Queueing Systems 44, 253–280, 2003

Queueing and Fluid Analysis of Partial Message Discarding Policy

PARIJAT DUBE pdube@us.ibm.com
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

EITAN ALTMAN altman@sophia.inria.fr
Project MISTRAL, INRIA, 2004, Route des Lucioles, B.P. 93, 06902, Sophia Antipolis Cedex, France

Received 6 February 2001; Revised 20 January 2003

Abstract. We consider a stream of packets that arrive at a queue with a finite buffer. A group of consecutive packets constitutes a frame. We assume that when an arriving packet finds the queue full, not only is the packet lost but also the future packets that belong to the same frame will be rejected. The first part of the paper deals with a detailed packet level queueing model; we obtain exact expressions for the stationary queue length distribution and the *goodput ratio* (i.e. the fraction of arriving frames that experience no losses). The second part deals with a fluid model and the fluid analysis leads to simple closed form expressions for the stationary workload process and the fluid goodput ratio.

Keywords: PMD policy, packet model, queues with vacation, goodput, fluid queue

1. Introduction

Often, a set of consecutive packets are grouped into a frame, and loss of one packet results in the loss of the whole frame. This situation is motivated by telecommunication systems in which often frames of information are fragmented into smaller entities (cells or packets) and when loosing one or more packets of a frame, the whole frame is corrupted. This is the case if we send Internet packets (typically of the size of 1 kbytes) over a wireless link (where packets have the size of around 300 bytes) or over the so called ATM networks (in which cells have the size of 52 bytes) [10]. Thus, packets of a frame, that arrive after a packet is lost from the same frame, are useless and it is advantageous to discard them thus achieving the twofold objective of congestion avoidance and goodput maximization. In the context of telecommunications, this approach for discarding is known as the Partial Message Discard (PMD) policy. This policy as well as other discarding approaches have been studied in several previous papers [6,7,9,11,13].

Our model could also be useful in other applications in which an entity is composed into different objects that queue up for service, and where the loss of one object makes the whole original entity useless. One may think of remote computing where a computer program is split into tasks that queue for service at some remote computer and if a task is lost then the whole program has to be re-initiated. One could also think of production lines, where if some component in the production of some entity is lost then the whole entity is useless.

In [7,13], the basic performance measure for the study of discarding policies is the *effective throughput*, which is the ratio of the amount of departures of good packets to the total outgoing flow. However, as argued in [11], a more suitable performance measure is the *goodput*, defined as the ratio of packets belonging to uncorrupteed frames and the total amount of packets that arrive at the network element's input.

The goal of this paper is to present *explicit expressions* for the stationary queue size distribution and the goodput of the PMD policy. Most of the previous works [9,11], deal with numerical studies of the performance of PMD policy. The first part of the paper considers a Markovian framework: a Poisson process of packet arrivals, geometrically distributed frame size, and exponentially distributed service times of packets. Explicit expressions for the queue size distribution and of the goodput are obtained based on recursions introduced in [11]. The originality of the first part is in providing closed form expressions for the stationary distribution of queue size and the goodput ratio.

As part of our packet level analysis, we propose an interpretation of the queueing model as equivalent to a dual of a vacation model (queue with service vacation). Using this interpretation, we provide a simpler analysis for the limiting heavy traffic regime (when buffer size approaches infinity).

In the second part of the paper we derive a fluid approximation which is valid for heavy traffic conditions. The input process (which may be quite general) is approximated by a fluid with a constant rate. We obtain explicit expressions for the workload process distribution and the goodput for the fluid approximation. The fluid analysis yields analytically tractable, simple expressions which will be helpful in analytical study of the sensitivity of the goodput to different parameters for, e.g., the message length, the buffer size, etc., which were studied numerically in earlier works [9,11]. Also, the nature of dependence of goodput to various parameters is clear, in particular the *goodput depends on the mean message length and the buffer size only through their product*. Our analytical results may be quite useful in dimensioning the buffer and/or capacity that is required for a given required goodput under PMD policy.

The structure of the paper is as follows. Section 2 analyzes the packet model. It is composed of the model description (section 2.1), the queue length distribution (section 2.2), an alternative modeling through a dual vacation model (section 2.3), of the vacation approach to the case of large buffer (section 2.4), analysis of the goodput (section 2.5) and numerical investigations (section 2.6). Section 3 analyzes the fluid model. It is composed of the model description (section 3.1), the derivation of the distribution of the workload process (section 3.2), the goodput analysis (section 3.3) and a numerical study of the fluid model (section 3.4). We then end with a concluding section. Some of the technical derivations are delayed to the appendix.

2. Packet model

2.1. Model description

We consider a single M/M/1/N queue.[1] The arrival rate is λ packets per second and the service rate is μ packets per second. Define the load $\rho = \lambda/\mu$.[2] A message length (in terms of packets) is considered to be geometrically distributed with parameter q. In PMD policy, if a packet arrives when the queue is full, it is discarded and all the subsequent packets belonging to the same message are also discarded, irrespective of the state of the queue upon their arrival epochs, until the head-of-message packet (i.e., a new message) arrives. To model the policy, two modes for working of the network element are defined: the *normal mode*, in which packets are admitted, and the *discarding mode*, in which arriving packets are discarded. The state transition diagram for PMD policy under this model is shown in figure 1. The packet model is the same as the one employed in [11]. Let $P_{i,j}$ $(0 \leqslant i \leqslant N,\ j = 0, 1)$ be the steady-state probability of having i packets in the system with the system in mode j ($j = 0$ for normal; $j = 1$ for discarding). Thus, we have the following set of equations for the steady-state probabilities [11] from figure 1:

$$\rho P_{0,0} = P_{1,0}, \tag{1}$$

$$q\rho P_{0,1} = P_{1,1}, \tag{2}$$

$$(\rho + 1)P_{i,0} = \rho P_{i-1,0} + P_{i+1,0} + q\rho P_{i-1,1}, \quad \text{for } 1 \leqslant i \leqslant N-1, \tag{3}$$

$$(q\rho + 1)P_{i,1} = P_{i+1,1}, \quad \text{for } 1 \leqslant i \leqslant N-1, \tag{4}$$

$$(\rho + 1)P_{N,0} = \rho P_{N-1,0} + q\rho P_{N-1,1}, \tag{5}$$

$$P_{N,1} = \rho P_{N,0}, \tag{6}$$

$$\sum_{i=0}^{N}(P_{i,0} + P_{i,1}) = 1. \tag{7}$$

Let $Q_j(z) = \sum_{i=0}^{N} z^i P_{i,j}$ ($j = 0$ for normal mode and $j = 1$ for discarding mode) and $Q(z) = Q_0(z) + Q_1(z) = \sum_{i=0}^{N}(P_{i,0} + P_{i,1})z^i$.

Remark 1. In many practical applications, the distribution of packet sizes and inter-arrival times may be more general. In particular, frequently packets have a constant

[1] Though we do a single node (router) analysis, we would like to comment that analysis by approximating the whole chain of routers (between the source and the destination) by one single router which experiences the maximum losses (*the bottleneck*) has both theoretical and experimental [3,4] justification (see also [5]). In this sense our single node should be looked on as the bottleneck node. The service time represents the time between the beginning of the transmission of a packet on the bottleneck interface until the beginning of the transmission of the next packet from the same flow.

[2] Although we consider the analysis of a single connection, our model could also be useful for the case of multiplexing. In the latter case, two packets of a flow can be spaced apart by a random number of packets from different flows; we may add this to the service time of a packet and use the exponential distribution as an (approximating) candidate for modeling the service times in an equivalent model with a single flow.

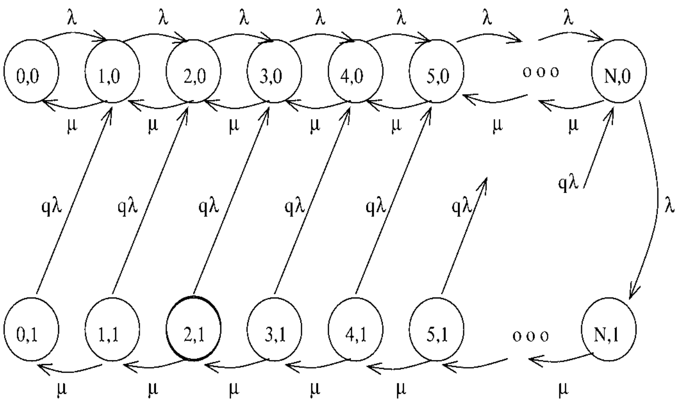

Figure 1. Transition structure under the PMD policy.

size (this is the case in ATM networks [11]). Yet our model can often serve as a good approximation for other distributions. In particular, a numerical investigation is provided in [11, section 5] that shows that the results for our model very well approximate those obtained for constant packet size. Moreover, we expect our model to be insensitive to both packet size distribution as well as to interarrival times distribution in the heavy traffic regime, for which we provide in section 3 a fluid limiting model (whose performance measures depend only on arrival and departure rates of packets and not on their distributions).

2.2. *PGF and distribution of the number of packets in the queue*

Proposition 1. The probability generating function $Q_j(z)$ is given by

$$Q_0(z) = P_{0,1} q \Bigg[\rho^{-1} \frac{1 - (\rho^{-1}(1+\rho q))^N}{1 - \rho^{-1}(1+\rho q)} + \sum_{i=1}^{N} \frac{[1 - (\rho^{-1}(1+\rho q))^{N-i+1}](1+\rho q)^{i-1}}{1 - \rho^{-1}(1+\rho q)} z^i \Bigg], \tag{8}$$

$$Q_1(z) = P_{0,1} \frac{(1 - z - z^{N+1}(q\rho + 1)^N q\rho)}{1 - z(q\rho + 1)} \tag{9}$$

with

$$P_{0,1} = \frac{(1-\rho)(\rho(1-q) - 1)}{D}$$

and

$$D = q\big(1 - \rho^{-N}(1+\rho q)^N\big) + \big(\rho(1-q) - 1\big)\big(1 - \rho(1+\rho q)^N\big)$$

The proof is given in the appendix.

By taking the inverse *z-transform* of equation (8) and (9) we obtain:

Corollary 1. The steady state probabilities are given by

$$P_{i,0} = \begin{cases} \dfrac{P_{0,1}q}{\rho}\dfrac{1-(\rho^{-1}(1+\rho q))^{N}}{1-\rho^{-1}(1+\rho q)} & \text{for } i = 0, \\ P_{0,1}q\dfrac{[1-(\rho^{-1}(1+\rho q))^{N-i+1}](1+\rho q)^{i-1}}{1-\rho^{-1}(1+\rho q)} & \text{for } 1 \leqslant i \leqslant N, \end{cases}$$

and

$$P_{i,1} = \begin{cases} \dfrac{(1-\rho)(\rho(1-q)-1)}{D} & \text{for } i = 0, \\ q\rho(1+q\rho)^{i-1}P_{0,1} & \text{for } 1 \leqslant i \leqslant N. \end{cases}$$

2.3. An equivalent vacation model

Next, we propose an interpretation of our model (denoted by P) as equivalent to a dual vacation model (denoted by P_d). This interpretation will especially be useful later, in considering the case of large buffers. Let $A_{t,N}$ be the number of packets in the P model at time t. We define a random variable, $Y_{t,N}$ as

$$Y_{t,N} = N - A_{t,N}.$$

$Y_{t,N}$[3] will be the number of packets in the equivalent vacation model. In other words, the number of packets in P_d equals the number of vacant places in the P. It thus follows that the service times in P_d are i.i.d. exponentially distributed with parameter λ, and that the arrival process to P_d is Poisson distributed with parameter μ. We note, however, that in the original process, arrivals are stopped during the *discarding period.* This *discarding period* will correspond to a *vacation period* in the dual model. More precisely, we define a discarding period in P as the duration from the instant that the buffer fills, till the next time the event B^1 occurs, where $B^1 = \{$service occurs and the next packet to arrive after that service is not discarded$\}$.

Note that with this definition, the number of packets discarded during a discarding period may be zero. Indeed, with probability $\mu/(\lambda + \mu) = 1/(1 + \rho)$, the first packet that arrives after the queue fills, will find the queue vacant and will not be discarded, and there are zero discarded packets in the discarding period.

A particularly important quantity in the equivalent vacation model is the distribution of the number of arrivals during a vacation denoted by V, or alternatively, the number of service periods during a discarding period in the original model. Note that by definition of the original discarding period $V \geqslant 1$.

Let T denote the instant of the beginning of a discarding period, and let S denote the instant when the first service completion occurs after T. Let V_1 (respectively V_2) be a r.v. distributed as V given that at time S, the message that is being transmitted is bad (respectively good). In other words, $V =_d V_2$ if and only if the next packet to arrive after

[3] N in subscript is for the buffer size N.

time S is not discarded. The latter occurs if and only if the following event B^2 occurs; $B^2 = \{$either there is no arrival during the interval (T, S), or there is at least one arrival but the last arrival in that interval is the last packet of a message$\}$. Let

$$\beta := \frac{(1-q)\rho}{1+\rho} \quad \text{thus } 1-\beta = \frac{1+q\rho}{1+\rho}.$$

Thus $V =_{\mathrm{d}} V_1$ with probability β and $V =_{\mathrm{d}} V_2 = 1$ with probability $1-\beta$.

Next we study the distribution of V_1. V_1 equals in distribution to one plus $\mathcal{A} :=$ the number of services that occur during the duration of M arrivals, where M is geometrically distributed with parameter q. Let $B(L)$ be the number of services in a random duration L. Then $B_L^*(z)$, the p.g.f. of $B(L)$, is given by

$$B_L^*(z) = E\left[\sum_{i=0}^{\infty} \mathrm{e}^{-\mu L}\frac{(\mu L)^i}{i!}z^i\right] = E\left[\mathrm{e}^{-\mu L(1-z)}\right] = L^*\big(\mu(1-z)\big)$$

where $L^*(\cdot)$ is the LST of L. Thus we need to evaluate $L^*(\cdot)$.

$$L^*(s) = E\left[\mathrm{e}^{-sL}\right] = E\left[E\left[\mathrm{e}^{-sT_i}\right]^M\right] = E\left[\big(T^*(s)\big)^M\right] = \mathcal{N}\big(T^*(s)\big)$$

where $T^*(\cdot)$ is LST of an interarrival time ($\sim$exp λ) and $\mathcal{N}(\cdot)$ is the p.g.f. of a geometrically distributed r.v. with parameter q. Observe that,

$$T^*(s) = \frac{\lambda}{\lambda+s} \quad \text{and} \quad \mathcal{N}(z_1) = \frac{qz_1}{1-(1-q)z_1}.$$

Thus, if we denote by $\alpha(z)$ the p.g.f. of $\mathcal{A}$, then

$$\begin{aligned}\alpha(z) &= \mathcal{N}\big(T^*\big(\mu(1-z)\big)\big) = \mathcal{N}\left(\frac{\lambda}{\lambda+\mu(1-z)}\right)\\ &= \frac{q}{\rho^{-1}(1-z)+q}\end{aligned}$$

The p.g.f. of V is given by $z((1-\beta)+\beta\alpha(z))$.

2.4. *The case of large buffer*

We use now the interpretation proposed in section 2.3 as the dual of a vacation model in order to study the behavior of our system as the buffer size N becomes large. Clearly, nontrivial distribution of $Y_{t,N}$ is obtained in the limit $N \to \infty$ only in the heavy traffic regime $\rho > 1$.

Observe that,

$$\lim_{N\to\infty} P(N - A_{t,N} = k) = \lim_{N\to\infty} P(Y_{t,N} = k) = P(Y_t = k).$$

In [8] (see also [12]) the authors have shown that the stationary number of customers present in a M/G/1 queueing system with generalized server vacation is a convolution of the distribution function of two independent positive random variables (*stochastic decomposition*), one of which being the stationary distribution of the number of customers in an ordinary M/G/1 queueing system without server vacations. The other corresponds to the p.g.f. of the number of packets at an arbitrary moment in a vacation. Let $\phi(\cdot)$ and $\pi(\cdot)$ be the p.g.f. for the stationary distribution of the number of customers at a random point in time in the in the vacation system and in the standard M/G/1 queueing system, respectively. Also, let $\hat{\alpha}(\cdot)$ denote the p.g.f. of the random variable V (i.e., the number of customers that arrive during a vacation). Then, with arrival rate μ (service rate in PMD queue) and departure rate λ (arrival rate in PMD queue), and $\rho = \lambda/\mu$, we have from [8]

$$\phi(z) = \frac{1 - \hat{\alpha}(z)}{\dot{\hat{\alpha}}(1)(1 - z)} \pi(z)$$

with

$$\pi(z) = \frac{(1 - \rho^{-1})(1 - z) B^*(\mu - \mu z)}{B^*(\mu - \mu z) - z}$$

where $B^*(\cdot)$ is the Laplace transform of the service time p.d.f. For our M/M/1/V case, $\pi(z)$ simplifies to $\pi(z) = (1 - \rho^{-1})/(1 - \rho^{-1} z)$. Using the result of the previous subsection, and inverting the p.g.f. $\phi(z)$, we get, finally,

$$P(Y = k) = \frac{q(\rho - 1)}{(\rho(1 - q) - 1)} \big(\rho(1 - q)(1 + \rho q)^{-(k+1)} - \rho^{-(k+1)}\big).$$

One can now check that this is indeed the limit obtained as $\lim_{N\to\infty} P(A_N = N - k)$ from corollary 1. We have, for $1 \leqslant k \leqslant N - 1$,

$$\begin{aligned}
&\lim_{N\to\infty} P(A_N = N - k) \\
&\quad = \lim_{N\to\infty} (P_{N-k,0} + P_{N-k,1}) \\
&\quad = \lim_{N\to\infty} \big[\rho(\rho - 1) q (1 + q\rho)^{N-(k+1)} \big[\rho(1 - q) - \rho^{-(k+1)} (1 + \rho q)^{-(k+1)}\big]\big] D^{-1} \\
&\quad = P(Y = k).
\end{aligned}$$

2.5. *Goodput ratio $\mathcal{G}$*

The goodput is defined in [11] as the ratio between total packets comprising good messages exiting the system and the total arriving packets at its input. Let $\mathcal{W}$ be the length (number of packets) of an arriving message, $\mathcal{Q}$ denote the queue length seen by the first packet of the arriving message and $\mathcal{V}$ be the random variable representing the success of a message ($\mathcal{V} = 1$ for a good message, and $\mathcal{V} = 0$ for a message which has one or more dropped packets). Then $\mathcal{G}$ can be expressed (see [11]), with

$S_{n,i} \triangleq P(\mathcal{V} = 1 \mid \mathcal{W} = n,\ \mathcal{Q} = i)$, as

$$\mathcal{G} = q \sum_{n=1}^{\infty} nq(1-q)^{n-1} \sum_{i=0}^{N} P(\mathcal{V} = 1 \mid \mathcal{W} = n,\ \mathcal{Q} = i) P(\mathcal{Q} = i). \tag{10}$$

In [11], recursions for evaluating these probabilities and hence $\mathcal{G}$ were given. We will present here an explicit expression for $\mathcal{G}$. To do this we will use the multidimensional generating function for probabilities $S_{n,i}$ which was obtained in a different context in [1]. By some abuse of notation let us denote $\overline{S}_i(x)$ $(= \sum_{n=1}^{\infty} S_{n,i} x^{n-1})$ and by $\overline{S}_n(y) (= \sum_{i=0}^{N} S_{n,i} y^i)$ as the generating function for probabilities $S_{n,i}$ $(1 \leqslant n \leqslant \infty$, $0 \leqslant i \leqslant N)$ for fixed i and fixed n, respectively. We also define the two-dimensional generating function of $S_{n,i}$ as $\overline{S}(x, y)$, i.e., $\overline{S}(x, y) = \sum_{n=1}^{\infty} \sum_{i=0}^{N} S_{n,i} y^i x^{n-1}$.

Proposition 2. The probability generating function $\overline{S}(x, y)$ can be expressed as $\overline{S}(x, y) = \sum_{i=0}^{N} c_i(x) y^i$ where, for $0 \leqslant i \leqslant N-1$,

$$c_i(x) = \begin{cases} 1 + K_3\big(A_1 - A_2 y_1^{N-(i+1)} - A_3 y_2^{N-(i+1)}\big) + K_4\big(B_1 y_1^{N-i} + B_2 y_2^{N-i}\big), \\ \quad 0 \leqslant i \leqslant N-1, \\ 0, \quad i = N, \end{cases}$$

with[4]

$$y_{1,2} = \frac{1 + \rho \pm \sqrt{(1+\rho)^2 - 4\rho x}}{2}, \qquad K_3 = -x\rho,$$

$$K_4 = \frac{x\rho(y_1^N - y_2^N)}{y_2^{N+1}(y_1 - \rho) - y_1^{N+1}(y_2 - \rho)}, \qquad A_1 = \frac{1}{(1 - y_1)(1 - y_2)},$$

$$A_2 = \frac{1}{(1 - y_1)(y_1 - y_2)}, \qquad A_3 = \frac{1}{(1 - y_2)(y_2 - y_1)} \quad \text{and} \quad B_1 = -B_2 = \frac{1}{y_1 - y_2}.$$

Proof. From [1], we have

$$\big[(1 - \alpha y)\alpha y - x\rho\alpha^2\big]\overline{S}(x, y) = \frac{1 - y^N}{1 - y}(1 - \alpha y)\alpha y - x\rho\alpha^2(\alpha y)^{N+1} K_1 + x\alpha^2(y - \rho)K_2$$

with

$$K_1 = \frac{\alpha^{-(N+1)}(y_1^N - y_2^N)}{y_2^{N+1}(y_1 - \rho) - y_1^{N+1}(y_2 - \rho)},$$
$$K_2 = \frac{1}{(y_1 - \rho)(y_2 - \rho)}\left[-1 + y_1^N + \frac{\rho y_1^{N+1}(y_2 - \rho)(y_1^N - y_2^N)}{y_2^{N+1}(y_1 - \rho) - y_1^{N+1}(y_2 - \rho)}\right]$$

[4] It should be noted that all the apparent constants $y_{1,2}$, $K_{3,4}$, $A_{1,2,3}$ and $B_{1,2}$ are functions of x.

where y_1 and y_2 are the roots of the equation $(1-\alpha y)\alpha y - x\rho\alpha^2 = 0$, i.e.,

$$y_{1,2}(x) = \frac{1+\rho \pm \sqrt{(1+\rho)^2 - 4\rho x}}{2}.$$

Also $y_1 + y_2 = 1+\rho$ and $y_1 y_2 = \rho x$. We will now represent $\overline{S}(x, y)$ as $\sum_{i=0}^{N} c_i(x) y^i$. Observe that,

$$\overline{S}(x, y) = G_1(y) - K_3 G_2(y) + K_4 G_3(y) - K_5 G_4(y)$$

where

$$G_1(y) = \frac{1-y^N}{1-y}, \qquad G_2(y) = \frac{(1-y^N)}{(1-y)(y-y_1)(y-y_2)},$$
$$G_4(y) = \frac{(y-\rho)}{(y-y_1)(y-y_2)}, \qquad G_3(y) = \frac{y^{N+1}}{(y-y_1)(y-y_2)}$$

and $K_3 = -x\rho$, $K_4 = x\rho\alpha^{N+1}K_1$ and $K_5 = -xK_2$. We shall now apply the partial fraction method and express the right-hand side of the last equation in the form of $(y^k - a^k)/(y-a)$ for some k and a. Thus we can write,

$$\begin{aligned}\overline{S}(x, y) = {}& (1+K_3A_1)\frac{1-y^N}{1-y} - K_3A_2\frac{y_1^N - y^N}{y_1 - y} - K_3A_3\frac{y_2^N - y^N}{y_2 - y} \\ & + K_4B_1\frac{y_1^{N+1} - y^{N+1}}{y_1 - y} + K_4B_2\frac{y_2^{N+1} - y^{N+1}}{y_2 - y} \\ & + \big(K_3A_2\big(1-y_1^N\big) + K_4B_1y_1^{N+1} + K_5C_1\big)\frac{1}{y-y_1} \\ & + \big(K_3A_3\big(1-y_2^N\big) + K_4B_2y_2^{N+1} + K_5C_2\big)\frac{1}{y-y_2}\end{aligned}$$

where

$$A_1 = \frac{1}{(1-y_1)(1-y_2)}, \qquad A_2 = \frac{1}{(1-y_1)(y_1-y_2)}, \qquad A_3 = \frac{1}{(1-y_2)(y_2-y_1)},$$
$$B_1 = \frac{1}{y_1-y_2}, \qquad B_2 = \frac{1}{y_2-y_1}, \qquad C_1 = \frac{y_1-\rho}{y_1-y_2} \quad \text{and} \quad C_2 = \frac{y_2-\rho}{y_2-y_1}.$$

But

$$K_3A_2\big(1-y_1^N\big) + K_4B_1y_1^{N+1+K_5C_1} = K_3A_3\big(1-y_2^N\big) + K_4B_2y_2^{N+1} + K_5C_2 = 0.$$

This is because $\overline{S}(x, y)$ is analytic in y, the left-hand side of equation (11) vanishes at $y = y_i$, $i = 1, 2$. Hence, the above equation can be written as

$$\overline{S}(x,y) = (1+K_3A_1)\frac{1-y^N}{1-y} - K_3A_2\frac{y_1^N - y^N}{y_1 - y} - K_3A_3\frac{y_2^N - y^N}{y_2 - y} + K_4B_1\frac{y_1^{N+1} - y^{N+1}}{y_1 - y} + K_4B_2\frac{y_2^{N+1} - y^{N+1}}{y_2 - y}.$$

Again, recalling that

$$\frac{a^k - y^k}{a - y} = a^{k-1} + a^{k-2}y + a^{k-3}y^2 + \cdots + ay^{k-2} + y^{k-1}$$

and grouping the coefficients of the same power of y we get $\overline{S}(x,y) = \sum_{i=0}^{N} c_i(x)y^i$. □

Having expressed $\overline{S}(x,y)$ as $\sum_{i=1}^{N} c_i(x)y^i$ (in proposition 2) we now proceed to obtain the expression for $\mathcal{G}$ using $\overline{S}(x,y)$.

Proposition 3. The goodput ratio, $\mathcal{G}$ can be written as

$$\mathcal{G} = q^2\sum_{i=0}^{N}\left(\frac{\mathrm{d}(xc_i(x))}{\mathrm{d}x}\right)_{x=(1-q)} P(Q=i) = q^2\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{i=0}^{N} xc_i(x)P(Q=i)\right)\right]_{x=(1-q)}.$$

Proof. We know by equation (10),

$$\mathcal{G} = q\sum_{n=1}^{\infty} nq(1-q)^{n-1}\sum_{i=0}^{N} S_{n,i}P(Q=i) = q^2\sum_{i=0}^{N}\sum_{n=1}^{\infty} S_{n,i}n(1-q)^{n-1}P(Q=i).$$

Also,

$$\overline{S}(x,y) = \sum_{i=0}^{N} y^i \sum_{n=1}^{\infty} S_{n,i}x^{n-1} = \sum_{i=0}^{N} c_i(x)y^i.$$

Thus, $c_i(x) = \sum_{n=1}^{\infty} S_{n,i}x^{n-1}$ and

$$\left(\frac{\mathrm{d}(xc_i(x))}{\mathrm{d}x}\right)_{x=(1-q)} = \sum_{n=1}^{\infty} nS_{n,i}(1-q)^{n-1}.$$

Thus,

$$\mathcal{G} = q^2\sum_{i=0}^{N}\sum_{n=1}^{\infty} S_{n,i}n(1-q)^{n-1}P(Q=i) = q^2\sum_{i=0}^{N}\left(\frac{\mathrm{d}(xc_i(x))}{\mathrm{d}x}\right)_{x=(1-q)} P(Q=i).$$

Thus, we can obtain the exact expression for the *goodput ratio* by knowing the coefficients $c_i(x)$ and $P(Q=i)$ $(= P_{i,0} + P_{i,1})$, for $0 \leqslant i \leqslant N$ (both being previously obtained in corollary 1 and proposition 2, respectively). Since the derivation as well as the final result are complex, we defer these to the appendix. We note that though the final closed form expression for $\mathcal{G}$ is complex, one can obtain significant insights into the dependence of $\mathcal{G}$ on different parameters of the network. In particular, the expression for $\mathcal{G}$ can be exploited to dimension the buffer size for QoS provisioning (like maximizing the goodput ratio). □

Remark 2. Suppose we add an economic feature to our model, by assigning a reward of γ per packet that is received and belongs to a good message, and a cost of ζ per packet that belongs to a bad message. Then the over all average rate of utility is

$$U = \lambda\big(\gamma\mathcal{G} - \zeta(1 - \mathcal{G})\big).$$

Observe that maximizing $\mathcal{G}$ also maximizes U. Thus measure $\mathcal{G}$ can be used as a user centric pricing scenario: the users only pay for the good messages that the network delivers and thus is an indication of quality perceived by the source and for which it can be charged by the network.

2.6. Numerical examples

An extensive numerical investigation of the packet model is available in [11]. In particular, it provides the buffer size needed so as to achieve a given throughput. It also shows that the model is robust to the distribution of packet size: the performance obtained for a fixed packet size are well approximated by our model.

The goal of this subsection is to briefly examine the dependence of the goodput on the message lengths and also on the buffer size in view of our observations in section 2.4.

Below we plot the $\mathcal{G}$ obtained with our explicit formula from equation (A.14) (in the appendix) with increasing load ρ. We first keep N fixed at 10 (50) and plot for $1/q = 5, 10, 15, 20, 25, 30$ with ρ varying from 0.1 to 3.0 (in steps of 0.1) in figure 2 (figure 3). Next we keep $1/q$ fixed at 20 (2) and plot for $N = 5, 10, 25, 50, 100, 200$, again with ρ varying from 0.1 to 3.0 (in steps of 0.1) in figure 4 (figure 5). We observe a limiting value of $\mathcal{G}$ as $1/q$ becomes large for fixed N in figures 2 and 3. Also for large values of N, the closeness to this limit is pronounced even at low values of mean message length, i.e., $1/q$. We also observe a limiting value of $\mathcal{G}$ as N becomes large for fixed $1/q$ in figures 4 and 5 and for small values of $1/q$, the closeness to this limit is pronounced even at low values of buffer sizes. This behavior supports the analysis in section 2.4 where we showed that there exists a nontrivial limiting behavior as N becomes large, while keeping all other parameters the same.

So far we have done packet-level performance evaluation of the PMD policy for a M/M/1/N queue model. The explicit expression for goodput is somewhat complex. In the next section we propose a fluid model for analyzing the PMD policy towards obtaining simple approximations for the goodput. The fluid model can be seen as a weak limit of the original packet model through a standard scaling argument.

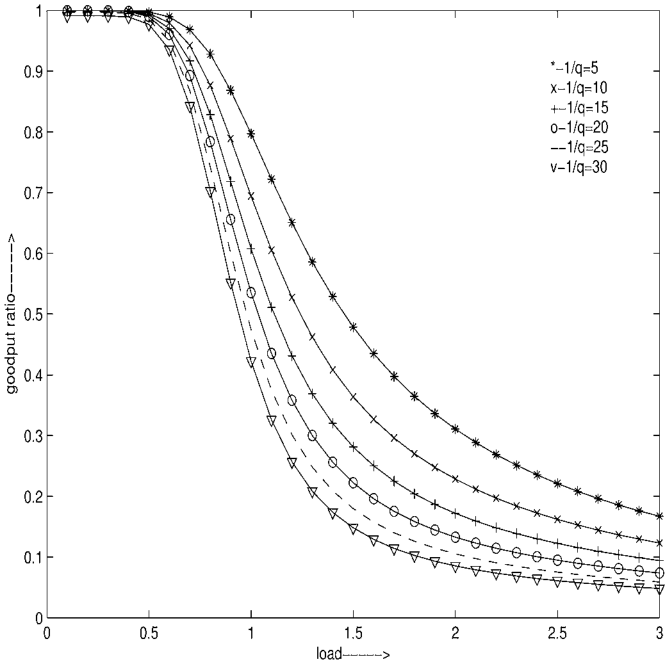

Figure 2. $\mathcal{G}$ vs. ρ for $1/q = 5, 10, 15, 20, 25, 30$ with $N = 10$.

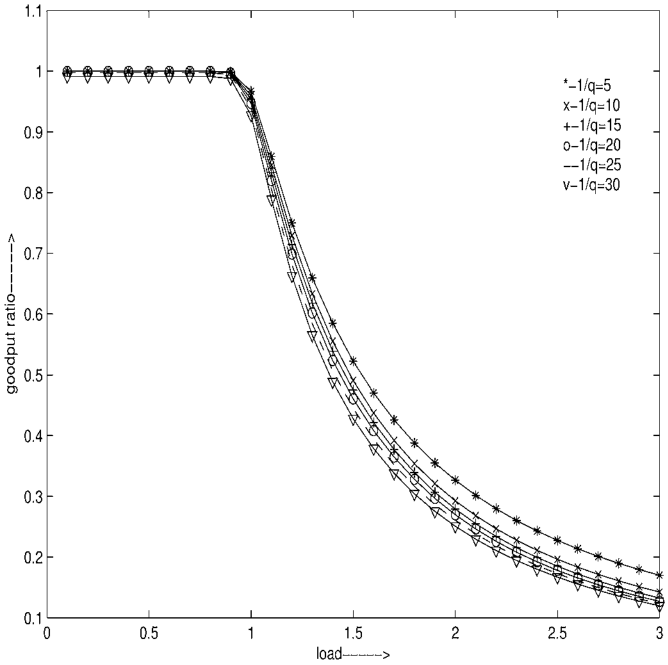

Figure 3. $\mathcal{G}$ vs. ρ for $1/q = 5, 10, 15, 20, 25, 30$ with $N = 50$.

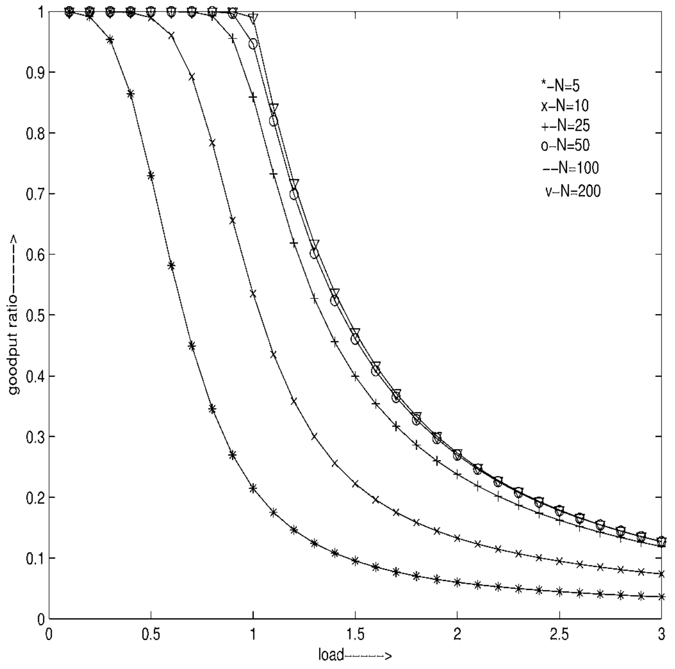

Figure 4. $\mathcal{G}$ vs. ρ for $N = 5, 10, 15, 20, 25, 30$ with $1/q = 20$.

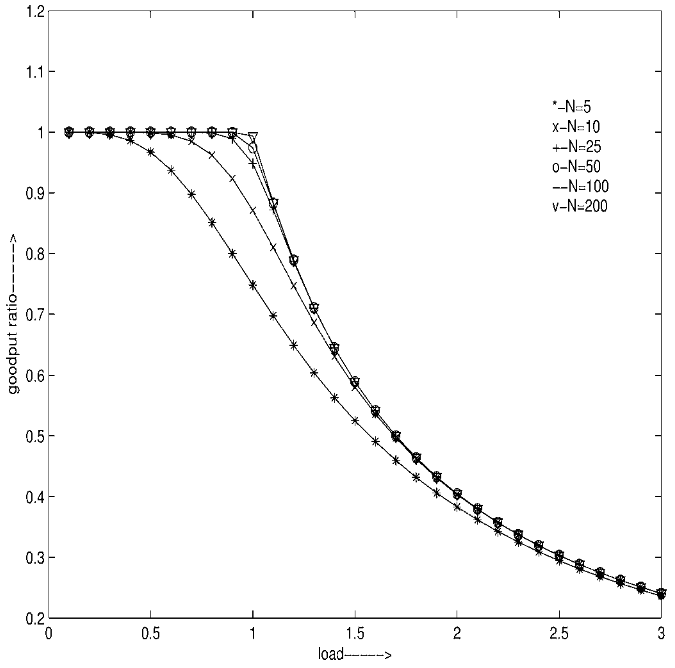

Figure 5. $\mathcal{G}$ vs. ρ for $N = 5, 10, 15, 20, 25, 30$ with $1/q = 2$.

3. Fluid approximation

3.1. Model description

Our fluid source always has messages to send and the capacity of the fluid buffer is finite, say B. The fluid buffer is served with a capacity c. The length of a message is assumed to be exponentially distributed with parameter η. If during the arrival of a message, the workload process $V(t)$ (alternatively the queue length, i.e., amount of fluid in the fluid buffer) reaches B, then all the remaining fluid corresponding to this message is dropped. Let the fluid arrival rate be h.

Remark 3. The fluid limit can be seen as a weak limit of the original model through a standard scaling. More precisely, consider n models, and add n as a superscript to the parameters of the nth model. Then the scaling is obtained as follows:

- Arrival rate: $\lambda^{(n)} := n\lambda$;
- Service rate: $\mu^{(n)} := n\mu$;
- Size of messages: geometrically distributed with parameter $q^{(n)} := q/n$;
- Buffer size: $N^{(n)} = nB$.

Let $X^{(n)}(t)$ be the queue length process of the nth model. Then, as $n \to \infty$, the process $X^{(n)}(t)/n$ weakly converges to our fluid process $V(t)$, with $h = \lambda$, $c = \mu$ and with $\eta = q\lambda$.

A typical evolution of $V(t)$ in our model is shown in figure 6. Also, let A be the event that the incoming fluid is accepted. To remove trivialities we assume that $c < h$.

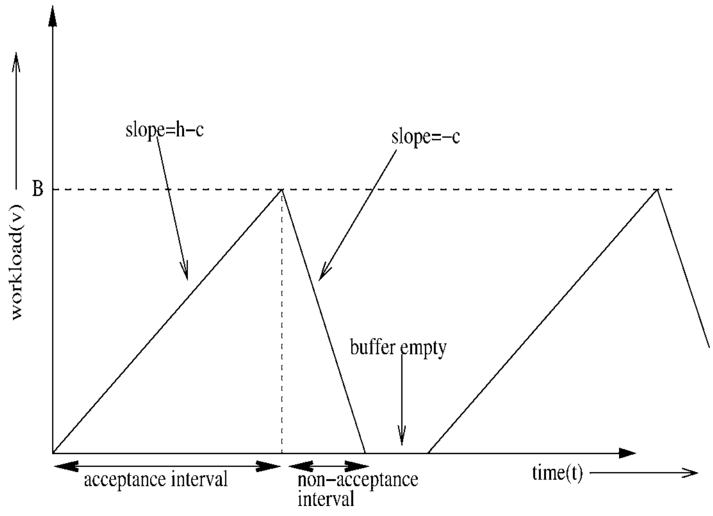

Figure 6. A typical evolution of the workload process $V(t)$ in our fluid model.

Let V_n be the random variable denoting the queue length at the end of nth non-acceptance period. The dynamics of V_n can be written as

$$V_{n+1} = V_n + (h - c)A_n - cA_n^c \tag{11}$$

where $A_n = (B - V_n)/(h - c)$ is the duration of the $(n + 1)$st acceptance period and $A_n^c = \min(X_n, B/c)$ is the duration upto which the process $V(t)$ will have a negative slope where X_n is the remaining length of the current incoming message at the epoch when $V(t)$ hits B, i.e., at the end of the $(n + 1)$th acceptance period and the start of the $(n + 1)$th non-acceptance period. Let T_n be the epoch of the commencement of the $(n+1)$st acceptance event. Thus, $T_{n+1} = T_n + A_n + X_n$ and $(V_n, A_n + X_n)$ can be viewed as a marked point process [2]. Thus,

$$V_{n+1} = B - cA_n^c.$$

Denote the steady state Laplace–Stieltjes Transform (LST) of V_n, i.e., $E[\mathrm{e}^{-s(B-cA_n^c)}]$ by $V_{\mathrm{p}}(s)$.[5] Let ρ_{p} be the probability density of V_n (the inverse of LST of V_n) in steady state.

3.2. *The workload process*

In this section we provide the distribution of the workload process.

Lemma 1. The LST and the probability density of V_n in steady state are given by

$$V_{\mathrm{p}}(s) = \mathrm{e}^{-sB}\eta\left[\frac{1-\mathrm{e}^{(B/c)(\eta-sc)}}{\eta - sc} + \frac{\mathrm{e}^{sB}\mathrm{e}^{-\eta B/c}}{\eta}\right],$$
$$\rho_{\mathrm{p}}(v) = \frac{\eta}{c}\mathrm{e}^{(\eta/c)(v-B)} + \mathrm{e}^{-\eta B/c}\delta(v)$$

(for $v \in [0, B)$).

In the above lemma we obtained the workload LST and probability density at the end of non-acceptance periods. Next we shall obtain these quantities at an arbitrary time, i.e. the LST and the probability density of the time stationary workload.

Proposition 4. The LST $V(s)$ and the probability density ρ of $V(t)$ in stationary regime are given by

$$V(s) = \frac{\eta h}{(h - c\mathrm{e}^{-B\eta/c})}\mathrm{e}^{-sB}\left(\frac{1-\mathrm{e}^{-B(\eta-sc)/c}}{\eta - sc}\right) - \frac{h(1-\mathrm{e}^{-\eta B/c})}{(h - c\mathrm{e}^{-B\eta/c})} + 1. \tag{12}$$

For $v \in [0, B)$,

$$\rho(v) = \rho^0(v) + \rho^1(v)$$

[5] The subscript p indicates the point process.

where $\rho^0(v) = (1 - h(1 - \mathrm{e}^{-\eta B/c})/(h - c\mathrm{e}^{-B\eta/c}))\delta(v)$, $\rho^1(v) = \eta h/(h - c\mathrm{e}^{-B\eta/c})(\mathrm{e}^{(\eta/c)(v-B)}/c)$ and $\rho(v) = 0$ for $v \geqslant B$. Finally, the mean stationary workload M_f is given by

$$M_\mathrm{f} = \frac{h}{h - c\mathrm{e}^{-B\eta/c}}\left(B - \frac{c}{\eta}\big(1 - \mathrm{e}^{-B\eta/c}\big)\right). \tag{13}$$

Proof. We shall now use the following inversion formula (see, e.g., [2, chapter 1, section 4]) to obtain the LST for the workload process $V(t)$ which we will then invert to get the probability density function

$$E\big[\mathrm{e}^{-sV(t)}\big] = \frac{E^0[\int_0^{T_1} \mathrm{e}^{-sV(t)}]}{E^0[T_1]}.$$

Thus,

$$\begin{aligned}
V(s) &= \frac{E^0\big[\int_0^{(B-V_0)/(h-c)} \mathrm{e}^{-s(V_0+(h-c)t)}\,\mathrm{d}t}{E^0[(B - V_0)/(h - c) + X_0]} \\
&\quad + \frac{\int_{(B-V_0)/(h-c)}^{(B-V_0)/(h-c)+A_0^c} \mathrm{e}^{-s(B-c(t-(B-V_0)/(h-c)))}\,\mathrm{d}t + X_0 - A_0^c\big]}{E^0[(B - V_0)/(h - c) + X_0]} \\
&= \frac{(h - c)\eta}{(h - c\mathrm{e}^{-B\eta/c})}\left[\frac{E^0[\mathrm{e}^{-sV_0}] - \mathrm{e}^{-sB}}{s(h - c)} + \frac{\mathrm{e}^{-sB}(E^0[\mathrm{e}^{scA_0^c}] - 1)}{sc} + E[X_0] - E[A_0^c]\right].
\end{aligned}$$

Observe that we have expressions for $E[\mathrm{e}^{-sV_0}]$ (i.e., $V_\mathrm{p}(s)$), $E[V_0]$ (can be obtained from $V_\mathrm{p}(s)$) and the expression for $E[\mathrm{e}^{scA_0^c}]$ and hence for $E[A_0^c]$ can be easily obtained by the definition of A_0^c. Thus, we get after some calculations (12). The inverse of the LST of the last equation gives $\rho(v)$. Finally, M_f is obtained by the integration:

$$M_\mathrm{f} = \int_0^B v\rho(v)\,\mathrm{d}v = \int_0^B v\frac{\eta h}{(h - c\mathrm{e}^{-B\eta/c})}\left(\frac{\mathrm{e}^{(\eta/c)(v-B)}}{c}\right)\mathrm{d}v$$

which implies (13). □

3.3. *The goodput ratio $\mathcal{G}_\mathrm{f}$*

We proceed with the model from the previous subsection and in particular, we continue to assume that $c < h$ in order to avoid trivialities. We define, the fluid analog of the *goodput ratio*, $\mathcal{G}_\mathrm{f}$ as the ratio of the total fluid comprising good messages (i.e., messages which do not suffer any fluid loss due to buffer overflow) exiting the node to the total arriving fluid at its input. Let $\mathcal{V}_\mathrm{f}$ be the random variable representing the success of a message, $\mathcal{V}_\mathrm{f} = 1$ for a good message, and $\mathcal{V}_\mathrm{f} = 0$ for a message which has lost some fluid. Let us define the sub-distribution function $F(w, 1)$ as the probability that a message is of length $\leqslant w$ and is good, i.e., $F(w, 1) = P(W \leqslant w,\ \mathcal{V}_{\mathrm{f}} = 1)$. Then we can write the goodput ratio as

$$\mathcal{G}_{\mathrm{f}} = \frac{\int_0^\infty w\,\mathrm{d}F(w,1)}{\int_0^\infty w\,\mathrm{d}F(w)}$$

where $F(w)$ is the message length distribution ($\sim\exp\eta$). Again, writing $F(w, 1)$ as,

$$\begin{aligned} F(w,1) &= P(\mathcal{V}_{\mathrm{f}} = 1 \mid W \leqslant w) P(W \leqslant w) \\ &= \int_0^B P(\mathcal{V}_{\mathrm{f}} = 1 \mid W \leqslant w,\ V = v)\rho(v)\,\mathrm{d}v \int_0^w f(u)\,\mathrm{d}u \end{aligned}$$

where $\rho(v)$ is the queue length density and V is the queue length at the epoch of the arrival of the message[6] and $f(x)$ is the message length density.

Proposition 5. The goodput is given by

$$\mathcal{G}_{\mathrm{f}} = \frac{c}{(h - c\mathrm{e}^{-B\eta/c})}\left[\mathrm{e}^{-B\eta h/((h-c)c)}\left(1 - \frac{c}{h}\right) + \left(\frac{c}{h} - \mathrm{e}^{-B\eta/c}\right)\right]. \qquad (14)$$

Proof. Observe that, for $w \in [0, (B - v)/(h - c)]$,

$$P(\mathcal{V}_{\mathrm{f}} = 1 \mid W \leqslant w,\ V = v) = 1$$

and, for $w > (B - v)/(h - c)$,

$$P(\mathcal{V}_{\mathrm{f}} = 1 \mid W \leqslant w,\ V = v) = P\left(W < \frac{B - v}{h - c} \;\middle|\; W < w\right).$$

Or in other words, for $w \in [0, B/(h - c)]$, if $v \in [0, B - w(h - c)]$

$$P(\mathcal{V}_{\mathrm{f}} = 1 \mid W \leqslant w,\ V = v) = 1,$$

else

$$P(\mathcal{V}_{\mathrm{f}} = 1 \mid W \leqslant w,\ V = v) = \frac{P(W < (B - v)/(h - c))}{P(W < w)}.$$

And for $w > B/(h - c)$,

$$P(\mathcal{V}_{\mathrm{f}} = 1 \mid W \leqslant w,\ V = v) = \frac{P(W < (B - v)/(h - c))}{P(W < w)}.$$

Thus we write, for $w \in [0, B/(h - c)]$, $F(w, 1) = F_1(w) + F_2(w)$, where

$$F_1(w) = \left(1 - \mathrm{e}^{-\eta w}\right)\int_0^{B-w(h-c)} \rho(v)\,\mathrm{d}v = \left(1 - \mathrm{e}^{-\eta w}\right)\left[1 - \frac{h(1 - \mathrm{e}^{-w\eta(h-c)/c})}{(h - c\mathrm{e}^{-B\eta/c})}\right].$$

[6] Due to PASTA the queue length distribution at the arrival epochs of messages, which come as a Poisson stream, is same as the stationary queue length distribution.

Further,

$$F_2(w) = \int_{B-w(h-c)^+}^{B} \big(1 - \mathrm{e}^{-(B-v)/(h-c)\eta}\big)\rho^1(v)\,\mathrm{d}v$$
$$= \frac{h}{(h - c\mathrm{e}^{-B\eta/c})}\left[\big(1 - \mathrm{e}^{-\eta w(h-c)/c}\big) - \frac{(h-c)}{h}\big(1 - \mathrm{e}^{-\eta h w/c}\big)\right]$$

and for $w > B/(h-c)$, $F(w) = \int_0^B (1 - \mathrm{e}^{-(B-v)\eta/(h-c)})\rho(v)\,\mathrm{d}v$. Thus we get

$$\mathrm{d}F_1(w) = \frac{\eta h}{(h - c\mathrm{e}^{-B\eta/c})}\Bigg(\bigg(\mathrm{e}^{-w\eta h/c} - \frac{c}{h}\mathrm{e}^{-\eta B/c}\mathrm{e}^{-\eta w}\bigg)$$
$$+ \bigg(1 - \frac{h}{c}\bigg)\big(\mathrm{e}^{-w\eta(h-c)/c} - \mathrm{e}^{-w\eta h/c}\big)\Bigg)\,\mathrm{d}w,$$
$$\mathrm{d}F_2(w) = \frac{h\eta(h-c)}{c(h - c\mathrm{e}^{-B\eta/c})}\big[\mathrm{e}^{-\eta(h-c)w/c} - \mathrm{e}^{-\eta h w/c}\big]\,\mathrm{d}w.$$

Thus, for $w \in [0, B/(h-c))$,

$$\mathrm{d}F(w,1) = \mathrm{d}F_1(w) + \mathrm{d}F_2(w)$$
$$= \frac{\eta h}{(h - c\mathrm{e}^{-B\eta/c})}\big(\mathrm{e}^{-w\eta h/c} - \frac{c}{h}\mathrm{e}^{-\eta B/c}\mathrm{e}^{-\eta w})\big)\,\mathrm{d}w$$

and for $w > B/(h-c)$, $\mathrm{d}F(w,1) = 0$. Hence we obtain

$$\mathcal{G}_\mathrm{f} = \frac{\eta^2 h}{(h - c\mathrm{e}^{-B\eta/c})}\int_0^{B/(h-c)} w\big(\mathrm{e}^{-w\eta h/c} - \frac{c}{h}\mathrm{e}^{-\eta B/c}\mathrm{e}^{-\eta w}\big)\,\mathrm{d}w$$

from which we obtain equation (14). □

Let us now observe the behavior of $\mathcal{G}_\mathrm{f}$ for extreme values of η, keeping all other parameters fixed. As η tends to zero we see from expression (14) that $\mathcal{G}_\mathrm{f}$ tends to zero. This can be explained by the fact that small η corresponds to very long frames, so that the probability that the queue will fill during the arrival of a message tends to one (since $h > c$).

For the other extreme, i.e. $\eta \to \infty$, the length of a message is very short; one could then expect that the goodput would be equal to the relative amount of fluid that is lost, since a message corresponds to an infinitesimal amount of fluid. This would give a goodput of c/h. This is however not the real limiting value of the goodput: we see, in fact, that as $\eta \to \infty$, we get $\mathcal{G}_\mathrm{f} \to c^2/h^2$ from expression (14). The reason that one could expect to have a goodput of c/h is that this indeed is the fraction of fluid that could be served. So this could give an expression for throughput. But even for a huge buffer, this does not take into account the fact that part of the fluid that is already in the queue corresponds to *bad* packets: they belong to messages in which some packets are dropped. In fact, all the queued fluid of a message that arrives when the amount of fluid hits the boundary is lost. We next provide an intuitive argument through an example that may justify this limiting behavior.

Consider $\eta = 1$, $h = 10$ and $c = 6$. We will show the approximate limit achievable for an expected behavior of our queue and see that it is close to c^2/h^2. We look at the case when a message is discarded, then the expected length of the subsequent discarding period will be 1 ($= 1/\eta$) (the expected remaining length of an $\exp(1)$ distributed random variable). During this period there will be an approximate[7] expected reduction of 6 ($= c/\eta$) in the fluid level. Then a new message, call it message a of expected length 1 starts arriving. The expected queue length at the end of the arrival of this message will approximately $B - 6 + (h - c) = B - 2$. However the expected amount of good fluid that was injected in the queue by this message is approximately 10. Then message b starts arriving, whose expected remaining length is again 1. Thus had message b been completely accepted the expected amount of fluid injected into the queue would have been 10, but because of buffer overflow the expected amount of fluid that can be accepted is approximately $h \times 2/(h - c) = 5$ units. Then the next discarding period starts, whose expected length will again be 1. And the expected amount of fluid that will be discarded in this discarding period will be approximately 10. Thus we have

$$\mathcal{G}_\mathrm{f} = \frac{10 + 10 + \cdots}{10 + 5 + 10 + 10 + 5 + 10 + \cdots}$$

which gives

$$\mathcal{G}_\mathrm{f} = \frac{10}{25} = 0.4 \approx \frac{c^2}{h^2} = 0.36.$$

Another interesting observation from the expression for $\mathcal{G}_\mathrm{f}$ is that *the dependence of $\mathcal{G}_\mathrm{f}$ on different parameters is only through two ratios, c/h and $B\eta/c$.* In particular, *$\mathcal{G}_\mathrm{f}$ is dependent on η and B only through their product.* Also, observe that as B tends to 0, $\mathcal{G}_\mathrm{f}$ tends to 0 and when B tends to ∞, $\mathcal{G}_\mathrm{f}$ tends to c^2/h^2.

3.4. Numerical examples

We shall first plot the density of the stationary workload process $\rho(v)$ and the goodput $\mathcal{G}_\mathrm{f}$ using our analytical expressions for an example with $c = 8$, $h = 12$ and $\eta = 0.6$. To compare the behavior of fluid approximation with the packet model we also plot the queue length distribution and the goodput ratio for the packet model. For the packet model we took $\lambda = h = 12$, $\mu = c = 8$, $q = \eta/\lambda = 0.05$. The plot for $\rho(v)$ for $B = N = 100$ is given in figure 7. We also plot the curves for $\mathcal{G}$ and $\mathcal{G}_\mathrm{f}$ as a function of buffer size in figure 8. We observe that the limiting value of goodput c^2/h^2 by the fluid model is close ot the actual limit of the goodput in the packet model. We next study the behavior of $\mathcal{G}_\mathrm{f}$ as we increase η. Again, we take $c = 8$, $h = 12$ and observe the behavior of $\mathcal{G}_\mathrm{f}$ for $B = 10, 30, 100$ as η increases from 0 to 15 in figure 9. The limiting

[7] Note that the reduction is cannot be greater than B units of fluid, hence, in fact, the reduction is $\simeq \min(B, Xc)$, where $X \simeq \exp(\eta)$. The use of word *approximate/approximately* in subsequent discussion is to highlight the fact that we are approximating a restricted exponential distribution as an exponential distribution.

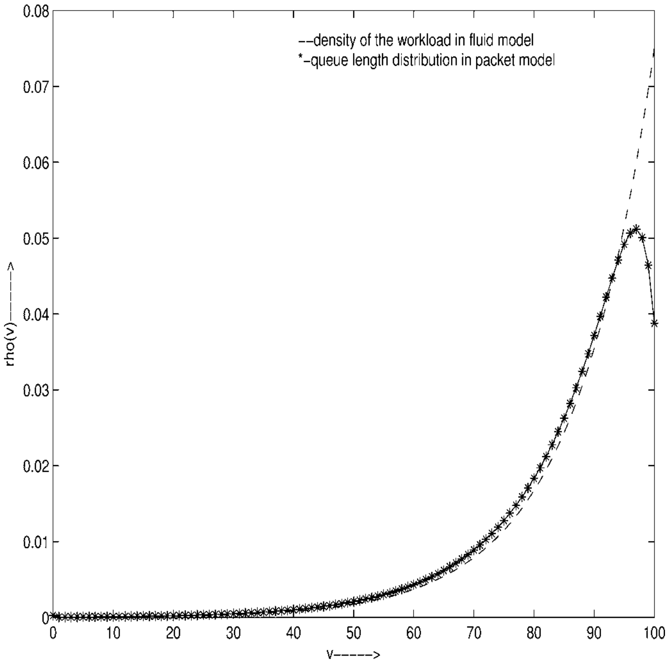

Figure 7. The probability density function $\rho(v)$ of the stationary workload process for the fluid model and the queue length distribution for the packet model.

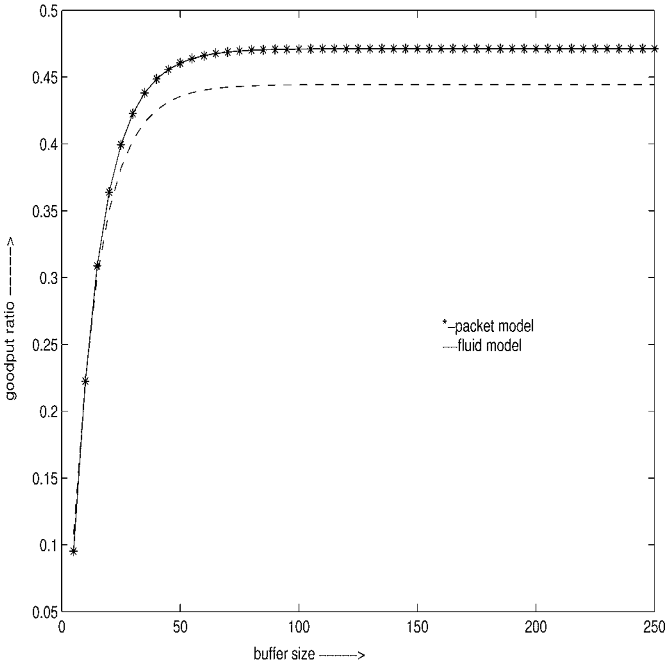

Figure 8. The goodput ratio as a function of buffer size for the packet model and its fluid approximation.

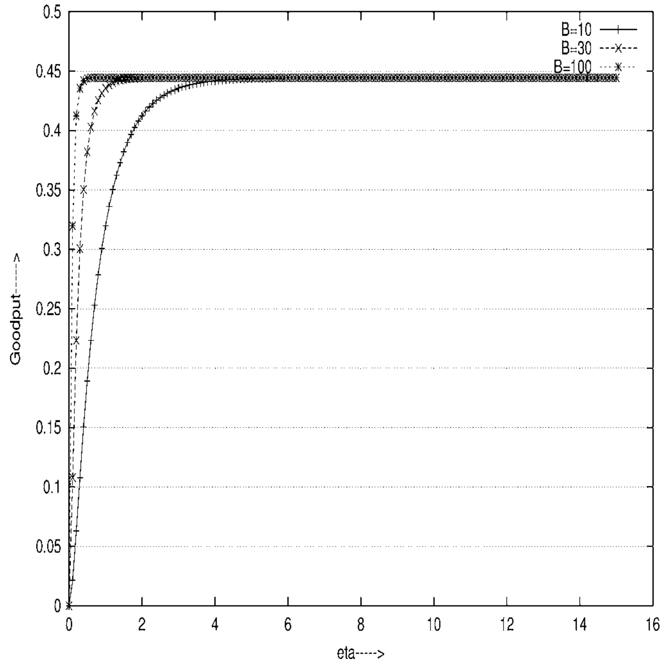

Figure 9. $\mathcal{G}_{\mathrm{f}}$ vs. η for different values of B with $c = 8$ and $h = 12$.

behavior ($\lim_{\eta\to\infty} \mathcal{G}_{\mathrm{f}} = c^2/h^2 = 0.44$) as η increases is seen even at low values of η as B increases. We next show the behavior of $\mathcal{G}_{\mathrm{f}}$ for $\eta = 0, 1, 2, 3, 4, 5$ as B increases from 5 to 30 in figure 10. Again the limiting behavior ($\lim_{B\to\infty} \mathcal{G}_{\mathrm{f}} = c^2/h^2 = 0.44$) as B increases is seen even at low values of B as η increases. Next we keep all other parameters same and take $h = 9$. Thus the limiting $\mathcal{G}_{\mathrm{f}}$ for large B (and also for large η) is 0.79. For different η (B) we plot $\mathcal{G}_{\mathrm{f}}$ in figure 11 (respectively 12) and observe again the limiting behavior for lower h/c.

Remark 4 (A network engineering problem). Consider the case where we want to dimension the buffer size at a network node so as to achieve the maximum goodput for a source when the node employs PMD policy for buffer management. From the expression for $\mathcal{G}_{\mathrm{f}}$ in equation (14) we have

$$\begin{aligned}\frac{\partial \mathcal{G}_{\mathrm{f}}}{\partial B} &= \frac{(\eta/h)\mathrm{e}^{-B\eta/c}\mathrm{e}^{-B\eta h/(h-c)c} - (\eta/h)\mathrm{e}^{-B\eta/c} + (h\eta/c^2)(\mathrm{e}^{-B\eta/c} - \mathrm{e}^{-B\eta h/(h-c)c})}{(h/c - \mathrm{e}^{-B\eta/c})^2}\\ &= \mathrm{e}^{-B\eta/c}\eta h \frac{(1/c^2)(1 - \mathrm{e}^{-B\eta/(h-c)}) - (1/h^2)(1 - \mathrm{e}^{-B\eta h/((h-c)c)})}{(h/c - \mathrm{e}^{-B\eta/c})^2} \geqslant 0.\end{aligned}$$

The non-negativity of $\partial\mathcal{G}_{\mathrm{f}}/\partial B$ follows as $h > c$. Thus the optimum buffer B size (at which the goodput is maximum) is the solution to:

$$\frac{1 - \mathrm{e}^{-B\eta h/((h-c)c)}}{1 - \mathrm{e}^{-B\eta/(h-c)}} = \frac{h^2}{c^2}.$$

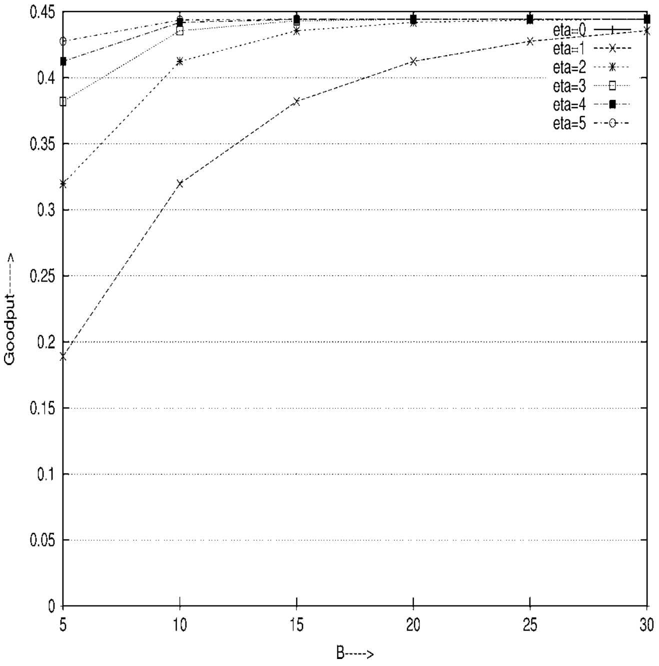

Figure 10. $\mathcal{G}_{\mathrm{f}}$ vs. B for different values of η with $c = 8$ and $h = 12$.

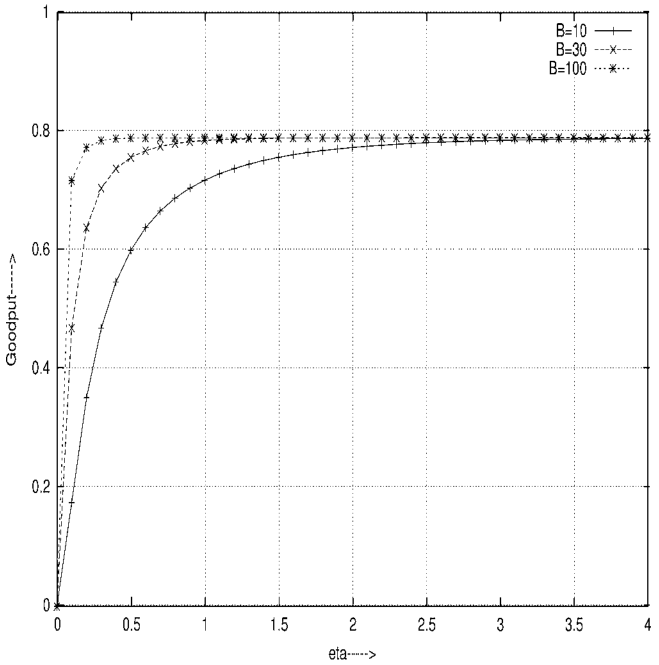

Figure 11. $\mathcal{G}_{\mathrm{f}}$ vs. η for different values of B with $c = 8$ and $h = 9$.

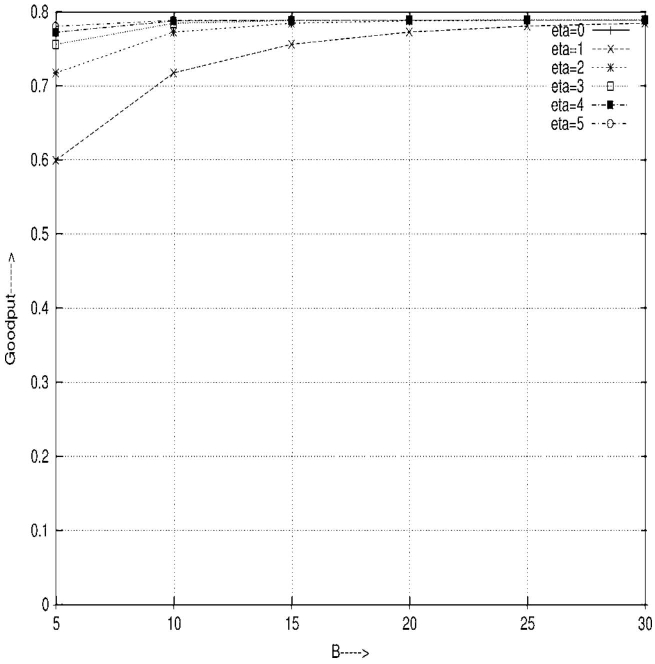

Figure 12. $\mathcal{G}_\mathrm{f}$ vs. B for different values of η with $c = 8$ and $h = 9$.

For multiple solutions we shoose the smallest one, as we have observed a limiting value of $\mathcal{G}_\mathrm{f}$ for large B. This solution (for fluid model) can provide useful engineering guidelines for designing buffer sizes for packet networks.

4. Conclusion

We have provided explicit expressions for the queue size distribution and for the goodput for the packet model based on recursions introduced in [11]. We then provided an alternative fluid approximation for studying the PMD policy and obtained the queue size distribution and goodput in this framework of fluid queue. Our analytical results will be quite useful in dimensioning the buffer size that should be used for a given required goodput under the PMD policy. Also the explicit expressions will be helpful in analytically studying the sensitivity of the goodput to various parameters. Our analysis shows the existence of nontrivial limits of the goodput for different regimes. An interesting exercise can be to obtain the limiting goodput expression (for different limiting behavior, large buffers, small mean message lengths, etc.) for the packet and the fluid model through our closed form expressions. The resulting simpler expressions (expecially for the packet model) can then be studied to provide guidelines for network designing. We are currently studying the potentials of the fluid model as an alternative to the packet model. Our ongoing work include generalizing the fluid model to include Markovian fluids and experimental validations of our models/findings.

Appendix A.

A.1. Proof of proposition 1

Proof. From equations (3) and (5) by applying *z-transform* we get

$$\begin{aligned} &Q_0(z)\big[(\rho+1)-\rho z-z^{-1}\big]-q\rho z Q_1(z)\\ &\quad= P_{0,0}\big[1+\rho-z^{-1}\big]-P_{1,0}-\rho z^{N+1}[P_{N,0}+qP_{N,1}]. \end{aligned} \tag{A.1}$$

Similarly, by applying *z-transform* to equation (4), we get

$$Q_1(z)=P_{0,1}\frac{(1-z-z^{N+1}(q\rho+1)^N q\rho)}{1-z(q\rho+1)}. \tag{A.2}$$

Now from equations (1), (2), (6) and (A.1), we get on solving for $Q_0(z)$

$$Q_0(z)=\frac{P_{0,0}[1-z^{-1}]-\rho z^{N+1}P_{N,0}[1+\rho q]+Q_1(z)q\rho z}{[(\rho+1)-\rho z-z^{-1}]}. \tag{A.3}$$

The zeros of the denominator of equation (A.3) are $z=1,\ \rho^{-1}$. At these values of z, the numerator of equation (A.3) should be equal to 0 because of the analyticity of $Q_0(z)$ (being a polynomial in z of degree less than or equal to N). Thus, substituting $z=1$ in the numerator of equation (A.3) and equating it to 0 we get an equation

$$P_{N,0}=(1+q\rho)^{N-1}qP_{0,1}. \tag{A.4}$$

Also, substituting $z=\rho^{-1}$ in the numerator of equation (A.3) and equating it to 0 we get another equation

$$P_{0,0}(1-\rho)-\rho^{-N}P_{N,0}(1+\rho q)+Q_1\big(\rho^{-1}\big)q=0. \tag{A.5}$$

From equations (A.2), (A.4) and (A.5), we get

$$P_{0,0}=P_{0,1}\frac{q(1-(p^{-1}+q)^N)}{[p(1-q)-1]}. \tag{A.6}$$

We shall interpret equation (7) as

$$Q_0(1)+Q_1(1)=1. \tag{A.7}$$

From equation (A.2)

$$Q_1(1)=(q\rho+1)^N P_{0,1} \tag{A.8}$$

and

$$\dot{Q}_1(1)=\frac{P_{0,1}[1+(\rho q+1)^N(N\rho q-1)]}{q\rho}. \tag{A.9}$$

From equation (A.3) differentiating the numerator and denominator and taking limit as $z \to 1$, we get

$$\begin{aligned} Q_0(1) &= \lim_{z\to 1} Q_0(z) \\ &= \lim_{z\to 1} \frac{z^{-2}P_{0,0} - \rho(N+1)z^N P_{N,0}(1+\rho q) + q\rho(z\dot{Q}_1(z) + Q_1(z))}{(-\rho + z^{-2})}. \end{aligned} \tag{A.10}$$

Thus from equations (A.8), (A.9), (A.4), (A.6) and (A.10) we get

$$Q_0(1) = \frac{P_{0,1}}{(1-\rho)}\left[\frac{q(1-(p^{-1}+q)^N)}{[p(1-q)-1]} + 1 - (1+\rho q)^N\right]. \tag{A.11}$$

Substituting equations (A.11) and (A.8) in equation (A.7), and solving for $P_{0,1}$, we get

$$P_{0,1} = \frac{(1-\rho)(\rho(1-q)-1)}{q(1-\rho^{-N}(1+\rho q)^N) + (\rho(1-q)-1)(1-\rho(1+\rho q)^N)}. \tag{A.12}$$

Knowing $P_{0,1}$ we have obtained the generating functions $Q_0(z)$ and $Q_1(z)$. However, we can further modify the expression for $Q_1(z)$ to a more meaningful form. From equation (A.3) we write, after some algebraic manipulations,

$$\begin{aligned} Q_0(z) = P_{0,1}q\Bigg[&(1+\rho q)^{N-1}\left(\frac{z^{N+1} - \rho^{-(N+1)}}{z - \rho^{-1}}\right) + (1+\rho q)^{N-2}\left(\frac{z^N - \rho^{-N}}{z-\rho^{-1}}\right) \\ &+ \cdots + \frac{z^2 - \rho^{-2}}{z - \rho^{-1}}\Bigg]. \end{aligned}$$

Observe that, each fraction inside the bracket on the right-hand side of the last equation is of the form $(x^k - a^k)/(x-a)$ which simplifies to

$$\frac{x^k - a^k}{x - a} = x^{k-1} + x^{k-1}a + x^{k-2}a^2 + \cdots + xa^{k-2} + a^{k-1},$$

thus we get

$$\begin{aligned} Q_0(z) = P_{0,1}q\rho\Bigg[&(1+\rho q)^{N-1}z^{N+1}\sum_{j=1}^{N+1}\left(\frac{1}{\rho z}\right)^j + (1+\rho q)^{N-2}z^N\sum_{j=1}^{N}\left(\frac{1}{\rho z}\right)^j \\ &+ \cdots + z^2\sum_{j=1}^{2}\left(\frac{1}{\rho z}\right)^j\Bigg]. \end{aligned}$$

Grouping the coefficients of the powers of z we get (8). □

A.2. Exact expression for $\mathcal{G}$

We shall first obtain an expression for $\sum_{i=0}^{N} c_i P(Q = i)$:

$$\begin{aligned}\sum_{i=0}^{N} c_i(x)P(Q = i) &= c_0(x)P(Q = 0) + \sum_{i=1}^{N-1} c_i(x)P(Q = i)\\ &= (1 + K_3A_1) + \big(c_0 - (1 + K_3A_1)\big)P(Q = 0)\\ &\quad + \big[K_4B_1y_1^N - K_3A_2y_1^{N-1}\big]\sum_{i=1}^{N-1} P(Q = i)y_1^{-i}\\ &\quad + \big[K_4B_2y_2^N - K_3A_3y_2^{N-1}\big]\sum_{i=1}^{N-1} P(Q = i)y_2^{-i}\\ &\quad - (1 + K_3A_1)P(Q = N).\end{aligned} \tag{A.13}$$

Observe that

$$\begin{aligned}\sum_{i=1}^{N-1} P(Q = i)y_1^{-i} &= Q_0\big(y_1^{-1}\big) + Q_1\big(y_1^{-1}\big) - P(Q = 0) - P(Q = N)y_1^{-N},\\ \sum_{i=1}^{N-1} P(Q = i)y_2^{-i} &= Q_0\big(y_2^{-1}\big) + Q_1\big(y_2^{-1}\big) - P(Q = 0) - P(Q = N)y_2^{-N}.\end{aligned}$$

Writing $Q(z) = Q_1(z) + Q_2(z)$, from the above equation the expression for $\sum_{i=1}^{N} c_i P(Q = i)$ simplifies to

$$\begin{aligned}&\sum_{i=0}^{N} c_i(x)P(Q = i)\\ &\quad = (1 + K_3A_1)\big(1 - P(Q = N)\big) + \big[K_4B_1y_1^N - K_3A_2y_1^{N-1}\big]Q\big(y_1^{-1}\big)\\ &\qquad + \big[K_4B_2y_2^N - K_3A_3y_2^{N-1}\big]Q\big(y_2^{-1}\big) + K_3\big(A_2y_1^{-1} + A_3y_2^{-1}\big)P(Q = N).\end{aligned}$$

And by proposition 3 we write

$$\mathcal{G} = q^2\left[(1-q)\left(\frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{i=0}^{N} c_i(x)P(Q = i)\right)\right)_{x=(1-q)} + \sum_{i=0}^{N} c_i(1-q)P(Q = i)\right]. \tag{A.14}$$

Thus we need to evaluate $(\mathrm{d}/\mathrm{d}x)(\sum_{i=0}^{N} c_i(x)P(Q = i))$. From the expression for $\sum_{i=0}^{N} c_i(x)P(Q = i)$ from equation (A.13), we write

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{i=0}^{N} c_i P(Q = i)\right)\\ &\quad = \frac{\mathrm{d}}{\mathrm{d}x}K_3A_1\big(1 - P(Q = N)\big) + \big[K_4B_1y_1^N - K_3A_2y_1^{N-1}\big]\frac{\mathrm{d}}{\mathrm{d}x}Q\big(y_1^{-1}\big)\end{aligned}$$

$$
\begin{aligned}
&+ Q\big(y_1^{-1}\big)\frac{\mathrm{d}}{\mathrm{d}x}\big[K_4B_1y_1^N - K_3A_2y_1^{N-1}\big] + \big[K_4B_2y_2^N - K_3A_3y_2^{N-1}\big]\frac{\mathrm{d}}{\mathrm{d}x}Q\big(y_2^{-1}\big)\\
&+ Q\big(y_2^{-1}\big)\frac{\mathrm{d}}{\mathrm{d}x}\big[K_4B_2y_2^N - K_3A_3y_2^{N-1}\big] + \frac{\mathrm{d}}{\mathrm{d}x}\big(K_3(A_2y_1^{-1} + A_3y_2^{-1})\big)P(Q = N).
\end{aligned}
$$

Thus, we need to obtain the derivative terms on the right side of the last equation. We have obtained these terms. The final expressions are provided here:

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}K_3A_1 = \frac{1}{q^2},\\
&\big(K_4B_1y_1^N - K_3A_2y_1^{N-1}\big) = \frac{y_1^{N-1}x\rho}{(1-y_1)\phi_{N+1}\delta}\big(\rho\phi_N - \delta_N(1-y_1)\big),\\
&\big(K_4B_2y_2^N - K_3A_2y_2^{N-1}\big) = -\frac{y_2^{N-1}x\rho}{(1-y_2)\phi_{N+1}\delta}\big(\rho\phi_N - \delta_N(1-y_2)\big),\\
&\frac{\mathrm{d}}{\mathrm{d}x}\big(K_4B_1y_1^N - K_3A_2y_1^{N-1}\big)\\
&\quad = \frac{y_1^{N-1}x\rho}{(1-y_1)\phi_{N+1}\delta}\Bigg[\rho\bigg(x(N-1)\beta_{N-2}\frac{\mathrm{d}y_1}{\mathrm{d}x} + \delta_{N-1} - N\beta_{N-1}\frac{\mathrm{d}y_1}{\mathrm{d}x}\bigg)\\
&\qquad - (1-y_1)N\beta_{N-1}\frac{\mathrm{d}y_1}{\mathrm{d}x} + \delta_N\frac{\mathrm{d}y_1}{\mathrm{d}x} + \big(\rho\phi_N - \delta_N(1-y_1)\big)\\
&\qquad \times\bigg(\frac{1}{x} + \frac{1}{(1-y_1)}\frac{\mathrm{d}y_1}{\mathrm{d}x} + \frac{(N-1)}{y_1}\frac{\mathrm{d}y_1}{\mathrm{d}x} - \frac{1}{\delta}2\frac{\mathrm{d}y_1}{\mathrm{d}x} - \frac{1}{\phi_{N+1}}\frac{\mathrm{d}\phi_{N+1}}{\mathrm{d}x}\bigg)\Bigg],\\
&\frac{\mathrm{d}}{\mathrm{d}x}\big(K_4B_2y_2^N - K_3A_3y_2^{N-1}\big)\\
&\quad = -\frac{y_2^{N-1}x\rho}{(1-y_2)\phi_N\delta}\Bigg[\rho\bigg(x(N-1)\beta_{N-2}\frac{\mathrm{d}y_2}{\mathrm{d}x} + \delta_{N-1} - N\beta_{N-1}\frac{\mathrm{d}y_2}{\mathrm{d}x}\bigg)\\
&\qquad - (1-y_2)N\beta_{N-1}\frac{\mathrm{d}y_2}{\mathrm{d}x} + \delta_N\frac{\mathrm{d}y_2}{\mathrm{d}x} + \big(\rho\phi_N - \delta_N(1-y_2)\big)\\
&\qquad \times\bigg(\frac{1}{x} + \frac{1}{(1-y_2)}\frac{\mathrm{d}y_2}{\mathrm{d}x} + \frac{(N-1)}{y_2}\frac{\mathrm{d}y_2}{\mathrm{d}x} - \frac{1}{\delta}2\frac{\mathrm{d}y_2}{\mathrm{d}x} - \frac{1}{\phi_{N+1}}\frac{\mathrm{d}\phi_{N+1}}{\mathrm{d}x}\bigg)\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
Q(y) = {}&(P_{0,0} + P_{0,1}) + \frac{qP_{0,1}}{(1+q\rho)(1-\rho^{-1}(1+\rho q))}\\
&\times\bigg[\rho(1-q)(1+\rho q)y\frac{1-(1+\rho q)^Ny^N}{1-(1+\rho q)y} - \frac{(1+\rho q)^{N+1}}{\rho^N}y\frac{1-(y\rho)^N}{1-(y\rho)}\bigg]
\end{aligned}
$$

and, finally,

$$
\frac{\mathrm{d}y_1}{\mathrm{d}x} = \frac{-\mathrm{d}y_2}{\mathrm{d}x} = \frac{\rho}{\sqrt{(1+\rho)^2 - 4\rho x}}.
$$

Thus, having obtained all the terms in equation (A.14) we have the explicit expression for $\mathcal{G}$.

References

[1] O. Ait-Hellal, E. Altman, A. Jean-Marie and I.A. Kurkova, On loss probabilities in presence of redundant packets and several traffic sources, Performance Evaluation 36/37 (1999) 485–518.

[2] F. Baccelli and P. Bremaud, *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrence* (Springer, Berlin, 1994).

[3] J.C. Bolot, End-to-end delay and loss behavior in the Internet, in: *Proc. of ACM SIGCOMM '93*, September 1993, pp. 289–298.

[4] J.C. Bolot, Analysis of audio packet loss in the Internet, NOSSDAV (1995).

[5] O.J. Boxma, Sojourn times in cyclic queues – the influence of the slowest server, in: *Computer Performance and Reliability*, eds. G. Iazeolla, P.J. Courtois and O.J. Boxma (Elsevier Science, Amsterdam, 1988) pp. 84–88.

[6] S. Floyd and V. Jacobson, Random early detection gateways for congestion avoidance, IEEE/ACM Trans. Networking 1(4) (1993) 25–39.

[7] S. Floyd and A. Romanow, Dynamics of TCP traffic over ATM networks, in: *Proc. of ACM SIGCOMM'94*, September 1994, pp. 79–88.

[8] S.W. Fuhrmann and R.B. Cooper, Stochastic decomposition in the M/G/1 queue with general vacations, Oper. Res. 33 (1985) 1117–1129.

[9] Y.H. Kim and S.Q. Li, Performance analysis of data packet discarding in ATM networks, IEEE/ACM Trans. Networking 7(2) (1999) 216–227.

[10] D. Kofman and M. Gagnaire, *Réseaux Haut Débit: Réseaux ATM, Réseaux Locaux et Reśeaux Tout-optiques* (Intereditions (groupe Masson), Paris, 1996).

[11] Y. Lapid, R. Rom and M. Sidi, Analysis of discarding policies in high-speed networks, IEEE J. Selected Areas Commun. 16(5) (1998) 764–777.

[12] J.G. Shanthikumar, On stochastic decomposition in $\mathrm{M}/G/1$ type queues with generalized server vacations, Oper. Res. 36(4) (1988) 566–569.

[13] J.S. Turner, Maintaining high throughput during overload in ATM switches, in: *Proc. of INFOCOM '96*, April 1996, pp. 287–295.