 1988.

In 1988, he joined the faculty of Computer Science and Electrical Engineering of the University of Vermont as an assistant and then associate professor. Since 1994, he has been an associate professor of Electrical and Computer Engineering of the Northeastern University, Boston. His research interests are in stochastic system modeling, teletraffic analysis and discrete-time queueing theory, with primary focus on the design and performance evaluation of Broadband Integrated Services Digital Networks (B-ISDN).

Dr. Stavrakakis is a senior member of IEEE and a member of the IEEE Communications Society, Technical Committee on Computer Communications. He has organized and chaired sessions, and has been a technical committee member, for conferences such as GLOBECOM, ICC and INFOCOM.

Sophia Tsakiridou received the Diploma in Electrical Engineering from the Aristotelian University of Thessaloniki, Greece, in 1988 and the Ph.D. degree in Electrical Engineering from the University of Vermont in 1994.

She is currently a Postdoctoral Fellow at Queen's University, Canada. Her research interests are in the areas of stochastic modeling and performance evaluation of communication networks.